Programa de Doctorado Interuniversitario en Criminología

EL FACTOR HUMANO Y LA TRANSFORMACIÓN DIGITAL 4.0 DEL SISTEMA DE JUSTICIA PENAL: ACEPTACIÓN SOCIAL E IMPACTO PROFESIONAL.

Sandra Pérez Domínguez

Director de la tesis

Dr. Fernando Miró Llinares

Codirector de la tesis

Dr. Francisco J. Castro Toledo

Universidad Miguel Hernández de Elche

- 2025 -

La presente Tesis Doctoral, titulada “El factor humano y la transformación digital 4.0 del sistema de justicia penal: aceptación social e impacto profesional”, se presenta bajo la modalidad de tesis convencional con los siguientes indicios de calidad:

1. Pérez Domínguez, S., Castro-Toledo, F. J., & Miró-Llinares, F. (2025). Sesgos y decisiones judiciales en tiempos de IA. ¿Qué sabemos, qué estudiamos?. *Revista Electrónica de Criminología*, 11.
2. Pérez Domínguez, S., & Simón Castellano, P. (2023). Attitudes and perceptions regarding algorithmic judicial judgement: barriers to innovation in the judicial system?. *IDP: Revista de Internet, Derecho y Política*, (39), 6.

El Dr. D. Fernando Miró Llinares, director, y el Dr. D. Francisco Javier Castro Toledo, codirector de la tesis doctoral titulada “El factor humano y la transformación digital 4.0 del sistema de justicia penal: aceptación social e impacto”,

INFORMAN:

Que Dña. Sandra Pérez Domínguez ha realizado bajo nuestra supervisión el trabajo titulado “El factor humano y la transformación digital 4.0 del sistema de justicia penal: aceptación social e impacto profesional” conforme a los términos y condiciones definidos en su Plan de Investigación y de acuerdo al Código de Buenas Prácticas de la Universidad Miguel Hernández de Elche, cumpliendo los objetivos previstos de forma satisfactoria para su defensa pública como tesis doctoral.

Lo que firmamos para los efectos oportunos, en Elche a 8 de enero de 2026.

Director de la tesis	Codirector de la tesis
Dr. D. Fernando Miró Llinares.	Dr. D. Francisco Javier Castro Toledo

El Prof. Dr. Fernando Miró Llinares, Coordinador del Programa de Doctorado en Criminología.

INFORMA:

Que Dña. Sandra Pérez Domínguez ha realizado bajo la supervisión de nuestro Programa de Doctorado el trabajo titulado "El factor humano y la transformación digital 4.0 del sistema de justicia penal: aceptación social e impacto profesional" conforme a los términos y condiciones definidos en su Plan de Investigación y de acuerdo al Código de Buenas Prácticas de la Universidad Miguel Hernández de Elche, cumpliendo los objetivos previstos de forma satisfactoria para su defensa pública como tesis doctoral.

Lo que firmo para los efectos oportunos, en Elche a 8 de enero de 2026.

Prof. Dr. Fernando Miró Linares

Coordinador del Programa de Doctorado en Criminología

FINANCIACIÓN

Esta tesis doctoral ha sido financiada mediante las ayudas para la formación predoctoral en colaboración con empresas, proporcionadas por el Vicerrectorado de Investigación de acuerdo con la Resolución Rectoral 05115/2021.

Los estudios incorporados en los capítulos 4, 5 (en concreto el estudio 2) y 6 se desarrollaron en el contexto del proyecto TED2021-129356B-I00 (Ius_Machina: Sobre las bases normativas y el impacto real de la utilización de algoritmos predictivos en los ámbitos judicial y penitenciario), financiado por MICIU/AEI/10.13039/501100011033 y por la "Unión Europea NextGenerationEU/ PRTR", cuyo investigador principal es Fernando Miró Llinares.

El estudio 1 incorporado en el capítulo 5 se desarrolló en el contexto del proyecto "Inteligencia digital aplicada a la seguridad ciudadana de ámbito local: identificación por profesionales de usos aceptables" en el marco del Convenio con la Diputación de Alicante y la Universidad de Alicante (DIPU-UA1-22X-8, PROYECTO 20).

Esta tesis doctoral ha sido financiada mediante las Ayudas Santander-UMH para estancias de investigación nacionales e internacionales para estudiantado de doctorado de la Universidad Miguel Hernández de Elche 2025 de acuerdo con la Resolución Rectoral 03206/2025

DEDICATORIA

A mis padres, por ser el
mayor ejemplo de esfuerzo y constancia.

Y a ti, solo a ti y para siempre.

AGRADECIMIENTOS

He dejado la redacción de estos agradecimientos prácticamente para el momento final. No por falta de ganas, sino porque, a lo largo de este proceso, han sido tantas las personas que se han preocupado por el desarrollo de esta tesis que sentía que cualquier intento se quedaba corto para devolver todo el cariño y apoyo que he recibido a lo largo de estos cinco años. Aun así, y con la adrenalina del último esfuerzo por finalizar este trabajo, quiero dedicar unas líneas a quienes han hecho que este camino haya sido posible.

Dada la dualidad que caracteriza esta tesis en su vertiente industrial, quiero expresar un agradecimiento profundo y sincero a mis directores de tesis, cuyo acompañamiento ha sido esencial para recorrer y comprender ambos mundos. A Fernando Miró Llinares, por brindarme las herramientas que han hecho posible este trabajo. Gracias por abrirme las puertas al mundo académico y especialmente, del Centro CRÍMINA, por permitirme nutrirme de compañeros, proyectos, congresos y nuevas experiencias, pero especialmente por ofrecerme un lugar donde ser y estar. Gracias por confiar y por seguir retándome con nuevos objetivos. Mi agradecimiento también a Francisco Javier Castro Toledo, mi codirector, por enseñarme que el mundo académico no está reñido con el empresarial, y que lo que hacemos desde la academia puede tener un impacto real más allá de ella. Gracias por creer en mí desde aquellas primeras prácticas y por apostar por la idea de que una tesis criminológica puede tener también una proyección práctica en el ámbito empresarial.

Asimismo, a todos aquellos que formáis parte de CRÍMINA, gracias por soportarme durante estos últimos meses y, sobre todo, por el apoyo incondicional que siempre me habéis dado. CRÍMINA está hecha de personas extraordinarias, y tengo la suerte de haber coincidido con muchas de ellas. Gracias Zora, Mar, Laura, Ana, Álvaro, Mario, Elena, Nacho, Aiala, Victoria y Sara, por estar, por acompañar y por hacer este camino mucho más amable de lo que habría sido sin vosotros. A Esther, por no dejar que pierda mi parte más psicológica, y recodarme porqué elegí estar hoy aquí. A mis *Totatlly Spies,* por entender perfectamente lo que este proceso supone y por seguir apoyándonos juntas, pero especialmente a Rocío, por ser mi salvavidas emocional y mi compañera de biblioteca, sin duda alguna, esta tesis tiene una gran parte de ti. A

Nieves, por ayudarme a transformar mis dudas en resultados significativos y por recordarme que, incluso en los momentos de mayor dispersión, siempre hay una tendencia que seguir. Junto a ellas, también encontré un espacio de refugio con Isa y Bertha en la Red Española de Jóvenes Investigadores en Criminología. Gracias a ellas y a tantos otros jóvenes investigadores he podido nutrirme de ideas que, sin duda, han inspirado y enriquecido parte de la presente tesis. Vorrei inoltre ringraziare espressamente Débora e Federica per aver reso il mio soggiorno a Modena come stare a casa mia. Sarò eternamente grata a entrambe e sappiate che, quando vorrete, potrete contare su di me per qualsiasi cosa.

Tengo la inmensa suerte de contar con una familia y unos amigos que, en gran medida, han hecho posible que esta tesis se haya materializado. Su cariño, su paciencia y su forma de sostenerme en cada paso han sido tan esenciales como cualquier página escrita. A mi *Squad,* para que este texto sea muestra de mis ausencias en momentos clave, y para que sepan que, han sido un gran pilar de desconexión y motivación, especialmente en estas últimas semanas, en esas llamadas durante las travesías que se volvieron un respiro y un refugio; muy especialmente, a esas llamadas. Y a Andrea, por darme la mano hace casi veinte años y no soltarla nunca. Porque, sin duda alguna, fuiste tú quien instauró en mí el esfuerzo y el sacrificio por el estudio en una etapa tan decisiva.

A mi familia. A todos y cada uno de vosotros, que habéis aprendido términos como “tesis”, “estancia”, “defensa” o “depósito” para acompañarme y hacerme sentir comprendida cuando hablaba de esta etapa. Pero, sobre todo, por recordarme cada día la maravillosa suerte que tengo de formar parte de una familia excepcional. A Loreto y a Salva, por confiar tanto en mí que, mucho antes de que esta tesis tomara forma, ya me pedisteis estar en su defensa. Gracias por creer en mí incluso antes de que yo misma lo hiciera. A mi Edi, por ser compañía en momentos de reclusión. Y a mis abuelos, por regalarme un amor tan inmenso como sencillo, por enseñarme valores que han marcado cada uno de mis pasos y por ser siempre ese lugar al que regresar cuando todo alrededor se hacía demasiado grande.

Pero si alguien se merece especialmente estos agradecimientos son ellos, mis padres. Mamá y papá, habéis sido, sois y seréis siempre el espejo donde mirarme,

porque nos habéis antepuesto siempre y habéis luchado por ofrecernos las mejores oportunidades, y muestra de ellos son estas páginas. Os quiero mucho más de lo que os podéis llegar a imaginar.

A Héctor, mi apoyo más incondicional durante los últimos diez años. Sin duda alguna, el orgullo con el que hablas de esta tesis, y de mí, ha sido uno de mis mayores motores. Porque siempre has confiado en mi capacidad para sacarla adelante y no me has dejado caer ni en los momentos más duros del camino. Por entender mis largas ausencias, por aguantarme en mis momentos malos y por, aun en medio de este caos, ofrecerte a construir nuestro hogar. Gracias por sostenerme, por acompañarme y por seguir a mi lado incluso cuando mi vida giraba alrededor de entregas, plazos y capítulos interminables.

Y, por último, a todas aquellas personas que, de un modo u otro, habéis formado parte de este camino. A quienes me ofrecisteis una palabra de ánimo, una conversación, un café, un abrazo o simplemente un respiro cuando más lo necesitaba. Puede que no os nombre uno a uno, pero cada gesto, cada detalle y cada acompañamiento están también en estas páginas. Gracias, de corazón.

INDICE DE CONTENIDOS

LISTADO DE TABLAS.

LISTADO DE FIGURAS.

RESUMEN

La transformación digital 4.0 del sistema de justicia penal, marcada por la digitalización y la incorporación de sistemas algorítmicos e inteligencia artificial, está modificando de forma profunda su funcionamiento. Este proceso, introduce herramientas capaces de analizar grandes volúmenes de datos y apoyar la toma de decisiones, lo que ha generado expectativas relacionadas con la eficiencia y la mejora en la gestión de la información. Sin embargo, también plantea interrogantes sobre la transparencia, la protección de derechos fundamentales y el riesgo de reproducir sesgos presentes en los datos históricos o en los propios modelos de aprendizaje automático. Tanto la ciudadanía como los profesionales perciben esta transición con una mezcla de interés y cautela, especialmente debido a la tensión entre la promesa de objetividad y la necesidad de preservar la supervisión efectiva. Estos elementos sitúan al factor humano como un componente esencial para comprender el alcance y la aceptación de la transformación tecnológica del sistema penal.

Es por ello, que la presente tesis plantea dos objetivos principales: en primer lugar, analizar las percepciones y actitudes de ciudadanía y operadores jurídicos respecto al uso de sistemas algorítmicos e inteligencia artificial en el sistema de justicia penal, y en segundo lugar, examinar la presencia de sesgos en la justicia humana y algorítmica. Para alcanzar estos objetivos se han realizado diversos estudios: una revisión sistemática de la literatura científica, orientada a identificar y clasificar los sesgos presentes tanto en la toma de decisiones humanas como en los sistemas algorítmicos aplicados al ámbito penal; dos investigaciones cualitativas con profesionales de los entornos policial, judicial y penitenciario, que permitieron profundizar en sus percepciones y actitudes ante los procesos de digitalización y la posible automatización de determinadas funciones; y finalmente, se elaboraron tres estudios cuantitativos dirigidos tanto a la ciudadanía como a operadores jurídicos, con el propósito de examinar la aceptación social de estas herramientas y analizar las variables asociadas a dicha aceptación.

Los resultados muestran que tanto ciudadanía como profesionales presentan una aceptación moderada y condicionada al nivel de automatización del uso de

algoritmos en el sistema de justicia penal, especialmente la supervisión humana efectiva, la imparcialidad y la objetividad de las decisiones automatizadas y supervisadas. Asimismo, tanto la ciudadanía como los propios profesionales señalan la necesidad de una formación adecuada, ya que el nivel de capacitación y la familiaridad tecnológica influyen de manera significativa en la aceptación de estas herramientas. Por su parte, los profesionales destacan la importancia de mantener la autonomía y discrecionalidad, comprender el funcionamiento de los sistemas apostando por modelos de colaboración humano – maquina (HITL) frente modelos totalmente automatizados (HOTL). El presente trabajo muestra que ni operadores jurídicos ni ciudadanía conciben el futuro de la justicia como un proceso de automatización plena, sino como un modelo de colaboración en el que la tecnología amplía las capacidades humanas sin sustituirlas. La aceptación social depende, por ello, de factores como la imparcialidad percibida, la transparencia, la proporcionalidad y la existencia de un control humano significativo, lo que demuestra que la confianza en estos sistemas no se sustenta únicamente en su rendimiento técnico. Por último, dado que la presente tesis cuenta con mención industrial, se ha desarrollado una herramienta destinada a evaluar la aceptación social de estas tecnologías en las administraciones públicas, permitiendo apoyar la toma de decisiones institucionales sobre su implantación.

Palabras clave: sistema de justicia penal 4.0; inteligencia artificial; factor humano; aceptación social; impacto profesional.

ABSTRACT

The emergence of "Justice 4.0," shaped by digitalization and the adoption of algorithmic systems and artificial intelligence, is transforming the criminal justice system in profound ways. These technologies introduce new tools capable of processing vast amounts of data and supporting decision-making, raising expectations around efficiency and improved information management. At the same time, they prompt important questions about transparency, the protection of fundamental rights, and the risk of reproducing biases embedded in historical data or in machine-learning models themselves. Both the public and justice professionals view this transition with a mix of interest and caution, aware of the tension between the promise of greater objectivity and the need to maintain meaningful human oversight. This makes the human factor central to understanding how far this technological shift can—and should—go, as well as how it is perceived and accepted.

Against this backdrop, the present dissertation sets out two core aims. The first is to examine how both the public and legal practitioners perceive and respond to the use of algorithmic and AI-driven tools within the criminal justice system. The second is to explore the presence of bias in human decision-making and in algorithmic systems. To address these aims, several studies were conducted: a systematic review of the scientific literature to identify and categorize the forms of bias present in human judgment and in algorithmic tools in criminal justice; two qualitative studies with professionals in policing, the judiciary, and the prison system, allowing for an in-depth exploration of their perceptions of digitalization and the potential automation of certain tasks; and three quantitative studies with members of the public and justice professionals to assess the social acceptance of these technologies and identify the factors shaping such acceptance.

The results indicate that both the public and professionals show moderate support for the use of algorithms, largely dependent on the level of automation involved. Effective human oversight, perceived fairness, and the objectivity of decisions—whether automated or human-supervised—emerge as key conditions for acceptance. Both groups also highlight the need for adequate training, noting that technological competence and familiarity strongly influence willingness to adopt

these tools. Professionals additionally stress the importance of preserving autonomy and discretion, understanding how algorithmic systems operate, and favoring human-in-the-loop models over fully automated ones. Overall, the findings suggest that neither justice professionals nor the general public envisage a future in which justice is fully automated. Instead, they support a collaborative model in which technology augments human capabilities rather than replaces them. Social acceptance thus depends on factors such as perceived impartiality, transparency, proportionality, and the presence of meaningful human control—showing that trust in these systems goes far beyond their technical performance.

Finally, given the industrial component of this dissertation, a tool was developed to assess the social acceptance of algorithmic and AI-based technologies within public administrations, offering guidance for institutional decision-making regarding their implementation.

Keywords: Criminal Justice System 4.0; artificial intelligence; human factor; social acceptance; professional impact.

INTRODUCCIÓN

En el imaginario colectivo, la justicia se ha representado durante siglos como una figura que sostiene una balanza en una mano y una espada en la otra, con los ojos vendados como símbolo de imparcialidad. Esta imagen ha transmitido la idea de una justicia objetiva, equitativa y guiada por la razón, libre de influencias externas. A nivel policial, el escudo representa la función de proteger a la sociedad frente al delito y la violencia, garantizando la seguridad ciudadana dentro de la legalidad y el respeto a los derechos; mientras que, a nivel penitenciario, la llave expresa la doble misión de privación de libertad y reinserción, pues no solo asegura el cumplimiento de la pena, sino que también abre la posibilidad de un nuevo comienzo orientado a la rehabilitación y la reintegración social.

Sin embargo, en los últimos años esa representación clásica y sus variaciones simbólicas se han visto desplazadas por nuevas imágenes: un robot que sostiene un mazo judicial, un dron patrullando las ciudades o algoritmos que deciden sobre procesos penitenciarios. Estos elementos reflejan cómo la transformación tecnológica comienza a condicionar dichos ámbitos y abren interrogantes cruciales acerca del papel que deben desempeñar la tecnología y el factor humano, pues las decisiones ya no dependen únicamente del juicio humano, sino que están cada vez más influenciadas por algoritmos y sistemas de inteligencia artificial, introduciendo una profunda reconfiguración de la justicia penal. Aquello que tradicionalmente se entendía como un espacio reservado al criterio humano y a los principios éticos del derecho, se enfrenta ahora a un escenario incierto, donde la incorporación de herramientas automatizadas tiende a desdibujar los límites entre la razón jurídica y la lógica algorítmica. Pero ¿cómo de factible es la idea de que la IA pueda reemplazar a los profesionales humanos?, ¿podemos esperar que el desarrollo de la IA asuma el rol que desempeñan los profesionales del sistema de justicia penal en un futuro cercano?; y la más importante, ¿cómo de dispuestos estamos a que eso ocurra?.

Como ya he anticipado, y dado que la transformación tecnológica resulta tan amplia, este trabajo la abordará en concreto en el ámbito denominado como *criminal justice system*, traducido al español como sistema de justicia penal. Este concepto hace referencia al conjunto de instituciones, normas y prácticas orientadas a investigar,

procesar y sancionar conductas delictivas. La traducción no es meramente literal: en inglés el énfasis recae en la noción de justicia como un sistema integral. Por ello, resulta relevante precisar que, en este trabajo, ambos términos se utilizan de manera equivalente, aunque se reconocerán las diferencias culturales y jurídicas que implica cada tradición. El sistema de justicia penal integra diversas fases y actores: la policía en la etapa de investigación, la fiscalía y los tribunales en la fase de enjuiciamiento, y las instituciones penitenciarias o alternativas comunitarias en la ejecución de sanciones (Waldron, Quarles, McElreath, Waldron & Milstein, 2009). Cada uno de estos componentes cumple funciones específicas, pero forman parte de un entramado que busca mantener el orden social y garantizar el respeto a los derechos fundamentales.

Pero esta transformación digital del sistema de justicia penal, no se limita únicamente a la incorporación de herramientas tecnológicas destinadas a la gestión de expedientes o a la modernización de los procesos administrativos. Por el contrario, esta transformación, impulsada por el incremento de la capacidad informática y la disponibilidad masiva de datos, ha propiciado la exploración del uso de técnicas de *machine learning* y otras formas de inteligencia artificial orientadas a mejorar la toma de decisiones en el ámbito de la justicia penal (Miró Llinares, 2020). En este nuevo paradigma, las tecnologías basadas en inteligencia artificial prometen, al menos en el plano teórico, reducir significativamente los tiempos de respuesta del sistema judicial, optimizar la asignación de recursos y proporcionar un cierto grado de objetividad en la valoración de los casos (Alarie, Niblett & Yoon, 2018; de Sousa et al., 2022; Farfán Intriago et al., 2023; Fine, Miller, & Le, 2024; Socol de la Osa & Remolina, 2024). Esta serie de promesas juntos con los problemas inherentes al juicio humano, como la influencia de emociones, sesgos implícitos y limitaciones cognitivas propias de la condición humana (Tversky & Kahneman, 1974; Fischhoff, 1975; Goodman-Delahunty & Sporer, 2010; Muñoz Aranguren, 2011; Maroney, 2011; Barry, 2021; Forza, Menegon & Rumiati, 2024), podría servir de caldo de cultivo para que las nuevas tecnologías se proyecten como un horizonte potencialmente más prometedor para la toma de decisiones en el ámbito penal.

Ahora bien, esta transformación digital del sistema judicial no está exenta de riesgos

ni de desafíos éticos. Lejos de garantizar decisiones libres de sesgos, la automatización puede, en muchos casos, reproducir e incluso amplificar prejuicios preexistentes al estar basada en datos históricos atravesados por desigualdades estructurales (Dou & Dou, 2025). Como han evidenciado Buolamwini y Gebru (2018), incluso sistemas ampliamente comercializados, como los de reconocimiento facial, presentan errores significativamente mayores al identificar a mujeres racializadas, lo que demuestra cómo la tecnología puede incorporar sesgos de género y raza si no se desarrolla con una perspectiva crítica. En el ámbito judicial, esta problemática se agrava, ya que las decisiones automatizadas influyen directamente en derechos fundamentales. Organismos internacionales como la UNESCO (2021) han alertado sobre la necesidad de garantizar la transparencia, la supervisión humana y el respeto a los principios de no discriminación en el diseño y uso de sistemas de inteligencia artificial. Así, la promesa de eficiencia y objetividad en la justicia digital debe ser equilibrada con un enfoque riguroso que evite que las tecnologías amplifiquen las injusticias que pretenden corregir (Krištofík, 2025).

En este escenario de tensiones entre eficiencia tecnológica y garantías jurídicas, la justicia algorítmica *(algorithmic fairness)* ha emergido como una propuesta orientada a corregir o mitigar los sesgos presentes en los sistemas automatizados de toma de decisiones (Lee, Jain, Cha, Ojha & 2019; Hellman, 2020; Berk, Heidari, Jabbari, Kearns & Roth, 2021; Mitchell, Potash, Barocas, D'Amour & Lum, 2021; Wang, Zhang & Zhu, 2022). Esta corriente busca garantizar que los algoritmos operen de manera equitativa, evitando la reproducción de desigualdades estructurales o la discriminación indirecta derivada del uso de datos históricos sesgados (Mitchell et al. 2005; Saavedra-Vera et al., 2023). No obstante, incluso cuando estas herramientas son diseñadas conforme a criterios de equidad, su aplicación en el ámbito penal entraña riesgos significativos, ya que pueden verse afectadas por la calidad de los datos de entrenamiento o por la selección de variables que, de forma indirecta, generan impactos discriminatorios. La ausencia de mecanismos efectivos de supervisión y transparencia agrava esta situación, comprometiendo la protección de derechos fundamentales. Por ello, el diseño y la implementación de algoritmos en la justicia penal requieren una evaluación crítica que considere su legitimidad, el impacto de sus decisiones y las garantías necesarias

para evitar resultados injustos (Cerezo-Martínez et al., 2024; Castro-Toledo, 2022; Slobogin, 2021).

Frente a estos riesgos y dilemas, las instituciones nacionales e internacionales han comenzado a articular respuestas normativas orientadas a mitigar los posibles efectos negativos del uso de la inteligencia artificial en el ámbito judicial. En el ámbito europeo, la regulación de la inteligencia artificial ha comenzado a tomar forma con iniciativas como el Reglamento Europeo de Inteligencia artificial (*IA Act*) y la Convención Marco sobre IA del Consejo de Europa, que buscan establecer un marco ético y jurídico que asegure un uso responsable y transparente de estas tecnologías.

Es con la formulación de estas normativas y directrices donde surge y se consolida el concepto de supervisión humana efectiva, entendido como la capacidad real de los operadores jurídicos para comprender, intervenir y, en su caso, corregir las decisiones generadas por sistemas automatizados. Esta supervisión se considera esencial en ámbitos de alto riesgo como la justicia penal, ya que pretende garantizar que el uso de la inteligencia artificial no sustituya el juicio crítico humano ni vulnere derechos fundamentales. Esta atención normativa al rol humano no es casual: responde a una inquietud social más profunda sobre el lugar de las personas en un entorno crecientemente automatizado. La irrupción de la digitalización y la IA no ha supuesto únicamente una transformación tecnológica, sino también una reacción profundamente humana: la necesidad de comprender qué es esta tecnología, cuál es su origen, hacia dónde se dirige y, especialmente, cómo incide en nuestras prácticas y decisiones. Esta inquietud pone de manifiesto el papel central del factor humano en el proceso de digitalización del sistema de justicia, donde no basta con incorporar nuevas herramientas: resulta imprescindible analizar su grado de aceptación, la manera en que son percibidas y los efectos que generan sobre quienes deben integrarlas en su labor cotidiana.

En este contexto, el factor humano se configura como un elemento clave para comprender tanto el presente como el futuro de la justicia digital. La aceptación o el rechazo de las tecnologías algorítmicas, la confianza en los sistemas automatizados, el nivel de comprensión sobre su funcionamiento y la percepción de imparcialidad

son variables que inciden de forma decisiva en su implementación efectiva (Mahmud, Islam, Ahmed, & Smolander, 2022). En esta tesis, el factor humano se entiende como el conjunto de percepciones, actitudes, competencias, experiencias y marcos normativos que operadores jurídicos y ciudadanía movilizan al interactuar con herramientas digitales y sistemas algorítmicos. En este contexto, el factor humano se manifiesta en distintas fases, desde la supervisión crítica de los resultados, el diseño y desarrollo de los sistemas por equipos multidisciplinares, el uso cotidiano de los operadores, la regulación normativa que delimita su alcance o la recepción ciudadana de sus efectos. Lejos de desempeñar un rol secundario, jueces, fiscales, cuerpos policiales y demás profesionales del sistema de justicia penal cumplen una función determinante en el uso de estas herramientas, ya que no solo interpretan los resultados generados por los sistemas algorítmicos, sino que, además, asumen las decisiones adoptadas por las herramientas. Esta responsabilidad no se limita al plano formal, pues variables personales como su experiencia, sus creencias o sus actitudes también median en la aceptación y el uso efectivo de dichas herramientas. No obstante, mientras que la literatura especializada ha prestado una atención considerable a la supervisión de resultados y al diseño de marcos regulatorios, las actitudes de los operadores como usuarios de estas herramientas y de la ciudadanía como destinataria final de sus decisiones han sido menos exploradas, pese a su relevancia para comprender la aceptación de la transformación tecnológica de la justicia.

La presente tesis doctoral centra su análisis precisamente en estas dos últimas dimensiones: las actitudes de los operadores jurídicos y de la ciudadanía, partiendo de la convicción de que la confianza, la percepción de imparcialidad y la aceptación social constituyen condiciones esenciales para la consolidación de las innovaciones tecnológicas en el sistema de justicia penal. La tesis se sustenta en la hipótesis de que el éxito de la transformación digital de la justicia penal no depende únicamente de los avances tecnológicos ni del desarrollo normativo, sino también, y de manera decisiva, del factor humano, entendido tanto como la capacidad de los profesionales para integrar estas herramientas en su práctica cotidiana, como por el nivel de aceptación social que generan sus efectos en quienes resultan directamente afectados por las decisiones. Ahora bien, dicha aceptación no puede concebirse

como un fenómeno pasivo, sino que requiere ser evaluada mediante mecanismos de participación que permitan a la ciudadanía expresar sus percepciones, expectativas y preocupaciones en torno al uso de herramientas algorítmicas en el sistema de justicia penal. En consecuencia, resulta imprescindible analizar empíricamente cómo son percibidas estas tecnologías, qué actitudes suscitan y cuáles son los factores que condicionan su aceptación o rechazo. La escasez de estudios que aborden esta cuestión, especialmente en el contexto español, constituye la principal motivación de la presente investigación.

Teniendo como ruta dicho objetivo, la presente tesis doctoral se estructura en tres partes principales, la *Primera Parte*, titulada "Fundamentos teóricos sobre la digitalización y el factor humano en el sistema de justicia penal 4.0.", desarrolla el contexto conceptual y normativo de la digitalización y algoritmización de la justicia penal. En el *Capítulo 1*, se presenta una aproximación a dicha transformación, atendiendo tanto a la evolución de los conceptos clave, digitalización, algoritmos e inteligencia artificial, como a los factores que han favorecido su impulso, entre ellos la necesidad de reforzar la confianza social en la justicia, la subjetividad humana o las deficiencias estructurales del propio sistema. Asimismo, se expone el panorama actual de la implementación tecnológica en los ámbitos judicial, penitenciario y policial, y se introducen las principales expectativas, beneficios y riesgos asociados a su uso. Finalmente, el capítulo cierra con una presentación de los marcos normativos y principios éticos más relevantes a nivel europeo e internacional que regulan estas herramientas.

El Capítulo 2 aborda el análisis del factor humano en la transformación tecnológica del sistema de justicia penal, se expone cómo esta transformación plantea tensiones entre la posible deshumanización de la justicia y el potencial apoyo a la toma de decisiones. En este marco, se presentan los principales modelos de interacción entre operadores jurídicos y herramientas digitales, con el fin de ilustrar los distintos grados de supervisión humana posibles en contextos algorítmicos. La última sección se centra en la aceptación social del uso de herramientas digitales en el sistema de justicia penal. Para ello, se examinan distintos modelos teóricos que explican la aceptación tecnológica y se analizan las actitudes sociales frente a la digitalización

del sistema. Se incorpora, además, el papel de la participación ciudadana como mecanismo para evaluar la aceptación de estas herramientas, subrayando su valor no solo consultivo, sino como vía esencial para una gobernanza algorítmica democrática. Finalmente, se identifican y desarrollan los principales factores que influyen en la aceptación de herramientas algorítmicas, como la transparencia y explicabilidad de los algoritmos, la existencia de supervisión humana y atribución de responsabilidad, la percepción de imparcialidad en los resultados, y el nivel de familiaridad tecnológica de los usuarios implicados.

La *Segunda Parte*, *"Estudios sobre el factor humano en el uso de algoritmos en el sistema de justicia penal"*, recoge la investigación empírica desarrollada. En el *Capítulo 3,* se presenta la justificación de la investigación, los objetivos y el enfoque científico, detallando el diseño metodológico y los instrumentos de análisis utilizados. Los *Capítulos 4, 5 y 6* recogen el desarrollo y los resultados de los estudios empíricos realizados en el marco de esta tesis, combinando metodologías cualitativas y cuantitativas para ofrecer un análisis exhaustivo del papel del factor humano en la digitalización del sistema de justicia penal.

El *Capítulo 4* se centra en el estudio de los sesgos presentes en la toma de decisiones judiciales, a partir de una revisión sistemática de la literatura científica. El objetivo principal es identificar y clasificar los distintos tipos de sesgos que afectan tanto a los operadores jurídicos como a las herramientas algorítmicas implementadas en el sistema de justicia penal. Para ello, se analizan cinco grandes categorías: los sesgos demográficos (como los basados en género, raza o estatus socioeconómico), los sesgos cognitivos (relacionados con heurísticos mentales), los sesgos digitales (derivados del uso de inteligencia artificial y algoritmos predictivos), los condicionamientos contextuales (vinculados al entorno social y organizacional del juez) y los sesgos lingüísticos (relacionados con la interpretación del lenguaje jurídico). A lo largo del capítulo, se evidencia una evolución del enfoque académico desde el análisis de los sesgos humanos hacia los riesgos emergentes vinculados a la automatización, sin que ello implique su superación. El capítulo concluye destacando la necesidad de desarrollar marcos interdisciplinares de análisis y líneas futuras de investigación centradas en la aceptación social y la gobernanza

responsable de la justicia digital.

En el *Capítulo 5*, se presentan dos estudios cualitativos centrados en explorar las percepciones y experiencias de los operadores jurídicos ante la incorporación de tecnologías digitales y algoritmos predictivos en su labor diaria. El primero de estos estudios se desarrolla en el ámbito de la seguridad ciudadana, y analiza el impacto de la digitalización en los procesos policiales, particularmente en relación con la digitalización de las tareas propias del ámbito de la seguridad. El segundo estudio se sitúa en el contexto judicial y penitenciario, abordando los posibles usos que la automatización y el uso de inteligencia artificial tienen en la práctica profesional de jueces y fiscales e identificando los principales desafíos, resistencias y oportunidades que perciben en relación con estas tecnologías.

El *Capítulo 6* incluye tres estudios que abordan las actitudes sociales hacia el uso de algoritmos y sistemas de inteligencia artificial en el ámbito de la justicia penal. El primer y segundo estudio, a través de diferentes marcos teóricos, profundiza en las barreras percibidas para la implementación de la justicia algorítmica, tanto desde la perspectiva de los operadores jurídicos como de la ciudadanía, e identifica los factores que pueden facilitar o dificultar la innovación tecnológica en el sistema de justicia penal. Finalmente, el tercer estudio emplea una metodología experimental para analizar las actitudes sociales que determinan la aceptación del uso de inteligencia artificial en la toma de decisiones judiciales y penitenciarias, considerando tanto la perspectiva de la ciudadanía como la de los profesionales, a través de diferentes escenarios de utilización de las herramientas y niveles de intervención humana.

La *Tercera Parte*, dedicada a la mención industrial y titulada *"Transferencia a la práctica profesional"*, se describe el know-how transferido a la empresa Plus Ethics, abordando cómo los resultados de la tesis se han convertido en una herramienta operativa diseñada para evaluar la aceptación del uso de la inteligencia artificial en la Administración Pública. Esta sección describe por qué era necesario desarrollar un instrumento de este tipo, cómo se ha construido a partir de la evidencia científica y de los modelos de aceptación tecnológica, y de qué manera su estructura, basada en módulos de definición del caso de uso, recogida de percepciones profesionales y

análisis automatizado, permite aplicarla en contextos reales. También detalla el proceso de funcionamiento, el papel de los distintos actores implicados, las modalidades de explotación y las condiciones técnicas y organizativas que permiten que la herramienta pueda ser utilizada por administraciones públicas para apoyar decisiones sobre la implantación responsable de sistemas de inteligencia artificial.

Finalmente, la Cuarta Parte, dedicada a la discusión y conclusiones, ofrece una síntesis integradora de los estudios desarrollados y expone las reflexiones generales de la investigación. Esta sección analiza cómo la digitalización, la algoritmización y la inteligencia artificial están reconfigurando el sistema de justicia penal; revisa los principales hallazgos empíricos sobre sesgos, percepciones y actitudes; y destaca el papel central del factor humano en la adopción y supervisión de estas tecnologías. Asimismo, identifica las limitaciones metodológicas del trabajo y plantea conclusiones orientadas a comprender las condiciones éticas, sociales y organizativas necesarias para una implementación legítima, responsable y alineada con los principios del Estado de derecho. Además, con el fin de obtener el doctorado con mención internacional, se incluye un apartado de conclusiones generales en inglés.

PARTE I. FUNDAMENTOS TEÓRICOS SOBRE LA DIGITALIZACIÓN Y EL FACTOR HUMANO EN EL SISTEMA DE JUSTICIA PENAL 4.0.

CAPÍTULO 1. LA TRANSFORMACIÓN DEL SISTEMA DE JUSTICIA PENAL 4.0.

1. La revolución del sistema de justicia penal ante la digitalización, algoritmización y el uso de Inteligencia artificial.

El concepto de "revolución" se define como un cambio profundo, radical y generalmente rápido en un sistema establecido, ya sea social, político, económico, cultural, científico o tecnológico. A lo largo de la historia, las sociedades han vivido grandes revoluciones que han permitido el avance y la transformación de sus estructuras fundamentales. En el contexto contemporáneo, nos encontramos ante la denominada Cuarta Revolución Industrial (Schwab, 2016), expresión acuñada en 2011 en la Hannover Messe[1] para describir la integración de tecnologías emergentes que están transformando los modos de producción, organización y comunicación a escala global. Este nuevo estadio se caracteriza por la interconexión de sistemas, la digitalización integral de procesos y el despliegue de nuevas tecnologías tales como la inteligencia artificial, el *big data*, el *machine learning*, la robótica avanzada, la computación en la nube y la interconexión de dispositivos en red, configurando así un entorno cada vez más digitalizado e interdependiente (Aguilar, 2021; Jain & Murugesan, 2021). La incorporación de estas innovaciones no solo ha reconfigurado las dinámicas sociales y económicas, sino que también ha impactado de manera significativa en sectores clave como la justicia, que enfrenta hoy el desafío de adaptarse a esta nueva realidad tecnológica (Hildebrandt, 2015; Susskind, 2019).

La incorporación de tecnologías emergentes en el sistema de justicia penal ha

[1] Se trata de una de las ferias industriales más grandes e importantes del mundo. Se celebra cada año en Hannover, Alemania, y reúne a empresas, expertos, gobiernos e instituciones de todo el mundo para presentar y debatir sobre los avances más recientes en tecnología industrial, automatización, digitalización, energía y manufactura avanzada.

impulsado el surgimiento de lo que, por analogía con la Cuarta Revolución Industrial, se denomina Justicia 4.0, concepto que en este contexto se materializa en la idea de un Sistema de Justicia Penal 4.0. Este concepto hace referencia a la aplicación de sus principales tecnologías al ámbito de la justicia, con el propósito de construir órganos judiciales más eficaces, accesibles y transparentes, capaces de responder con agilidad a las demandas ciudadanas y sociales (Barona, 2019a). Esta transformación tecnológica no se limita a la incorporación instrumental de nuevas herramientas, sino que implica una reconfiguración profunda de las formas en que se conciben, organizan y gestionan los procesos judiciales. Se trata de un cambio paradigmático que afecta tanto a la dimensión estructural como a la funcional del sistema penal, en la medida en que transforma los procedimientos, redefine las funciones de los actores involucrados e introduce nuevos dilemas vinculados con la legitimidad, las garantías procesales y los riesgos asociados al uso de sistemas algorítmicos. Esta nueva transformación representa un cambio de tal magnitud que abordarla en su totalidad resultaría inabarcable y, además, excedería los objetivos específicos planteados en esta tesis doctoral. No obstante, con el fin de contextualizar adecuadamente el objeto de estudio, resulta imprescindible realizar un análisis general de su evolución en el contexto del sistema de justicia penal, así como de las implicaciones que conlleva en términos de beneficios, riesgos y marcos regulatorios. Este es, precisamente, el propósito del presente capítulo: ofrecer una visión comprensiva del contexto en el que se está configurando la digitalización del sistema de justicia penal.

Históricamente, el sistema de justicia penal se ha caracterizado por un fuerte componente humano, basado en la presencialidad de las audiencias, el uso de expedientes en papel y la realización de trámites burocráticos y formales que, en muchos casos, resultaban lentos e ineficientes. Sin embargo, la irrupción del universo digital ha puesto de manifiesto las limitaciones de este modelo, evidenciando la necesidad de repensar sus estructuras y prácticas. De este modo, la transformación digital del sistema de justicia penal no solo supone la transición desde un soporte físico hacia sistemas informáticos, sino que plantea la oportunidad y el desafío de articular procedimientos más ágiles, accesibles y fundamentados en la evidencia. En este sentido, uno de los avances más significativos ha sido la

incorporación de soluciones algorítmicas capaces de gestionar, procesar y analizar grandes volúmenes de información con altos niveles de precisión. Tales capacidades han alimentado un intenso debate sobre la conveniencia de integrar las nuevas tecnologías en la administración de justicia. De hecho, investigaciones recientes han mostrado que la digitalización y la innovación en justicia se orientan principalmente hacia la eficiencia, la accesibilidad y la participación ciudadana, aunque no están exentas de desafíos relacionados con la compatibilidad organizativa, la aceptación de los profesionales y la heterogeneidad de los resultados alcanzados (Correia, Pereira, & Bilhim, 2024). Asimismo, la transformación digital también se extiende en los ámbitos policial y penitenciario. En el ámbito penitenciario, la digitalización ha impulsado la gestión inteligente de establecimientos a través de sistemas de información penitenciaria, videovigilancia avanzada, control biométrico y programas de seguimiento electrónico (Fedorczyk, 2024; Güerri Ferrández, Martí Barrachina & Pedrosa, 2021; López Lorca, 2023). Estas innovaciones permiten optimizar la clasificación, el seguimiento y la reinserción de las personas privadas de libertad, además de facilitar la coordinación interinstitucional entre juzgados, fiscalías y servicios penitenciarios. En el ámbito policial, la adopción de tecnologías predictivas y de análisis de datos masivos, como los modelos de predictive policing, han transformado los procesos de patrullaje, investigación y prevención del delito (Perry, 2013; Ferguson, 2017). Tales herramientas permiten anticipar zonas o individuos con mayor probabilidad de implicación delictiva, mejorando la asignación de recursos y la toma de decisiones estratégicas, aunque también generan preocupaciones sobre sesgos algorítmicos, transparencia y respeto a los derechos fundamentales (Brantingham, Valasik & Mohler, 2018).

En este contexto de modernización y búsqueda de soluciones más ágiles, han surgido también aplicaciones específicas como las denominadas Herramientas de Valoración del Riesgo (HVR), diseñadas para estimar la probabilidad de reincidencia o de comisión de nuevos delitos por parte de las personas evaluadas (Brandariz García, 2016). Estas herramientas buscan dotar de mayor objetividad y rigor empírico a procedimientos históricamente marcados por la subjetividad inherente al juicio humano. La evolución tecnológica, particularmente el crecimiento exponencial en la capacidad de almacenamiento y procesamiento de datos ha

permitido perfeccionar estas soluciones, dando lugar a sistemas más avanzados sustentados en técnicas de machine learning y métodos de inteligencia artificial que abren nuevas posibilidades para la toma de decisiones en el ámbito penal (Miró Llinares, 2020). Dichas tecnologías no solo permiten identificar patrones complejos en grandes volúmenes de datos, sino también realizar análisis predictivos de alta precisión, anticipando comportamientos o riesgos potenciales y contribuyendo así a optimizar la asignación de recursos y la adopción de medidas judiciales o penitenciarias más eficaces. En esta línea, las soluciones algorítmicas aplicadas a la determinación judicial del delito o de la pena han adquirido un protagonismo creciente. Bajo los conceptos de *algorithmic justice, predictive sentencing* o *evidence-based sentencing* se agrupan sistemas actuariales algorítmicos de apoyo a la decisión judicial que operan mediante la evaluación de factores de riesgo de reincidencia, especialmente en delincuentes adultos, como lo demuestran los estudios pioneros de Hanson y Bussiere (1998) y de Gendreau, Little y Goggin (1996). Estos sistemas buscan proporcionar información adicional a los tribunales para fundamentar decisiones críticas tales como la concesión de la libertad provisional o condicional, la clasificación penitenciaria o la reubicación de personas privadas de libertad en distintos regímenes penitenciarios (Rizer & Watney, 2018). La disponibilidad de conjuntos de datos robustos ha sido clave para consolidar este enfoque predictivo, inicialmente desarrollado para delitos violentos y posteriormente extendido a otros ámbitos delictivos. Entre las metodologías más relevantes se encuentran el HCR-20 para la violencia general (Douglas et al., 2013), el SAVRY para la violencia juvenil (Borum, Bartel, & Forth, 2005), el SVR-20 para la violencia sexual (Boer et al., 1997) y el SARA para la violencia contra la pareja (Kropp et al., 1995). De forma complementaria, en los sistemas penitenciarios se observa un proceso paralelo de tecnificación orientado a la prevención de incidentes y la reinserción social. La implementación de plataformas de gestión integral de internos, sistemas de monitoreo electrónico y programas de evaluación del riesgo de violencia institucional han demostrado mejorar la seguridad y la eficiencia operativa. En el entorno policial, la analítica de datos y la inteligencia artificial se emplean cada vez más para la detección temprana de patrones criminales, el análisis de redes delictivas y la gestión de recursos, marcando el tránsito hacia modelos de policía basada en la evidencia (Ratcliffe, 2019).

No obstante, la transición hacia una justicia digital no está exenta de desafíos. El paso de procedimientos manuales a sistemas digitalizados supone una revolución en la administración judicial, pero enfrenta resistencias culturales e institucionales, así como limitaciones derivadas de la falta de capacitación de los operadores jurídicos (Bertot, Jaeger, & Grimes, 2010). Además, si bien la digitalización puede mejorar la eficiencia y la accesibilidad, su implementación debe llevarse a cabo con cautela para no socavar la aceptación del sistema ni comprometer las garantías procesales (van den Bos, 2001).

El caso de España resulta ilustrativo en este sentido, pues la transformación digital del sistema judicial se ha consolidado como una prioridad en las estrategias gubernamentales recientes, enmarcadas en la necesidad de superar las limitaciones del modelo tradicional y ofrecer un servicio público más ágil, interoperable y adaptado a la era digital. Entre los principales instrumentos estratégicos de modernización[2] se encuentra el Plan Justicia 2030, promovido por el Ministerio de Justicia de España, que busca alinear la transformación del sistema con los principios de la Agenda 2030 de Naciones Unidas y con las estrategias digitales de la Unión Europea (Ministerio de Justicia, 2020). Dicho plan se articula en torno a tres grandes objetivos: la mejora de la eficiencia interna del sistema judicial mediante la incorporación de herramientas digitales y modelos de gestión flexibles y sostenibles; el fortalecimiento de la relación entre la justicia y la ciudadanía, garantizando transparencia, accesibilidad e inclusión; y la vinculación del servicio público de justicia con los retos estratégicos nacionales e internacionales, tales como la reactivación económica, la transición ecológica, la cohesión social, la educación, la protección de los derechos humanos y la prevención de la corrupción. De esta manera, la digitalización del sistema de justicia penal no se concibe únicamente como un mecanismo técnico de modernización, sino como un agente transformador capaz de redefinir el papel de la justicia en la sociedad contemporánea. Este proceso

[2] Además de las iniciativas estratégicas y programáticas aquí expuestas, el marco normativo también ha experimentado un proceso de adaptación progresiva para dar soporte a la digitalización del sistema judicial. Estas disposiciones legales y reglamentarias se examinarán en detalle en el apartado específico dedicado a la digitalización del sistema de justicia penal (pág: 18)

no implica únicamente la incorporación de nuevas herramientas tecnológicas, sino que supone una reconfiguración profunda de la forma en que se administran y gestionan los procesos y tareas propias del sistema de justicia penal.

Podemos afirmar que la irrupción del Sistema de Justicia Penal 4.0 en el marco de la Cuarta Revolución Industrial ha supuesto una transformación para el sistema de justicia penal, caracterizada tanto por promesas de mayor eficiencia, transparencia y accesibilidad, como por riesgos vinculados con la integración, la equidad y la protección de derechos fundamentales. Esta doble dimensión obliga a replantear las bases mismas sobre las que se organiza y se proyecta la administración de justicia en la era digital. Ahora bien, para comprender con mayor claridad el alcance de estos cambios y sus implicaciones concretas, resulta necesario detenerse en los pilares conceptuales que sostienen este proceso: la digitalización, la algoritmización y la inteligencia artificial. El siguiente apartado, por tanto, estará dedicado a delimitar y precisar estos conceptos, con el fin de establecer un marco analítico riguroso que permita abordar el objeto de estudio de la presente tesis doctoral.

1.1. Precisiones conceptuales y su diferenciación en un contexto de transformación digital.

Como hemos podido comprobar, la irrupción de lo que para la presente tesis hemos denominado sistema de justicia penal 4.0, representa una transformación sustancial del modelo tradicional de justicia, implicando cambios profundos en su concepción, estructura y funcionamiento. Este nuevo paradigma exige una comprensión rigurosa de los conceptos fundamentales asociados a la revolución tecnológica, tradicionalmente ajenos a este sector, así como de las implicaciones que conlleva su integración en los procesos y funciones del sistema de justicia penal. La creciente incorporación de tecnologías emergentes y la consiguiente proliferación de términos vinculados a la transformación de la justicia han derivado, en ocasiones, en un uso impreciso o ambiguo de la terminología, empleando en múltiples ocasiones términos como sinónimos que realmente no lo son. Esta falta de claridad conceptual dificulta tanto la delimitación del verdadero alcance de cada tecnología como la adecuada valoración de sus implicaciones jurídicas, técnicas y prácticas.

En este escenario, resulta imprescindible establecer una delimitación conceptual precisa y consistente que permita construir un marco teórico sólido, capaz de identificar y diferenciar los diversos elementos que conforman el Sistema de Justicia Penal 4.0. Alcanzar este objetivo implica consensuar definiciones claras que eviten ambigüedades terminológicas y favorezcan un entendimiento común de los conceptos que serán empleados a lo largo de la presente tesis doctoral. Debe advertirse, sin embargo, que dicha delimitación no puede realizarse de manera completamente aislada, puesto que los conceptos en cuestión se encuentran estrechamente interrelacionados. Por esta razón, su análisis se desarrollará de forma progresiva: partiendo de nociones tecnológicas más básicas e introductorias, para avanzar gradualmente hacia aquellas que presentan un mayor nivel de complejidad teórica y técnica.

1.1.1. Conceptualización del término digitalización y su concreción en el sistema de justicia penal.

En este marco, el primer concepto que debe abordarse es el de digitalización, entendido como el proceso de conversión de información analógica en formatos digitales que posibilitan su almacenamiento, procesamiento y análisis mediante tecnologías computacionales (Brennen & Kreiss, 2016). En otras palabras, la digitalización consiste en el uso de hardware y software para recibir y transmitir comunicaciones, procesar textos, organizar archivos y ejecutar operaciones que previamente se realizaban de manera manual. De este modo, la digitalización no solo representa una herramienta de modernización tecnológica, sino también un cambio significativo en la forma en que se gestionan y manipulan los datos.

Ahora bien, conviene precisar que en la literatura académica y profesional se emplean términos que, aunque en apariencia podrían considerarse sinónimos, presentan matices diferenciadores. Así, el concepto *digitization* hace referencia al proceso de digitalización en sentido estricto, esto es, a la mera conversión de lo físico a lo digital. Por otro lado, la noción de *transformación digital* remite a un cambio organizativo profundo, impulsado por las posibilidades que ofrecen las tecnologías digitales, y que excede la simple digitalización (Nazareno, 2023). En este trabajo, sin embargo, se tomará como punto de partida la definición propuesta por Brennen y

Kreiss (2016), centrada en la digitalización como proceso técnico que sirve de base para desarrollos posteriores organizativos. Esta precisión resulta especialmente relevante si se considera que la propia literatura científica ha mostrado una notable falta de consenso en torno a estos términos, Reis, Amorim, Melão, Cohen y Rodrigues (2020), a partir de una revisión sistemática de 121 artículos indexados en Scopus, evidencian que los conceptos de digitalización, *digitization* y transformación digital se emplean de manera indistinta en numerosos estudios, lo que ha generado un considerable grado de ambigüedad. Sus hallazgos subrayan que la digitalización no puede entenderse únicamente como un fenómeno técnico, sino como un proceso transversal que afecta a las dimensiones sociales, económicas y organizativas. Incluir esta perspectiva enriquece el análisis, ya que permite comprender la digitalización como la base de transformaciones más amplias, en las que convergen la innovación tecnológica, la reestructuración institucional y la creación de valor en distintos sectores.

La distinción entre digitalización y transformación digital cobra especial relevancia en el sistema de justicia penal. Mientras la primera se manifiesta en la implementación de sistemas de gestión electrónica de expedientes, bases de datos centralizadas y plataformas de comunicación digital, la segunda implica una reconfiguración más profunda de los procesos y estructuras organizativas, con un impacto directo en la manera en que se conciben y prestan los servicios judiciales. Comprender esta diferencia resulta fundamental para analizar el alcance de las iniciativas emprendidas en el contexto español. En este sentido, cabe señalar que la literatura académica ha subrayado que la digitalización constituye el punto de partida técnico sobre el cual se articulan procesos de transformación organizativa más amplios. Así lo demuestra la revisión de Pereira, Pinheiro de Lima, Gonçalves Machado y Gouvêa da Costa (2020), quienes analizan la convergencia entre Industria 4.0 y manufactura sostenible y evidencian que la digitalización posibilita transformaciones de mayor calado orientadas a la eficiencia y sostenibilidad. Este hallazgo resulta extrapolable al sector judicial, donde la digitalización técnica de expedientes y comunicaciones constituye la base sobre la cual se despliegan procesos de transformación institucional más complejos.

En el caso concreto de España, la digitalización de la Administración de Justicia ha estado fuertemente impulsada por la legislación a lo largo de más de dos décadas. Un primer hito normativo se encuentra en la reforma de 1994 de la Ley Orgánica del Poder Judicial (Ley Orgánica 16/1994, 1994), que, a través de la modificación de su artículo 230, reconoció por primera vez la validez jurídica de la actividad judicial realizada mediante nuevas tecnologías, sentando así las bases para la informatización de los trámites judiciales. Posteriormente, se aprobaron normas específicas destinadas a promover el uso de medios electrónicos; entre ellas, la Ley 18/2011, de 5 de julio (LUTICAJ), que reguló de manera expresa la incorporación de las tecnologías de la información y comunicación en la Administración de Justicia.

En esta misma línea, la aprobación de la Ley 42/2015, de 5 de octubre, de Enjuiciamiento Civil, supuso un paso relevante al introducir la celebración de subastas judiciales electrónicas, así como la posibilidad de establecer comunicaciones telemáticas con la Administración de Justicia. Dichas previsiones se materializaron mediante el Real Decreto 1065/2015, de 27 de noviembre, que reguló las comunicaciones electrónicas en el ámbito territorial dependiente del Ministerio de Justicia y estableció el funcionamiento de LexNET, plataforma segura destinada al envío y recepción telemática de escritos, notificaciones y resoluciones judiciales.

La crisis sanitaria derivada de la COVID-19 aceleró de manera decisiva este proceso, impulsando no solo la digitalización, sino también la transformación digital de la Administración de Justicia y de la sociedad en su conjunto. En este contexto, la Ley 3/2020 introdujo medidas procesales y organizativas orientadas a afrontar las consecuencias de la pandemia, incorporando avances significativos en la regulación de las actuaciones procesales realizadas por medios telemáticos. Más recientemente, esta trayectoria legislativa ha continuado con la aprobación de nuevas normativas vinculadas a la denominada “eficiencia digital” en la justicia, entre ellas la Ley Orgánica 1/2025 (2025), que refuerza la consolidación de un modelo más tecnológico y adaptado a las necesidades actuales.

La digitalización del sistema de justicia penal constituye, además, un prerrequisito indispensable para la posterior implementación de sistemas algorítmicos e

inteligencia artificial. Estas herramientas dependen de la existencia de datos estructurados, interoperables y de calidad que permitan su análisis y procesamiento automatizado. Como señalan Barocas, Hardt y Narayanan (2023), los algoritmos aprenden a partir de conjuntos de datos previamente recopilados y digitalizados, lo que convierte a la digitalización en un cimiento técnico esencial para el desarrollo de modelos predictivos y sistemas de apoyo a la decisión. La ausencia de una infraestructura digital adecuada no solo limitaría el potencial de estas tecnologías, sino que además aumentaría los riesgos de sesgos derivados de registros incompletos o de difícil procesamiento.

Desde una perspectiva institucional, Reiling (2020) destaca que la digitalización también cumple una función cultural, en tanto que prepara a los operadores jurídicos para interactuar con entornos tecnológicos. La adopción progresiva de sistemas como LexNET en España ha contribuido a normalizar el uso de plataformas electrónicas en la gestión judicial, lo que facilita la aceptación posterior de herramientas más avanzadas, incluidas aquellas basadas en inteligencia artificial. Por ello, la digitalización puede entenderse como la fase preliminar y necesaria en la transición hacia la *Justicia 4.0*, caracterizada por la incorporación de sistemas algorítmicos y de inteligencia artificial en la gestión de expedientes, la predicción de riesgos y el apoyo a la toma de decisiones (Sourdin, 2018). El desarrollo de la digitalización en el ámbito judicial español evidencia, en consecuencia, un proceso progresivo sustentado en una base normativa sólida y en la implementación de infraestructuras tecnológicas que han permitido una gestión más ágil y segura de los procedimientos. No obstante, la diferencia conceptual con la transformación digital subraya que aún existen retos vinculados a la reorganización estructural y cultural de la Administración de Justicia, necesarios para aprovechar plenamente las oportunidades que ofrecen las tecnologías digitales.

1.1.2. Conceptualización del término "algoritmo" y sus aplicaciones en el sistema de justicia penal.

En su acepción más general, un algoritmo puede definirse como una secuencia finita y ordenada de instrucciones o reglas bien definidas que, aplicadas de manera secuencial, permiten resolver un problema o ejecutar una tarea de forma automática

(Cormen, Leiserson, Rivest, & Stein, 2009). Este concepto, que integra las propiedades de finitud, precisión y eficacia, constituye la base de todo proceso computacional. El origen moderno de esta noción se encuentra en la obra de Alan Turing, quien en su influyente artículo *On Computable Numbers, with an Application to the Entscheidungsproblem*, publicado en 1936, introdujo la idea de una máquina abstracta, la denominada Máquina de Turing, capaz de ejecutar operaciones lógicas elementales de manera indefinida. Con ello, Turing sentó las bases de la informática teórica y de la concepción contemporánea del cómputo automático (Turing, 1936).

Con el advenimiento de la era digital, el concepto de algoritmo ha experimentado una expansión significativa, trascendiendo la lógica de los programas clásicos para dar cabida a formas más complejas y adaptativas. Mientras que los algoritmos tradicionales se limitaban a ejecutar reglas fijas predefinidas, los desarrollos contemporáneos en aprendizaje automático (*machine learning*) y aprendizaje profundo (*deep learning*) han incorporado una capacidad inédita de ajuste dinámico. Estos modelos, basados en la explotación de grandes volúmenes de datos (*big data*), modifican iterativamente sus parámetros internos para mejorar su capacidad predictiva o de recomendación sin necesidad de intervención humana directa (Domingos, 2015; Jordan & Mitchell, 2015). De este modo, los algoritmos modernos ya no se conciben únicamente como secuencias estáticas de instrucciones, sino como sistemas en constante evolución capaces de modificar sus patrones de funcionamiento en respuesta a la incorporación de nueva información.

La introducción de estas tecnologías en el ámbito judicial constituye uno de los campos más controvertidos y, al mismo tiempo, más significativos de su aplicación. En términos generales, un algoritmo judicial puede definirse como un procedimiento formal, finitamente descrito, que transforma datos relativos a sujetos y hechos en indicadores cuantificables con el propósito de orientar la decisión judicial y estandarizar la comparación entre casos (Angwin, Larson, Mattu, & Kirchner, 2016). Esta definición permite comprender cómo, en el contexto de la justicia, los algoritmos funcionan como mecanismos de formalización y objetivación de la información, aportando criterios que, en principio, buscan reforzar la consistencia y la transparencia de las decisiones adoptadas por los operadores

jurídicos.

La materialización de esta lógica se expresa en dos grandes tipologías de instrumentos algorítmicos (Hannah-Moffat, 2013; Harcourt, 2019). La primera corresponde a los instrumentos actuariales tradicionales, desarrollados desde la década de 1970, los cuales combinan variables de carácter histórico, como la edad al primer delito o el número de condenas previas, con factores de tipo dinámico, entre ellos la situación laboral o los vínculos familiares. Gracias a esta combinación de elementos, dichos instrumentos permiten elaborar informes presentenciales y orientar decisiones relacionadas con la concesión de la libertad condicional. La segunda tipología está constituida por los algoritmos basados en *big data* y *machine learning*, que integran registros judiciales, policiales y administrativos, y reajustan de manera autónoma los pesos otorgados a cada variable a medida que incorporan nueva información. La diferencia entre ambas categorías resulta fundamental, ya que los algoritmos actuariales tradicionales se apoyan en reglas explícitas y en metodologías propias de las ciencias sociales, mientras que los modelos fundamentados en *big data* se caracterizan por una naturaleza opaca, a menudo descrita como "caja negra". Además, presentan un elevado grado de variabilidad y un tratamiento masivo y heterogéneo de datos que, en muchos casos, carece de las garantías metodológicas suficientes (Hannah-Moffat, 2013).

En la práctica profesional, estos algoritmos encuentran múltiples aplicaciones, siendo una de las más frecuentes el uso de instrumentos actuariales de evaluación de riesgo en distintas fases del sistema de justicia penal, como la libertad provisional o la estimación de la probabilidad de reincidencia. Estos instrumentos convierten datos relativos a personas y hechos en indicadores cuantificables, como puntajes o categorías de riesgo, que pretenden orientar la decisión judicial y homogeneizar la comparación entre casos (Kleinberg et al., 2018; Monahan & Skeem, 2016; Brennan, Dieterich & Ehret , 2009; DeMichele et al., 2020; Brittain, Georges, & Martin, 2021).

La literatura informática clásica refuerza esta perspectiva[3] al señalar que los algoritmos son procedimientos formales, finitos y bien especificados, cuyo propósito es transformar entradas en salidas de manera sistemática (Knuth, 1968; Cormen et al., 2009). Trasladado al ámbito de la justicia penal, este procedimiento se manifiesta en la combinación mecánica de variables observadas, como el historial delictivo, la edad de la persona o determinadas características procesales, con el fin de generar puntajes que orienten las decisiones judiciales. La investigación empírica ha puesto de relieve que los métodos algorítmicos tienden a producir comparaciones más consistentes y estandarizadas que aquellas derivadas exclusivamente de la intuición profesional (Grove et al., 2000; Zeng, Ustun, & Rudin, 2016). A ello se suma la evidencia que respalda tanto la amplia implantación de estas herramientas como su rendimiento predictivo, el cual suele oscilar entre niveles bajos y moderados en función del contexto en que se aplican y del tipo de instrumento utilizado. En consecuencia, lo que se observa no es una sustitución de la decisión judicial por parte de los algoritmos, sino la provisión de medidas estandarizadas que contribuyen a fundamentarla y a dotarla de mayor coherencia (Dressel & Farid, 2018; Fazel et al., 2022; Andrews & Bonta, 2010).

Esta comprensión permite advertir que, pese a las limitaciones inherentes a sus modelos predictivos, la definición formal de algoritmo y la literatura sobre evaluación actuarial coinciden en describir procedimientos explícitos, finitos y verificables que convierten datos relativos a sujetos y hechos en indicadores cuantificables. El propósito central de estos procedimientos es orientar las decisiones judiciales y facilitar la comparación entre casos, reforzando así la coherencia y la trazabilidad de la justicia penal (Kleinberg , Lakkaraju, Leskovec, Ludwig & Mullainathan, 2018; Monahan & Skeem, 2016).

La aplicación de algoritmos en la justicia no se restringe al ámbito penal, sino que se

[3] En este contexto, se refiere a la concepción previamente expuesta de los algoritmos en el ámbito judicial, entendidos como procedimientos formales que transforman datos relativos a sujetos y hechos en indicadores cuantificables destinados a orientar la decisión judicial y a estandarizar la comparación entre casos (Angwin, Larson, Mattu, & Kirchner, 2016).

extiende prácticamente a todas las ramas del sistema judicial y administrativo. En la jurisdicción penal, por ejemplo, destacan herramientas como COMPAS o la *Public Safety Assessment* (PSA) en Estados Unidos, que generan puntuaciones de riesgo utilizadas para orientar decisiones relacionadas con fianzas, libertad condicional o la determinación de la cuantía de la pena. En el ámbito policial, los modelos de predicción delictiva (*predictive policing*), como PredPol, HunchLab o el sistema Gotham de Palantir, procesan series de datos georreferenciados sobre delitos para planificar patrullajes y despliegues estratégicos (Angwin et al., 2016). A nivel penitenciario, sistemas como RisCanvi en Cataluña, OASys en Inglaterra y Gales o el Level of Service Inventory-Revised (LSI-R) se utilizan para determinar el régimen de vida y los programas de tratamiento asignados a la población reclusa. En la fase de ejecución de la pena, los sistemas de monitoreo electrónico emplean algoritmos basados en geocercas dinámicas, que ajustan automáticamente los radios de control en función de la localización histórica del penado.

Más allá del ámbito penal, la Administración pública y la protección social han recurrido también a sistemas algorítmicos de manera intensiva. Un ejemplo paradigmático es el sistema SyRI en Países Bajos, diseñado para detectar riesgos de fraude en los subsidios sociales, o los módulos de análisis de fraude de la Seguridad Social española, que aplican técnicas de detección de anomalías para identificar posibles irregularidades. En la jurisdicción contencioso-administrativa, destacan algoritmos de reparto de expedientes, como el prototipo SOLON del Consejo General del Poder Judicial en España o el sistema desarrollado por el Conseil d'État francés, cuya implementación ha suscitado debates sobre la transparencia en la distribución de cargas de trabajo. El abanico de aplicaciones expuesto pone de relieve la ubicuidad del cálculo algorítmico en la justicia contemporánea y plantea la necesidad de establecer marcos normativos que aseguren la trazabilidad, la auditabilidad y la neutralidad de estas herramientas (Mittelstadt, Russell, & Wachter, 2016; Završnik, 2021). En efecto, la progresiva incorporación de algoritmos en todas las dimensiones del sistema judicial confirma que nos encontramos ante un cambio estructural en la forma en que se toman decisiones legales, lo cual exige un análisis crítico de sus fundamentos, alcances y limitaciones.

1.1.3. Conceptualización del término inteligencia artificial: de la formulación teórica a las definiciones de los marcos regulatorios contemporáneos.

La definición del término inteligencia artificial ha experimentado una profunda evolución desde su formulación inicial a mediados del siglo XX hasta las definiciones normativas y técnicas que predominan en la actualidad, y, aunque se han propuesto múltiples enfoques, no existe aún una definición generalmente aceptada del concepto (Russell & Norvig, 2021). El término fue introducido por primera vez por John McCarthy, un matemático e informático estadounidense, en la propuesta para la Conferencia de Dartmouth (Dartmouth Summer Research Project on Artificial Intelligence), celebrada en 1956 en el Dartmouth College (Hanover, New Hampshire). Dicha propuesta, elaborada junto con Marvin Minsky, Nathaniel Rochester y Claude Shannon, partía de la premisa de que la inteligencia humana podía describirse con tal precisión que sería posible su simulación mediante máquinas, en concreto indican que "El estudio debe proceder sobre la base de la conjetura de que cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, describirse con tal precisión que pueda construirse una máquina capaz de simularlo" (McCarthy, Minsky, Rochester, & Shannon, 1955, p. 1).[4] Este planteamiento no ofrecía una definición cerrada, sino que configuraba un programa de investigación interdisciplinar para explorar la reproducción artificial de funciones cognitivas humanas, tales como el razonamiento, el aprendizaje y la resolución de problemas.

En las décadas posteriores, el concepto fue objeto de reformulaciones que reflejaban tanto el avance técnico como la diversificación de enfoques teóricos. Durante la década de 1980, Elaine Rich (1983) definió la IA como el proceso de *"hacer que las máquinas realicen tareas que, si las hiciera un humano, se considerarían inteligentes"*, adoptando un criterio funcional centrado en la imitación de habilidades humanas.

[4] Texto original: "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy, Minsky, Rochester, & Shannon, 1955, p. 1).

Posteriormente, Russell y Norvig (1995) introdujeron un enfoque más formal al conceptualizar la IA como el estudio de agentes racionales, definidos como sistemas que perciben su entorno y actúan sobre él con el objetivo de maximizar sus posibilidades de éxito, independientemente de si su comportamiento imita o no el humano.

En la investigación contemporánea sobre IA, una corriente relevante ha buscado comprender cómo replicar las capacidades cognitivas humanas, apoyándose en disciplinas como la psicología cognitiva y la neurociencia. Según Nilsson (2009), la IA debe entenderse no solo como un proceso de emulación de dichas capacidades, sino también como un sistema en constante evolución que interactúa con datos y entornos complejos. Esta visión ha facilitado el desarrollo de algoritmos de *machine learning* y *deep learning*, que permiten a las máquinas aprender y adaptarse de forma autónoma a nuevas situaciones sin intervención humana directa. Desde un punto de vista técnico, la inteligencia artificial se considera una subdisciplina de la informática dedicada al desarrollo de sistemas capaces de ejecutar tareas que, hasta hace poco, solo podían ser realizadas por personas (Russell & Norvig, 1995). Pero lejos de ser una entidad única, la IA constituye un campo multidisciplinario que abarca desde la imitación de comportamientos humanos hasta la construcción de sistemas autónomos cuyas habilidades pueden, en ocasiones, superar las capacidades humanas (Boden, 2016).

Más recientemente, en un intento de unificar las diferentes definiciones de la inteligencia artificial, Sheikh, Prins y Schrijvers (2023) integraron y reformularon diversos planteamientos previos, tomando como referencia la definición propuesta por el High-Level Expert Group on Artificial Intelligence de la Comisión Europea (2019). Esta formulación describe la IA como "sistemas que muestran un comportamiento inteligente al analizar su entorno y tomar acciones con cierto grado de autonomía para alcanzar objetivos específicos" (p. 16). El valor de esta definición radica en su capacidad para recoger los elementos centrales presentes en conceptualizaciones anteriores, como la actuación adecuada al contexto, la orientación hacia objetivos, la capacidad de aprendizaje y la autonomía relativa, y al mismo tiempo mantener un alcance lo suficientemente amplio para abarcar

desarrollos actuales y futuros sin restringirse a un dominio técnico concreto. Entre estas conceptualizaciones previas se encuentra la de Nilsson (2009), quien la define como un sistema que "funciona de manera adecuada y con previsión en su entorno" (p. 16), destacando así la importancia de la capacidad de actuar de forma correcta y anticipada. También DenkWerk (2018) incorpora en su definición la idea de que la IA supone "la capacidad de percibir, de perseguir objetivos, de iniciar acciones y de aprender a partir de un bucle de retroalimentación" (p. 16). En términos muy cercanos, el propio AI HLEG (2019) formula la IA como un sistema que combina el análisis del entorno, la autonomía relativa y la orientación a fines específicos, rasgos que más tarde retoman Sheikh et al. (2023) en su propuesta de integración. Esta convergencia permite disponer de un concepto operativo de IA que la distingue de los algoritmos en general y de la digitalización en sentido amplio, asegurando su utilidad para el estudio de su inserción social y su tratamiento regulatorio.

En el ámbito jurídico, el avance de nuevas herramientas tecnológicas aplicadas al contexto judicial ha abierto oportunidades, pero también ha generado importantes retos éticos y regulatorios. Estos desafíos no solo pueden abordarse desde una perspectiva normativa tradicional, sino que requieren definiciones claras y consensuadas que orienten su desarrollo y uso. En este sentido, en 2019 la Organización para la Cooperación y el Desarrollo Económicos (en adelante, OECD, por sus siglas en inglés) propuso una definición que describe los sistemas de inteligencia artificial como:

> *"*Un sistema de inteligencia artificial es un sistema basado en máquinas que, para objetivos explícitos o implícitos, infiere, a partir de la información que recibe, cómo generar resultados como predicciones, contenidos, recomendaciones o decisiones que pueden influir en entornos físicos o virtuales. Diferentes sistemas de IA varían en sus niveles de autonomía y capacidad de adaptación tras su implementación.*"* (OECD, 2019, p.7).[5]

[5] Texto original: *An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or*

Esta definición de la OECD incorpora elementos que amplían la concepción tradicional de los sistemas de inteligencia artificial. En primer lugar, reconoce que estos sistemas pueden presentar niveles variables de autonomía, lo que implica que su capacidad de operar sin intervención humana puede oscilar desde una dependencia total de instrucciones externas hasta una actuación altamente independiente tras su despliegue. En segundo lugar, subraya que los resultados generados por la inteligencia artificial ya sean predicciones, contenidos, recomendaciones o decisiones, tienen la capacidad de influir tanto en entornos físicos como virtuales, lo que refleja la diversidad de ámbitos en los que estas tecnologías pueden operar, desde la robótica hasta los servicios digitales. No obstante, mantiene como elemento central la referencia a que los objetivos de funcionamiento son definidos por humanos, explícita o implícitamente, lo que preserva la noción de que la inteligencia artificial actúa en última instancia en función de fines establecidos por sus desarrolladores o usuarios.

Más recientemente, el Reglamento Europeo de Inteligencia Artificial (AI Act), aprobado en 2024, define la IA como:

> "Un sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras el despliegue, y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales" (Parlamento Europeo & Consejo de la UE, 2024, p.46).

La definición del *AI Act* introduce dos elementos diferenciadores respecto de formulaciones previas de inteligencia artificial: en primer lugar, la capacidad de adaptación tras el despliegue y, en segundo lugar, la inferencia algorítmica como

---

decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment. (OECD, 2019, p. 7).

mecanismo para generar salidas (predicciones, contenidos, recomendaciones o decisiones) a partir de datos de entrada**.** La primera alude a que ciertos sistemas pueden modificar su funcionamiento mediante procesos de autoaprendizaje una vez operativos, mientras que la segunda destaca la facultad de deducir modelos o algoritmos, o ambos, a partir de la información recibida, superando el mero procesamiento de datos. Ambos rasgos responden a la necesidad de regular tecnologías que no son estáticas, sino que evolucionan y aprenden en contextos reales, lo que plantea retos adicionales.

Como se ha podido observar, las definiciones de inteligencia artificial son múltiples y han evolucionado desde aproximaciones iniciales centradas en la emulación de capacidades humanas (McCarthy et al., 1955; Rich, 1983) hasta formulaciones más recientes que incorporan criterios funcionales, técnicos y normativos (Russell & Norvig, 2021; OECD, 2019; Parlamento Europeo & Consejo de la UE, 2024). Esta diversidad refleja la naturaleza multidisciplinar del campo y la necesidad de contar con un concepto operativo que, sin perder rigor, resulte lo suficientemente amplio para abarcar desarrollos presentes y futuros.

En el contexto de la presente investigación, se entenderá por inteligencia artificial a un sistema basado en máquinas, diseñado para operar con distintos niveles de autonomía y, potencialmente, con capacidad de adaptación tras su despliegue, que para objetivos explícitos o implícitos (definidos por desarrolladores, reguladores o usuarios) infiere, a partir de datos de entrada, resultados tales como predicciones, contenidos, recomendaciones o decisiones, capaces de influir en entornos físicos o virtuales (Parlamento Europeo & Consejo de la UE, 2024; OECD, 2019). Esta definición integra elementos procedentes de las conceptualizaciones técnicas, funcionales y normativas contemporáneas, incorporando la orientación a objetivos, la actuación adecuada al contexto, el aprendizaje autónomo y la inferencia algorítmica como rasgos esenciales. En esta definición, el factor humano no solo actúa como origen, estableciendo los fines, diseñando los modelos y seleccionando los datos, sino también como elemento de control y legitimación, dado que la supervisión, interpretación y aplicación de los resultados generados por la IA dependen en última instancia de la intervención y el juicio profesional de las

personas.

Ahora que ya se ha establecido la definición de inteligencia artificial, resulta fundamental comprender que, derivado de este concepto, existen múltiples herramientas y metodologías diseñadas para responder a finalidades específicas y con distintos niveles de complejidad técnica. Para la presente tesis, no constituye un objetivo analizar exhaustivamente la totalidad de metodologías existentes en el campo de la inteligencia artificial, ni centrar el enfoque en la descripción pormenorizada de cada una de ellas. Más bien, se parte de la premisa de que el interés de esta investigación no radica en el desarrollo interno de los algoritmos, sino en los efectos que generan en ámbitos como el judicial y penitenciario. No obstante, la inclusión de conceptos como *machine learning*, *deep learning* o *procesamiento del lenguaje natural* resulta pertinente, en tanto que permite dimensionar el abanico de enfoques técnicos que integran este campo y, con ello, hacerse una idea más clara del alcance que puede tener la inteligencia artificial en la transformación de los sistemas judiciales y penitenciarios. En este sentido, aunque la investigación no se centra en los aspectos más especializados del diseño algorítmico, el reconocimiento de estas metodologías proporciona un marco de referencia útil para comprender cómo tecnologías desarrolladas en entornos informáticos o comerciales se trasladan progresivamente al ámbito jurídico (Russell & Norvig, 2021; LeCun, Bengio, & Hinton, 2015).

El aprendizaje automático (*machine learning*, ML) constituye una de las subdisciplinas centrales de la inteligencia artificial y se define como la capacidad de los sistemas computacionales para mejorar su desempeño en una tarea a partir de la experiencia, sin necesidad de ser programados de manera explícita (Mitchell, 1997). Dentro de este campo se distinguen dos aproximaciones principales: por un lado, el aprendizaje supervisado utiliza datos previamente etiquetados con el fin de entrenar modelos que posteriormente sean capaces de realizar predicciones o clasificaciones sobre nuevos casos, y por otro lado, el aprendizaje no supervisado se orienta a descubrir patrones y estructuras ocultas en conjuntos de datos no etiquetados, lo que permite identificar regularidades sin necesidad de información previa (Russell & Norvig, 2021). A partir de esta metodología se desarrolla el

aprendizaje profundo (*deep learning*, DL), concebido como una rama del machine learning que se distingue por la utilización de redes neuronales artificiales conformadas por múltiples capas interconectadas. Estas arquitecturas posibilitan el procesamiento de grandes volúmenes de información y la construcción de representaciones jerárquicas de los datos, lo que amplía significativamente la capacidad de los modelos para identificar relaciones complejas. El aprendizaje profundo ha demostrado una eficacia sobresaliente en tareas como el reconocimiento de imágenes, el procesamiento del lenguaje natural y la generación de predicciones en escenarios caracterizados por elevados niveles de incertidumbre (LeCun, Bengio, & Hinton, 2015).

En lo relativo a la aplicación de la inteligencia artificial en la justicia, uno de los campos que puede resultar prometedor es el procesamiento del lenguaje natural (*natural language processing*, NLP), que se centra en el estudio de la interacción entre los ordenadores y el lenguaje humano. Esta disciplina abarca aplicaciones como la traducción automática, el análisis de sentimientos y, en el ámbito jurídico, la automatización de tareas relacionadas con la redacción de contratos, la clasificación de sentencias o la extracción de información de grandes volúmenes de documentos legales (Ashley, 2017). Otro desarrollo relevante lo constituyen los sistemas de apoyo a la decisión (decision support systems, DSS), entendidos como herramientas interactivas que combinan información, modelos analíticos y técnicas de inteligencia artificial con el propósito de asistir a los profesionales en procesos de decisión complejos. Estos sistemas pueden apoyar a jueces y abogados en la evaluación de pruebas, la identificación de precedentes relevantes o la selección de estrategias procesales adecuadas). Asimismo, la inteligencia artificial ha impulsado el desarrollo de estrategias de predicción delictiva (*predictive policing*), que utilizan modelos estadísticos y técnicas de análisis de datos para anticipar riesgos criminales y asignar de manera más eficiente los recursos policiales. Del mismo modo, el ámbito jurídico se ha beneficiado del desarrollo de la jurimetría o *legal analytics*, que consiste en la aplicación de modelos estadísticos y algoritmos de inteligencia artificial para analizar de forma masiva resoluciones judiciales. Estas herramientas permiten identificar patrones en la toma de decisiones, calcular probabilidades de éxito en litigios e incluso optimizar las estrategias procesales de los profesionales

del derecho (Aletras, Tsarapatsanis, Preoţiuc-Pietro, & Lampos, 2016). Finalmente, se han consolidado mecanismos de resolución de disputas en línea (online dispute resolution, ODR), que utilizan plataformas digitales para gestionar y resolver conflictos de manera no presencial. Estos sistemas suelen basarse en algoritmos de negociación y mediación asistida y han mostrado especial utilidad en controversias relacionadas con el comercio electrónico y la protección de consumidores, constituyendo un ejemplo claro de justicia digital orientada a la accesibilidad y eficiencia (Katsh & Rabinovich-Einy, 2017).

1.1.4. Recapitulación y delimitación conceptual de la digitalización, los algoritmos y la inteligencia artificial como nuevas tecnologías en el sistema de justicia penal.

A modo de recapitulación, y con el propósito de ofrecer una aclaración final sobre el marco conceptual de esta investigación, concretamos el sentido en que utilizamos los tres términos centrales. En primer lugar, entendemos la digitalización como el proceso técnico de conversión de información analógica en formatos digitales que permiten su almacenamiento, tratamiento y análisis mediante tecnologías computacionales (Brennen & Kreiss, 2016). Este concepto se diferencia de la *digitization*, entendida como la mera conversión de lo físico a lo digital, y de la transformación digital, que alude a cambios organizativos y culturales de mayor alcance (Nazareno, 2023; Reis et al., 2020). En nuestra tesis, empleamos el término digitalización en este sentido técnico, como base imprescindible sobre la cual se articulan procesos de modernización institucional en la Administración de Justicia.

En segundo lugar, definimos algoritmo como un procedimiento formal, finito y verificable que transforma datos en resultados destinados a orientar decisiones (Cormen et al., 2009). En el ámbito judicial, consideramos tanto los modelos actuariales tradicionales como los desarrollos contemporáneos basados en *machine learning* y *big data*. En nuestra investigación, cuando hablamos de algoritmos nos referimos a este conjunto de procedimientos aplicados en contextos judiciales y penitenciarios con finalidades de evaluación, predicción o asignación, reconociendo su capacidad para aportar coherencia y estandarización, aunque también sus limitaciones en términos de opacidad y sesgo (Angwin et al., 2016; Hannah-Moffat,

2013).

Por último, asumimos la inteligencia artificial como sistemas basados en máquinas capaces de operar con distintos niveles de autonomía y adaptación, diseñados para inferir a partir de datos de entrada resultados tales como predicciones, recomendaciones o decisiones con incidencia en entornos físicos o virtuales (OECD, 2019; Parlamento Europeo & Consejo de la UE, 2024). Esta definición abarca metodologías como el *machine learning*, el *deep learning* o el procesamiento del lenguaje natural, que permiten dimensionar el alcance transformador de la IA en la justicia. En esta tesis, el término inteligencia artificial se utiliza para designar aquellas tecnologías que superan la lógica algorítmica clásica al incorporar capacidad de aprendizaje y autonomía relativa.

En definitiva, cuando hablamos de nuevas tecnologías o de transformación tecnológica lo hacemos en un sentido amplio y genérico que integra estos tres conceptos, y de manera específica, entendemos digitalización como la base técnica, los algoritmos como procedimientos de formalización y orientación de decisiones, y la inteligencia artificial como un estadio más avanzado que introduce autonomía y aprendizaje, todas ellas ofreciendo tanto oportunidades de eficiencia como desafíos éticos y normativos.

2. La necesidad de la transformación tecnológica en el sistema de justicia penal.

La transformación digital del sistema de justicia penal no debe entenderse como un fenómeno accesorio ni como un proceso meramente técnico de modernización. Se trata, más bien, de una necesidad estratégica que responde a la obligación de garantizar el acceso efectivo a la justicia, proteger los derechos fundamentales y reforzar la confianza ciudadana en las instituciones judiciales (Cordella, 2020). En un contexto global en el que los servicios públicos se digitalizan con rapidez y en el que la ciudadanía demanda una administración más eficiente, accesible y transparente, la justicia no puede permanecer al margen de estas dinámicas. De hacerlo, corre el riesgo de profundizar la brecha entre los tribunales y la sociedad, comprometiendo su legitimidad democrática (Tyler, 2006). La digitalización, en consecuencia, no se limita a la implementación de nuevas herramientas tecnológicas, sino que implica repensar de manera integral la organización y el funcionamiento del sistema de justicia penal, con el objetivo de superar déficits históricos y situar al Estado de derecho en el centro de la vida democrática.

En el caso español, los déficits estructurales del sistema judicial han sido reiteradamente denunciados en informes nacionales e internacionales (Consejo General de la Abogacía Española, 2024; European Commission, Directorate-General for Justice and Consumers, 2025; Fundación Aranzadi La Ley, 2024), lo que ha generado un consenso sobre la urgencia de reformas profundas. Problemas como la lentitud de los procedimientos, la excesiva burocratización, la sobrecarga de trabajo en los juzgados y la insuficiencia de recursos materiales y humanos han configurado un escenario que compromete el derecho fundamental a un proceso sin dilaciones indebidas. Ante esta situación, la transformación tecnológica se presenta como una herramienta para mejorar la eficiencia procesal, reducir las cargas burocráticas y garantizar una justicia más cercana, comprensible y útil para la ciudadanía.

Ahora bien, los factores que explican la necesidad de transformación digital en el sistema penal son múltiples, pero podrías articularse en tres grandes ejes: la crisis de confianza ciudadana en las instituciones judiciales, las limitaciones cognitivas del juicio humano y las deficiencias estructurales que afectan a la organización de la

justicia. Estos tres elementos, interconectados entre sí, permiten comprender por qué la digitalización no constituye un lujo tecnológico, sino un recurso imprescindible para garantizar una justicia más legítima, imparcial y eficiente.

El primer factor que motiva la digitalización del sistema de justicia penal es la falta de confianza ciudadana en la institución judicial. La confianza pública es un componente central de la legitimidad democrática, en tanto que influye en la disposición de los ciudadanos a aceptar y acatar las resoluciones judiciales, incluso cuando estas resultan desfavorables (Tyler, 2006). En el caso español, esta confianza se ha visto erosionada en los últimos años por diversos fenómenos, entre ellos la percepción de politización de la justicia, la excesiva dependencia de los ritmos mediáticos y la proliferación de filtraciones de sumarios y grabaciones que convierten procesos judiciales en espectáculos públicos (Gil Robles, 2025). Estos episodios han debilitado la credibilidad de la justicia y han reforzado la idea de que los tribunales no actúan de manera independiente ni imparcial. La digitalización puede contribuir a revertir esta tendencia mediante la implementación de sistemas de gestión de la información más seguros, el fortalecimiento de la transparencia procesal y la reducción de la opacidad en la toma de decisiones judiciales. Herramientas como el expediente judicial electrónico o la publicación en línea de resoluciones, acompañadas de adecuados mecanismos de accesibilidad, representan vías concretas para acercar la justicia a la ciudadanía y recuperar su confianza.

El segundo factor que explica la urgencia de la transformación digital está relacionado con las limitaciones cognitivas y emocionales del juicio humano. La psicología cognitiva ha demostrado que los operadores jurídicos, como cualquier otro individuo, recurren a heurísticos y atajos mentales que, aunque útiles en la vida cotidiana, pueden generar sesgos en contextos de alta complejidad, como el proceso penal (Tversky & Kahneman, 1974; Kahneman, 2011). Estos sesgos influyen en la valoración de pruebas, la estimación de riesgos y la imposición de sanciones, lo que introduce un grado de incertidumbre e inconsistencia en las decisiones judiciales. La digitalización ofrece aquí una oportunidad para reforzar la imparcialidad y la coherencia de las resoluciones a través de sistemas de apoyo a la decisión basados

en datos empíricos, como los algoritmos de predicción de reincidencia o las herramientas de análisis automatizado de pruebas digitales. Aunque estos recursos no deben sustituir el juicio humano, sí pueden complementarlo, ofreciendo información más robusta y reduciendo la influencia de prejuicios individuales en decisiones que tienen un fuerte impacto social (Završnik, 2020).

El tercer factor se refiere a las deficiencias estructurales que afectan al sistema judicial. La justicia penal española presenta una sobrecarga crónica que se traduce en retrasos prolongados, una burocracia excesiva y un uso ineficiente de los recursos disponibles. Estas carencias estructurales responden a un modelo organizativo obsoleto que ha sido incapaz de adaptarse a las demandas sociales y jurídicas de la actualidad. La digitalización, mediante la implementación del expediente judicial electrónico, la interoperabilidad entre operadores y la automatización de trámites procesales, se configura como un instrumento idóneo para aliviar la carga burocrática y garantizar un acceso más equitativo a la justicia. No se trata únicamente de trasladar a formato digital lo que ya existía en papel, sino de repensar de manera integral la organización de los procesos judiciales y de introducir nuevas formas de gestión que optimicen los recursos y mejoren la calidad del servicio público de justicia (Susskind, 2019).

2.1. La desconfianza en la justicia como factor de impulso hacia la transformación tecnológica.

La falta de confianza ciudadana en el poder judicial constituye uno de los principales factores que explican la necesidad de avanzar en su digitalización. La literatura ha mostrado que la legitimidad de los sistemas judiciales no depende exclusivamente de su capacidad coercitiva, sino de la percepción de imparcialidad y justicia procedimental que transmiten a la ciudadanía (Tyler, 2006). Sin esta confianza, el cumplimiento voluntario de las decisiones judiciales se debilita, poniendo en riesgo la cohesión democrática y la eficacia de las instituciones (Gibson & Caldeira, 1995). Por ello, la digitalización se presenta como una vía para revertir estos déficits, ya que puede mejorar la eficiencia procesal, incrementar la transparencia y reforzar la percepción de imparcialidad. No obstante, la investigación reciente advierte que la confianza constituye también una condición indispensable para aceptar

innovaciones tecnológicas en la justicia. Así, la ciudadanía solo percibe como legítimas las herramientas de inteligencia artificial cuando las considera transparentes, explicables y respetuosas con la dignidad de las personas (Binns et al., 2018; Fine, Berthelot & Marsh, 2025; Wirtz et al., 2018). En consecuencia, la digitalización judicial no debe entenderse como un mero proceso técnico, sino como una estrategia institucional orientada a restaurar la confianza pública y garantizar la legitimidad de la justicia en la era digital (Wang & Siau, 2019).

En el caso español, la relación entre ciudadanía y justicia ha estado marcada por un recorrido complejo, en el que se entrelazan momentos de avance institucional con fases de profunda desconfianza. Durante la dictadura franquista (1939-1975), el poder judicial carecía de independencia real y se hallaba subordinado al régimen, lo que deterioró gravemente su legitimidad. La Transición democrática (1975-1978) y la aprobación de la Constitución de 1978 representaron un punto de inflexión, pues establecieron la independencia judicial en el artículo 117 y crearon el Consejo General del Poder Judicial (CGPJ) como órgano de gobierno de jueces y magistrados. Sin embargo, la confianza ciudadana en los primeros años de democracia continuó siendo limitada, en parte porque numerosos jueces procedentes del régimen anterior se mantuvieron en sus cargos y porque la cultura judicial tardó en adaptarse plenamente a los valores democráticos (Toharia, 1975). De hecho, ya entonces distintos autores y estudios sociológicos subrayaron la urgencia de una modernización de la justicia para fortalecer su legitimidad pública (Toharia, 1975; Habermas, 1973).

En la década de 1980, el debate sobre la politización de la justicia cobró fuerza tras la reforma de 1985, que modificó el sistema de elección de los vocales del CGPJ para que fueran designados en su totalidad por las Cortes Generales. Aunque esta medida se concibió como un mecanismo de democratización, fue interpretada por algunos sectores como el inicio de la partidización del gobierno judicial (Maravall & Przeworski, 2003). Aun así, los años ochenta y noventa también fueron testigos de avances significativos, como la aprobación de la Ley Orgánica del Poder Judicial (1985) y la celebración de macrojuicios contra tramas de terrorismo de Estado y de corrupción política, que reforzaron el papel institucional de la justicia.

No obstante, hacia finales de los años noventa y comienzos de los 2000, la imagen pública de la justicia seguía siendo deficitaria. Carmena (1997) describió la administración judicial como un sistema en "desorden" que debía reinventarse para responder a las exigencias democráticas, mientras que Nieto (2004) denunció un "desgobierno judicial" que evidenciaba fallas estructurales y una brecha entre la justicia formal y la percibida por la ciudadanía. Si bien se impulsaron reformas como el Pacto de Estado para la Reforma de la Justicia (2001), orientadas a agilizar procedimientos y modernizar juzgados, los efectos sobre la confianza ciudadana resultaron limitados. A ello se sumó el impacto de la crisis económica de 2008 y la proliferación de escándalos de corrupción política, que deterioraron aún más la imagen institucional del poder judicial. De acuerdo con Martínez i Coma y Sanz-Labrador (2009), más de la mitad de los españoles declaraba tener poca o ninguna confianza en la justicia, el 85% consideraba que el trato judicial variaba en función de la persona y alrededor del 40% opinaba que el sistema funcionaba mal o muy mal. Estas cifras confirmaban que la aceptación social de la justicia dependía menos de factores ideológicos que de su desempeño efectivo en términos de rapidez, equidad e integridad, lo cual resulta especialmente relevante en un poder cuyos miembros no son elegidos directamente por la ciudadanía. Así mismo, la percepción social del sistema judicial se caracteriza por bajos niveles de confianza vinculados a la ineficiencia, la politización y la desigualdad en el trato, lo que limita de manera estructural su legitimidad social (Martínez i Coma & Sanz-Labrador, 2009).

Los acontecimientos más recientes han contribuido a profundizar esta crisis de legitimidad que atraviesa el poder judicial. En este sentido, la prolongada parálisis en la renovación del CGPJ desde 2018, derivada de la falta de acuerdos políticos entre los principales partidos, ha sido calificada por expertos como una "crisis de legitimidad" del poder judicial (Pendás, 2020). Esta situación descrita se refleja también en los datos del EU Justice Scoreboard de 2024, según los cuales España se sitúa entre los países de la Unión Europea con peor percepción ciudadana de la independencia judicial: únicamente un 35% de los encuestados considera independiente a la justicia nacional, cifra muy por debajo de la media comunitaria, que supera el 50% (Comisión Europea, 2024). La visión empresarial es incluso más negativa, con apenas un 30% de confianza. Los ciudadanos atribuyen esta

desconfianza principalmente a las interferencias políticas (55%) y a la percepción de que los jueces no actúan con total imparcialidad (47%). A ello se añade la baja eficiencia procesal, pues los procedimientos civiles y comerciales de primera instancia duran en promedio 300 días y pueden alcanzar los 600 con apelaciones, situando a España entre los sistemas judiciales más lentos del continente.

En este contexto histórico de avances institucionales combinados con déficits de legitimidad, la transformación digital del sistema judicial se presenta como una oportunidad y un desafío. Por un lado, puede contribuir a resolver problemas estructurales vinculados a la lentitud procesal, la sobrecarga de los juzgados o la falta de transparencia. Por otro, constituye un instrumento simbólico de modernización que podría favorecer la regeneración institucional. Sin embargo, su implementación no es neutra, ya que su éxito dependerá de cómo se integre en la práctica judicial y de si logra transmitir confianza a la ciudadanía (Minaggia, 2023). La literatura advierte que, en contextos donde la legitimidad institucional es frágil, la introducción de herramientas tecnológicas puede generar escepticismo si no se acompaña de garantías éticas y jurídicas claras. En particular, el uso de algoritmos en el ámbito penal puede perpetuar o incluso reforzar sesgos ya existentes si no se somete a adecuados controles de evaluación, transparencia y explicabilidad (Minaggia, 2023).

La aceptación social de la digitalización de la justicia también depende de factores culturales e institucionales. García-Sánchez, Rodríguez-Ariza y Frías-Aceituno (2013) sostienen que la confianza en sistemas digitales complejos está fuertemente condicionada por la confianza previa en las instituciones y por el nivel de alfabetización institucional y digital de la población. En ausencia de transparencia y comunicación clara, difícilmente se logrará consolidar la confianza en la justicia digital. En esta misma línea, Gutterman (2023) subraya que la confianza en entornos tecnológicos no se deriva únicamente de la eficacia de las herramientas, sino también de la calidad de la comunicación institucional. Aplicado al ámbito judicial, ello significa que la digitalización, desde la implementación del expediente electrónico hasta el desarrollo de sistemas predictivos, debe ir acompañada de pedagogía institucional, rendición de cuentas y salvaguardias que garanticen un

control humano significativo. Por ello, la incorporación de nuevas tecnologías en el sistema judicial español debe entenderse como una estrategia de reconstrucción institucional frente a décadas de erosión de legitimidad. Como señalan Viana y Arranz (2021), cualquier mejora en la eficiencia debe vincularse a un cambio cultural y organizativo que refuerce el papel de la justicia como garante del contrato social. La transformación digital, en este sentido, solo se convertirá en un verdadero vector de legitimación si responde tanto a exigencias de productividad como a demandas más profundas de justicia procedimental, igualdad de trato y control democrático.

2.2. La subjetividad humana en sistema de justicia penal.

El segundo factor que podría estar impulsando la transformación tecnológica en la administración de justicia es la subjetividad humana. Los operadores del sistema de justicia penal, lejos de actuar con absoluta racionalidad y objetividad, han estado históricamente condicionados por diversos elementos, entre los que destacan los sesgos cognitivos, las emociones y los prejuicios sociales vinculados al género, la raza, la clase social o la nacionalidad. Estos factores inciden de manera significativa en las decisiones adoptadas por jueces, fiscales, agentes penitenciarios, policías y demás actores, generando incertidumbre, parcialidad y desigualdad en el acceso a la justicia (Fenoll, 2025). Asimismo, esta subjetividad no opera de manera aislada, sino dentro del marco de la discrecionalidad judicial**,** entendida como el margen de decisión que las normas conceden a los operadores jurídicos para interpretar hechos, valorar pruebas y seleccionar cursos de acción. Diversos estudios han mostrado que la discrecionalidad incorpora necesariamente un componente subjetivo, pues las decisiones no pueden desprenderse por completo de la experiencia, percepciones o intuiciones del decisor (Edlin, 2017). Sin embargo, esta misma subjetividad puede transformarse en arbitrariedad cuando no está sometida a criterios de control, motivación y revisión, permitiendo que los sesgos cognitivos y los prejuicios sociales influyan de manera no reconocida en la resolución de los casos. La literatura en justicia penal advierte que factores como las creencias implícitas, las emociones y los atajos mentales tienden a operar en esa zona gris entre la discrecionalidad legítima y la decisión arbitraria, generando inconsistencias y desigualdad en la práctica judicial (Guthrie, Rachlinski, & Wistrich, 2001; Rachlinski & Wistrich, 2017). De este modo, la subjetividad humana se convierte en el vínculo crítico entre discrecionalidad y sesgo: es la base que hace posible la adaptación flexible de la norma, pero también el mecanismo a través del cual se introducen desviaciones sistemáticas que comprometen la imparcialidad y la racionalidad del sistema de justicia penal. A continuación, se examinan estas dos dimensiones clave de la subjetividad en la justicia penal: (1) el recurso a heurísticos y sesgos cognitivos y (2) las discriminaciones o prejuicios hacia ciertos grupos.

La psicología cognitiva ha permitido el estudio de los procesos mentales implicados

en el conocimiento, la percepción, el aprendizaje, la memoria, el razonamiento y la resolución de problemas, contribuyendo al desarrollo de modelos explicativos sobre la manera en que los individuos procesan datos sensoriales, cómo construyen esquemas de interpretación del entorno y de qué modo aplican estrategias para la toma de decisiones (Anderson, 2015; Eysenck & Keane, 2020).

Dentro de la psicología cognitiva, uno de los aportes más influyentes ha sido la identificación y el estudio de los heurísticos y sesgos cognitivos. Estos conceptos, introducidos en la década de 1970 por Amos Tversky y Daniel Kahneman, buscan explicar cómo los individuos emplean atajos mentales o estrategias cognitivas para simplificar la toma de decisiones en contextos de incertidumbre. Dichos mecanismos permiten reducir la carga de procesamiento y generar respuestas rápidas y, en muchos casos, eficaces para la vida cotidiana. Sin embargo, esta misma simplificación puede dar lugar a errores sistemáticos de juicio, conocidos como sesgos cognitivos, que afectan la precisión y racionalidad de las decisiones (Tversky & Kahneman, 1974; Kahneman, 2011). Los heurísticos cumplen una función adaptativa, ya que de manera general en nuestro contexto cotidiano se suele requerir de decisiones rápidas más que exactitud absoluta. Desde la perspectiva evolutiva, resulta más eficiente contar con mecanismos mentales que ofrezcan estas respuestas rápidas en la mayoría de los casos, aunque no siempre óptimas (Gigerenzer, Todd & ABC Research Group, 2000). En trabajos posteriores, Kahneman (2011) constituye una síntesis fundamental de las investigaciones desarrolladas durante décadas sobre los procesos de juicio y decisión. El autor plantea un modelo de doble sistema de pensamiento: el Sistema 1, caracterizado por su rapidez, intuición y carga emocional, y el Sistema 2, definido por su carácter analítico, lento y consciente. La interacción entre ambos sistemas explica cómo los individuos pueden responder de manera eficaz en la vida cotidiana, aunque a costa de incurrir en errores recurrentes cuando las circunstancias requieren un análisis más profundo. Kahneman (2011) demuestra que, si bien el Sistema 1 permite la adaptación rápida al entorno mediante la activación de heurísticos, este mismo mecanismo es responsable de sesgos que generan desviaciones sistemáticas en la evaluación de riesgos, probabilidades o alternativas. Por su parte, el Sistema 2 actúa como un mecanismo de control más preciso y deliberativo, pero su activación

implica un esfuerzo cognitivo elevado, por lo que no siempre se pone en marcha de manera efectiva.

En cuanto a los heurísticos y sesgos, encontramos que, en su primer trabajo, Tversky y Kahneman (1974) identificaron, a través de una serie de experimentos, tres heurísticos fundamentales: representatividad, disponibilidad y anclaje con ajuste.

El heurístico de representatividad, descrito por Tversky y Kahneman (1974), hace referencia a la tendencia de las personas a evaluar la probabilidad de un evento en función del grado de semejanza que guarda con un prototipo o estereotipo, y uno de los experimentos más conocidos para ilustrarlo consistió en presentar a los participantes descripciones de individuos con ciertos rasgos de personalidad, entre ellos un personaje ficticio denominado "Steve", caracterizado como alguien tímido y reservado, siempre dispuesto a ayudar, pero con escaso interés en las personas o en el mundo real, descrito además como un alma dócil y ordenada, con necesidad de estructura y una marcada pasión por los detalles. A partir de este perfil se solicitó a los sujetos que estimaran la probabilidad de que Steve ejerciera una determinada profesión, seleccionando entre opciones como granjero, vendedor, piloto de avión, bibliotecario o médico, de modo que la cuestión central era establecer cómo se ordenaban dichas ocupaciones de la más a la menos probable. Los resultados evidenciaron el funcionamiento del heurístico de representatividad, ya que la probabilidad atribuida a que Steve fuera bibliotecario dependía principalmente de la correspondencia percibida entre su descripción y el estereotipo socialmente asociado a esa profesión, lo que mostró que los participantes no organizaban las opciones en función de la probabilidad estadística real, sino de la similitud percibida con los estereotipos disponibles. Este hallazgo permite comprender cómo este heurístico, pese a que resulta adaptativo al simplificar el procesamiento de información en condiciones de incertidumbre, también conduce a errores sistemáticos ampliamente confirmados en diversos experimentos. Entre ellos, se encuentra la insensibilidad a las probabilidades base, observada en un estudio en el que se presentaron descripciones de personas supuestamente seleccionadas de grupos con distintas proporciones de ingenieros y abogados, y aun cuando se manipuló explícitamente la proporción real (70/30 frente a 30/70), los

participantes ignoraron dichos datos y fundamentaron sus juicios únicamente en la semejanza de la descripción con el estereotipo profesional. Otro error recurrente es la insensibilidad al tamaño de la muestra, ejemplificada en una investigación sobre las tasas de nacimientos en dos hospitales, uno grande y otro pequeño, donde, a pesar de que la teoría estadística predice que los hospitales pequeños presentan mayor variabilidad en la proporción de nacimientos, la mayoría de los participantes consideró que ambos tenían la misma probabilidad de registrar días con más del 60% de nacimientos masculinos. Asimismo, la insensibilidad a la predictibilidad quedó demostrada en estudios en los que se pedía a los participantes predecir el desempeño futuro de estudiantes o empresas a partir de descripciones breves, resultando que sus juicios eran tan extremos como las valoraciones iniciales derivadas de las descripciones, aun cuando estas ofrecían una capacidad predictiva claramente limitada. A su vez, la llamada ilusión de validez se constató al observar que la confianza en los juicios aumentaba en la medida en que la información coincidía con un estereotipo, incluso cuando dicha información era escasa o irrelevante, fenómeno que puede observarse con especial claridad en contextos aplicados como las entrevistas de selección de personal. Finalmente, la malinterpretación de la regresión a la media se ilustró en situaciones como el entrenamiento de pilotos, donde los instructores concluían erróneamente que la crítica mejoraba el desempeño y los elogios lo empeoraban, cuando en realidad los cambios observados respondían a un fenómeno estadístico inevitable, evidenciando con todos estos experimentos que aunque el heurístico de representatividad cumple una función adaptativa al reducir la complejidad de los juicios, también limita la incorporación de conceptos estadísticos fundamentales y genera sesgos sistemáticos que afectan la precisión del razonamiento humano.

El segundo de los heurísticos descrito por Tversky y Kahneman (1974) fue el de disponibilidad, y se refiere a la tendencia de las personas a estimar la frecuencia o la probabilidad de un evento en función de la facilidad con que ejemplos o instancias relacionadas pueden ser recordadas o imaginadas. Uno de los experimentos que utilizaron para ilustrar el presente heurístico consistió en presentar a las participantes listas compuestas por nombres de hombres y mujeres. Cuando los nombres masculinos pertenecían a personalidades más famosas que los femeninos,

los sujetos tendían a juzgar erróneamente que la lista contenía un mayor número de hombres, pese a que la proporción entre ambos géneros era en realidad equivalente. En otra tarea, se pidió a los sujetos decidir si había más palabras en inglés que comenzaban con la letra "r" o más palabras que contenían "r" en la tercera posición. A pesar de que lo segundo es más frecuente, la mayoría respondió lo primero, dado que es más sencillo recuperar palabras a partir de su primera letra. Entre los sesgos derivados del heurístico de disponibilidad, los autores señalaron la recuperabilidad de instancias, según la cual los eventos más fáciles de recordar parecen más frecuentes de lo que son; la efectividad del conjunto de búsqueda, que influye en la facilidad con que se generan ejemplos; la imaginabilidad, que lleva a sobrestimar la probabilidad de sucesos que pueden visualizarse con claridad, como desastres o accidentes, aun cuando sean raros; y la correlación ilusoria, donde la asociación mental entre dos fenómenos (por ejemplo, sospecha y "ojos extraños" en pruebas proyectivas) genera la falsa impresión de que coocurren con mayor frecuencia de la real.

Finalmente, definieron el heurístico de anclaje y ajuste como la tendencia a emitir estimaciones numéricas tomando como referencia un valor inicial, que actúa como ancla y condiciona el resultado final debido a ajustes insuficientes (Tversky y Kahneman, 1974). Uno de los sesgos más frecuentes es el ajuste insuficiente, es decir, la tendencia a permanecer demasiado cerca del ancla original, incluso cuando este carece de relevancia. En este caso se expuso a los participantes a una rueda de la fortuna que determinaba un número aleatorio y posteriormente pedirles que estimaran el porcentaje de países africanos pertenecientes a la ONU. Los resultados mostraron que quienes recibieron un número bajo en la rueda tendieron a dar estimaciones cercanas al veinticinco por ciento, mientras que quienes recibieron un número alto ofrecieron estimaciones cercanas al cuarenta y cinco por ciento, evidenciando que la cifra inicial influía decisivamente en los juicios aun cuando carecía de relación con el problema planteado. Otro error característico se observa en tareas que implican cálculos numéricos incompletos, este estudio consistió en solicitar a los sujetos calcular en apenas unos segundos el producto factorial de ocho a uno, presentado en orden descendente para un grupo (8 × 7 × 6…) y en orden ascendente para otro (1 × 2 × 3…), los resultaros mostraron que aquellos que se les

había presentado el grupo descendiente presentaban valores más altos que las del grupo ascendente, aunque el resultado correcto era idéntico en ambos casos (40.320). Este hallazgo mostró que los individuos suelen detenerse en los cálculos iniciales y ajustan de manera insuficiente el valor final. Del mismo modo, en el ámbito de la probabilidad se identificaron dos sesgos complementarios. Por un lado, la sobreestimación de eventos conjuntivos, que ocurre cuando la gente cree que es más probable de lo que realmente es que varios sucesos pasen uno tras otro, porque parten de la probabilidad de un solo suceso y no tienen en cuenta lo mucho que baja esa probabilidad al combinar varios. El segundo es la subestimación de eventos disyuntivos, que aparece cuando se piensa que es menos probable de lo real que ocurra al menos uno de varios sucesos posibles, ya que las personas se quedan demasiado cerca de la probabilidad de cada evento por separado y no ajustan hacia arriba al considerar que basta con que se cumpla uno. Finalmente, en los estudios sobre la construcción de intervalos de confianza, como aquellos en los que se pedía a los participantes predecir el valor del índice Dow Jones, se observó que las estimaciones resultaban demasiado estrechas, lo que reflejaba un exceso de confianza en la propia capacidad de predicción y una clara subvaloración de la incertidumbre real.

En estos primeros experimentos realizados por Tversky y Kahneman (1974) demostraron que los heurísticos, si bien permiten economizar esfuerzo cognitivo y facilitan respuestas rápidas en contextos de incertidumbre, conducen a errores recurrentes que comprometen la validez de los juicios probabilísticos. Este trabajo evidenció la naturaleza sistemática de los sesgos cognitivos e inauguró una línea de investigación fundamental en la psicología cognitiva y en las ciencias del comportamiento, con importantes implicaciones para la comprensión de la toma de decisiones humanas. Posteriormente, a partir del desarrollo de esta nueva perspectiva, surgieron investigaciones que profundizaron en la clasificación y análisis de los distintos heurísticos y sesgos que guían la conducta humana. Así, junto a los heurísticos iniciales de representatividad, disponibilidad y anclaje identificados por Tversky y Kahneman, comenzaron a definirse otros sesgos igualmente influyentes en los procesos de juicio y decisión. Entre ellos se encuentran el sesgo de confirmación, que refleja la tendencia a privilegiar la

información que corrobora las creencias previas y a ignorar la evidencia contraria (Nickerson, 1998); el sesgo de grupo o *ingroup bias*, que consiste en favorecer sistemáticamente a los miembros del propio grupo frente a los ajenos, con claras repercusiones en la cohesión social y la discriminación (Tajfel & Turner, 1986); o el sesgo retrospectivo (*hindsight bias*), que alude a la inclinación a percibir los acontecimientos pasados como más previsibles de lo que realmente fueron, reconfigurando la memoria de los sujetos y reforzando una ilusión de control (Fischhoff, 1975).

Ahora bien, la consolidación de esta línea de investigación encontró un nuevo punto de inflexión con la obra *Thinking, Fast and Slow* (*Pensar rápido, pensar despacio*), en ella Kahneman (2011) analiza de manera crítica cómo las personas emiten juicios de probabilidad sin disponer de un conocimiento preciso sobre el concepto estadístico de probabilidad. En sus palabras: "Nos preguntamos cómo la gente puede hacer juicios de probabilidad sin conocer con precisión lo que es la probabilidad. Concluimos que la gente tiene que simplificar de algún modo esta tarea imposible, y nos pusimos a buscar cómo lo hace. Nuestra respuesta fue que cuando se le pide juzgar probabilidades, la gente realmente juzga algo y cree que ha juzgado sobre probabilidad" (p. 133). Esta afirmación refleja la tendencia del Sistema 1 a sustituir preguntas complejas por otras más simples, a través de un mecanismo conocido como sustitución heurística.

En términos prácticos, ello implica que cuando un individuo, incluidos los operadores jurídicos, debe estimar probabilidades en contextos de incertidumbre, a menudo recurre a juicios intuitivos basados en variables accesibles, aunque no necesariamente pertinentes. Así, la tarea de calcular la probabilidad de un evento, como la reincidencia delictiva, se sustituye por una evaluación heurística relacionada, que resulta más fácil de procesar. Kahneman (2011) señala que este proceso no es consciente, sino automático, lo que explica por qué los individuos suelen creer que están evaluando probabilidades de manera objetiva, cuando en realidad están respondiendo a una pregunta distinta y más simple.

Este fenómeno adquiere una relevancia especial en el sistema de justicia penal, ya que el trabajo de jueces, fiscales, agentes penitenciarios y policías implica de manera

constante la valoración de riesgos. En este marco, resulta evidente que, al igual que cualquier otro ser humano, estos profesionales no son inmunes a los heurísticos y sesgos cognitivos que afectan el razonamiento bajo incertidumbre, lo que sugiere que sus decisiones, lejos de apoyarse siempre en cálculos probabilísticos exactos, pueden verse condicionadas por atajos mentales que simplifican la tarea pero que, al mismo tiempo, introducen errores sistemáticos en la valoración de los casos (Kahneman, 2011).

En este sentido, la investigación empírica ha mostrado que el sesgo retrospectivo puede alterar la percepción judicial sobre la previsibilidad de los resultados de un registro policial, aunque no siempre cambie la decisión legal final, lo que refleja un desafío cognitivo en la supervisión de la actuación policial (Rachlinski, Guthrie y Wistrich, 2011), y también que factores extralegales como la demografía, las emociones o las intuiciones inciden en la imparcialidad idealizada del rol judicial (Rachlinski y Wistrich, 2017). Del mismo modo, se ha documentado que los jueces recurren a heurísticos como el anclaje, la disponibilidad o la representatividad, lo cual los hace susceptibles a errores sistemáticos similares a los observados en la población general (Guthrie, Rachlinski y Wistrich, 2001), siendo uno de los ejemplos más estudiados el efecto de anclaje en las sentencias, ya que investigaciones han mostrado que las propuestas de condena del fiscal influyen en la determinación de la pena (Englich y Mussweiler, 2001), que incluso números irrelevantes o generados al azar afectan a jueces experimentados (Englich, Mussweiler y Strack, 2006). Además, Peer y Gamliel (2013) analizan cómo los jueces, pese a su formación y experiencia, no son inmunes a heurísticos como el sesgo de confirmación, el sesgo retrospectivo o la falacia de conjunción, y señalan que incluso al evaluar evidencia inadmisible o al dictar sentencias bajo condiciones de fatiga cognitiva reproducen patrones similares a los de los jurados, lo que demuestra que la racionalidad judicial está limitada por los mismos procesos cognitivos que afectan al resto de la población (Peer y Gamliel, 2013)

A su vez, la neurociencia ha permitido comprender que la toma de decisiones judiciales combina procesos intuitivos y deliberativos, mostrando cómo las emociones, el estrés y la empatía afectan la imparcialidad y cómo la comunicación

no verbal influye en la valoración probatoria, lo que evidencia que el conocimiento de los procesos cerebrales puede ayudar a reducir sesgos y a mejorar la calidad de las sentencias (Dal Santo, 2024). Esta perspectiva se complementa con los aportes de la filosofía y la psicología cognitiva, que han mostrado que los sesgos implícitos atraviesan todo el proceso judicial, desde la investigación policial hasta la decisión final, y que es necesario implementar estrategias de concienciación y programas de debiasing para reforzar la legitimidad racional de la justicia (Arena, Luque y Moreno Cruz, 2021).

En el contexto español también se ha puesto de manifiesto que heurísticos como la representatividad, la disponibilidad o el sesgo de confirmación generan distorsiones en la práctica jurisdiccional, especialmente en el proceso penal, donde la inmediación y la valoración subjetiva del comportamiento de acusados y testigos resultan determinantes (Muñoz Aranguren, 2011). De manera complementaria, Cunliffe (2014) ha mostrado que los jueces tienden a construir narrativas para determinar los hechos en juicios complejos, como los de agresión sexual o homicidio infantil, lo que los hace igualmente vulnerables a prejuicios implícitos y estereotipos que pueden conducir tanto a condenas erróneas como a absoluciones injustificadas. En la misma línea, estudios de archivo en España han revelado que más de la mitad de las sentencias analizadas estaban influidas por un anclaje decisional que condicionaba la motivación legal (Fariña, Arce y Novo, 2002).

En cuanto al segundo factor que afecta a la subjetividad en el sistema de justicia penal encontramos que existen una serie de condicionantes sociodemográficos que inciden de manera significativa en jueces, fiscales, policías y agentes penitenciarios. Lejos de constituir un proceso puramente técnico y neutral, las decisiones judiciales se ven atravesadas por prejuicios implícitos asociados con el sexo, la edad, la etnia, la religión, la orientación sexual y la condición socioeconómica. Tales sesgos no deben entenderse como anomalías individuales sino como expresiones de estructuras sociales más amplias que reproducen desigualdades históricas en el acceso a derechos. De acuerdo con la teoría de la discriminación estructural, la justicia refleja y perpetúa jerarquías sociales preexistentes valiéndose de atajos cognitivos y mecanismos automáticos que condicionan la igualdad procesal

(Williams & Law, 2012; Wofford, 2017).

En relación con la etnia y la raza, la evidencia empírica demuestra que los estereotipos sobre peligrosidad y culpabilidad influyen de manera decisiva en las decisiones judiciales y fiscales. En Kenia, por ejemplo, los jueces mostraron una tendencia entre un tres y un cinco por ciento mayor a conceder apelaciones a acusados de su mismo grupo étnico, lo que refleja un claro favoritismo endogrupal con implicaciones directas para la imparcialidad del proceso (Choi, Harris, & Shen-Bayh, 2022). En Estados Unidos, investigaciones sobre solicitudes de fianza han mostrado que los fiscales tienden a exigir montos más elevados para acusados negros y latinos, lo que incrementa la probabilidad de detención preventiva y condiciona negativamente las posibilidades de defensa, generando un efecto acumulativo en la trayectoria procesal (Concannon & Na, 2023). Estas desigualdades se refuerzan en el ámbito de los delitos de drogas, donde la etiqueta de "ofensor peligroso" impacta de manera desproporcionada en varones negros, en particular en casos vinculados al crack, lo que confirma la persistencia de estereotipos raciales en la imposición de penas (Spohn & Sample, 2013).

La interacción entre raza, edad y género constituye uno de los ejes de mayor severidad en el sistema penal. Jóvenes varones negros han sido identificados como el grupo más castigado, en parte por el estereotipo de que forman parte de una clase peligrosa y porque los jueces consideran que soportan mejor la prisión que otros acusados, lo que configura un diferencial de castigo estructural (Steffensmeier, Ulmer, & Kramer, 1998). Esta constatación se relaciona con la teoría de las "focal concerns", que sostiene que las decisiones judiciales se guían por percepciones de culpabilidad, peligrosidad y costos prácticos, y que en contextos de incertidumbre los jueces construyen un "atajo perceptual" basado en categorías sociales como la raza, el género o la edad (Albonetti, 1991).

El género, por su parte, presenta dinámicas complejas. La presencia de juezas en tribunales colegiados incrementa la probabilidad de que demandas por acoso o discriminación sexual resulten favorables y, además, modifica el comportamiento decisional de sus colegas varones, lo que confirma la importancia de la diversidad para mitigar sesgos institucionales (Peresie, 2005). Sin embargo, la experiencia

judicial no neutraliza del todo los prejuicios, ya que incluso jueces experimentados reproducen ideologías tradicionales de género en casos de custodia o discriminación laboral (Miller, 2018). En el ámbito sucesorio, la noción de "influencia indebida" ha sido aplicada con más frecuencia para invalidar testamentos que favorecen a parejas del mismo sexo o amistades, reforzando normas heteropatriarcales y limitando formas familiares alternativas (Recupero, Hargrave, & McClure, 2015). Asimismo, en procesos migratorios se ha identificado que los jueces varones tienden a aplicar patrones de masculinidad y caballerosidad que derivan en un trato más indulgente hacia las mujeres solicitantes de asilo y más severo hacia varones en condiciones semejantes, lo que muestra cómo la concepción judicial de género condiciona la equidad procesal (Gill, Kagan, & Marouf, 2019).

La indigeneidad y la condición socioeconómica también influyen en la administración de justicia. En Australia se ha observado que, aunque el estatus indígena no siempre se traduce de manera directa en una mayor probabilidad de encarcelamiento, sí genera disparidades cuando se combina con la reincidencia o con el género, situando a los hombres indígenas en una situación de particular vulnerabilidad frente al sistema penal (Bond & Jeffries, 2011). En Estados Unidos, la composición racial de los tribunales de apelación en casos de pena de muerte resulta decisiva, ya que la sola presencia de un juez afroamericano aumenta de manera significativa la probabilidad de conceder alivio a acusados negros, lo que evidencia cómo la diversidad interna puede contrarrestar sesgos estructurales (Kastellec, 2020). La condición socioeconómica refuerza estos patrones: los sistemas de fianza monetaria penalizan de manera indirecta a los acusados pobres, quienes enfrentan mayores probabilidades de detención preventiva y de recibir condenas más severas (Concannon & Na, 2023). Del mismo modo, los instrumentos de evaluación de riesgo utilizados en libertad condicional tienden a asignar niveles más altos de peligrosidad a acusados negros, debido a la correlación de los indicadores con la pobreza y con historiales penales más extensos (Lowder, Desmarais, & Baucom, 2018).

Los factores ideológicos y religiosos también tienen influencia en este contexto, diversos estudios han mostrado que las convicciones religiosas influyen en la

interpretación judicial de derechos fundamentales y en los fallos relativos a la libertad religiosa, mientras que la orientación política de los jueces se relaciona con patrones diferenciados de voto en cortes superiores (Sisk & Heise, 2005; Reyes & Reyes, 2019; Epstein, Landes, & Posner, 2013).

Finalmente, la investigación sobre sesgos implícitos ha cuestionado el mito de la objetividad judicial. Estudios basados en el *Implicit Association Test* han demostrado que la mayoría de las personas, incluidos los jueces, reproducen de manera automática asociaciones negativas hacia grupos históricamente desfavorecidos, incluso cuando se consideran imparciales. Estos sesgos son sistemáticos y robustos, se manifiestan en fases clave del proceso penal y afectan por igual a la policía, los fiscales y los jueces (Gravett, 2017). Desde la psicología social y la neurociencia se ha demostrado que la racionalidad judicial está limitada por los mismos procesos cognitivos que afectan al resto de la población (Arena, Luque, & Moreno Cruz, 2021).

2.3. Las deficiencias estructurales del sistema de justicia penal.

El tercer gran factor que impulsa la necesidad de transformación digital en el sistema de justicia penal se vincula directamente con sus déficits estructurales, pues la sobrecarga de trabajo en los órganos judiciales, la insuficiencia de personal y la persistencia de una burocracia excesiva configuran un escenario de ineficiencia crónica que no solo provoca dilaciones en la tramitación de los procedimientos, sino que también acentúa las desigualdades territoriales y erosiona la confianza ciudadana. Tales déficits, documentados de manera sistemática en informes oficiales tanto nacionales como europeos, que coinciden en señalar la necesidad de una modernización integral de la justicia para garantizar su eficacia y legitimidad (CEPEJ, 2024; Comisión Europea, 2024; Consejo General de la Abogacía Española, 2024; Eurojust & eu-LISA, 2021; HLEG/CEPS, 2021; Poder Judicial de España, 2024).

Uno de los principales problemas detectados en los diferentes informes es la sobrecarga de trabajo en los juzgados, que afecta tanto al rendimiento institucional como a la salud laboral de jueces y magistrados. El Primer informe de evaluación de riesgos psicosociales en la Carrera Judicial, elaborado por el Consejo General del Poder Judicial (CGPJ), reveló que el 84 % de los miembros de la carrera se encuentra en la zona de riesgo "muy elevado" en cuanto a carga de trabajo, porcentaje que se eleva al 88 % en los juzgados unipersonales y alcanza el 94 % en jurisdicciones como la mercantil (Consejo General del Poder Judicial, 2018). Este diagnóstico demuestra que la saturación de expedientes no solo repercute en la calidad de las resoluciones, sino que también constituye un riesgo para la salud laboral de quienes sostienen el sistema (Gil-Monte, López-Vílchez, Llorca-Rubio & Sánchez Piernas, 2016). El problema tiene además una dimensión de género, ya que otro informe del CGPJ identificó que más del 65 % de juezas y magistradas considera que el volumen actual de trabajo, sumado al que supondría asumir cargos de mayor responsabilidad, constituye un obstáculo significativo para el ascenso profesional, lo que se traduce en un freno estructural a la igualdad en la carrera judicial (Consejo General del Poder Judicial, 2025).

La insuficiencia de personal judicial y administrativo agrava este escenario, pues mientras la litigiosidad ha crecido de forma sostenida, los recursos humanos no lo

han hecho en la misma medida. El Informe *La Justicia en España 2024*, elaborado conjuntamente por el Gabinete de Estudios del CGPJ y el Consejo General de Procuradores de España (CGPE), evidencia que entre 2018 y 2023 los asuntos en trámite en los juzgados españoles aumentaron un 52,29 %, mientras que el número de juzgados solo se incrementó en un 3,67 % en ese mismo período, lo que explica que la media de asuntos pendientes por órgano judicial ascendiera a 952 en 2023, con notables diferencias territoriales (CGPJ & CGPE, 2024). En consecuencia, los datos muestran un desfase estructural entre la demanda y la capacidad de respuesta del sistema, que únicamente puede corregirse mediante mecanismos capaces de redistribuir cargas de forma dinámica, algo inalcanzable sin la incorporación de herramientas digitales avanzadas. La estadística oficial refuerza esta idea, el resumen de datos estadísticos por partidos judiciales 2024, publicado por el CGPJ, muestra cómo la carga de trabajo no se distribuye de manera homogénea entre los distintos órganos, produciendo grandes desigualdades territoriales (Consejo General del Poder Judicial, 2025). Estas diferencias acentúan las ineficiencias, ya que determinados juzgados acumulan retrasos crónicos mientras otros mantienen una carga asumible. Sin un sistema digital integrado que permita asignar recursos con criterios objetivos y en tiempo real, estas disfunciones tienden a perpetuarse.

La burocracia excesiva y la rigidez organizativa constituyen otro obstáculo estructural. El diseño institucional de la justicia española, caracterizado por la coexistencia de competencias entre el Ministerio de Justicia, las comunidades autónomas con traspaso competencial, el CGPJ y la fiscalía general, ha generado fragmentación y duplicidades que dificultan la modernización. La falta de homogeneidad en los sistemas de gestión procesal y la dependencia de trámites manuales limitan los beneficios de herramientas digitales como LexNet, cuyo alcance sigue siendo parcial. En este sentido, la Ley Orgánica 1/2025, de 2 de enero, de medidas de eficiencia del servicio público de justicia, reconoce expresamente estas disfunciones, señalando la falta de especialización de los juzgados, la proliferación de órganos con competencias idénticas en cada partido judicial y la desigual distribución de cargas como factores que requieren una reorganización estructural apoyada en la digitalización (Boletín Oficial del Estado, 2025).

En el contexto internacional, los informes europeos confirman la magnitud del problema. El CEPEJ Evaluation Report 2022, elaborado por la Comisión Europea para la Eficiencia de la Justicia, muestra que España presenta indicadores de congestión y tiempos de resolución superiores a la media europea, con un gasto público en justicia por habitante situado por debajo de otros Estados miembros con mayor capacidad de respuesta (CEPEJ, 2024). Estos datos evidencian que la ineficiencia del sistema español no es únicamente una percepción interna, sino una realidad objetivable en el marco europeo, lo que refuerza la urgencia de implementar cambios tecnológicos y organizativos profundos.

El propio Plan Nacional de Estadística Judicial 2021-2024, coordinado por el CGPJ, el Ministerio de Justicia, la Fiscalía General del Estado, las comunidades autónomas competentes y el Instituto Nacional de Estadística, reconoce la necesidad de avanzar hacia sistemas normalizados y accesibles de información judicial, subrayando la importancia de que los datos sean reutilizables y transparentes para favorecer la planificación y la evaluación del sistema (Consejo General del Poder Judicial et al., 2021). Este esfuerzo pone de relieve que incluso la medición de la realidad judicial ha estado marcada por la fragmentación y la ausencia de homogeneidad, lo cual solo puede resolverse plenamente mediante procesos de digitalización integrales.

Es por ello, que la transformación tecnológica de la justicia debe entenderse como una estrategia que no solo moderniza los procedimientos, sino que responde directamente a problemas estructurales documentados en informes oficiales. Al automatizar tareas repetitivas, establecer sistemas de gestión interoperables, implantar mecanismos de asignación dinámica de asuntos y permitir un acceso ciudadano más directo a los expedientes, se pueden reducir los cuellos de botella, mejorar la eficiencia institucional y garantizar la igualdad en el acceso a la justicia. Además, la transición digital ofrece un impacto positivo en el bienestar laboral de jueces y personal administrativo, al reducir los niveles de sobrecarga identificados como un riesgo psicosocial de primer orden. Los déficits estructurales del sistema de justicia penal español han alcanzado un nivel crítico que compromete tanto la efectividad de los derechos fundamentales como la legitimidad del propio Estado de derecho. La evidencia empírica aportada por los informes del CGPJ, de la CEPEJ y del

Plan Nacional de Estadística Judicial demuestra que la congestión, la falta de personal, la burocracia y la desigualdad territorial son realidades persistentes. Frente a ello, la digitalización se configura no como una opción, sino como una necesidad histórica e institucional. Solo a través de una transformación tecnológica acompañada de reformas organizativas y legales será posible garantizar que la justicia penal en España sea eficiente, equitativa y transparente, cumpliendo así su función esencial de protección de los derechos y libertades de la ciudadanía.

Todos estos aspectos, entre los que se encuentran la falta de confianza ciudadana en las instituciones, los sesgos y discriminaciones que atraviesan las decisiones judiciales y las deficiencias estructurales que limitan la igualdad de acceso a la justicia, evidencian que la imparcialidad idealizada del sistema penal se encuentra en cuestión. Y pueden ser precisamente todas estas debilidades las que han incentivado el interés por la incorporación de nuevas herramientas tecnológicas, concebidos como mecanismos capaces de ofrecer mayor consistencia y transparencia en las resoluciones judiciales y de responder a los problemas de legitimidad que afectan al sistema en su conjunto (Aletras et al., 2016).

3. Nuevas tecnologías en el sistema de justicia penal: estado actual y perspectivas futuras.

3.1. Panorama actual del uso de nuevas tecnologías en el sistema de justicia penal: digitalización, algoritmos e inteligencia artificial.

Hasta este punto, el análisis ha recorrido tanto las proyecciones futuras como los antecedentes que han dado forma al proceso de transformación tecnológica. Corresponde ahora situarnos en el presente: en el escenario actual en el que dichas transformaciones se materializan. En este sentido, resulta necesario examinar el panorama de las herramientas disponibles actualmente, tanto en el ámbito nacional como en el contexto internacional, ya que conforman el ecosistema que sostiene los nuevos modelos de gestión del sistema de justicia penal, marcando el tránsito hacia un sistema caracterizado por la integración de sistemas avanzados de información, inteligencia artificial y automatización de procesos.

La digitalización y la inteligencia artificialse han convertido en tecnologías transversales con capacidad para transformar profundamente la justicia penal en todos sus ámbitos. La incorporación de sistemas digitales y algoritmos inteligentes promete agilizar trámites, mejorar la gestión de información y optimizar la toma de decisiones en los procesos penales (Januário, 2023; Završnik, 2020). En los tribunales, por ejemplo, ya se emplean herramientas informáticas para administrar grandes volúmenes de documentación judicial y hasta para predecir resultados judiciales basándose en patrones extraídos de casos previos (Aletras et al., 2016; Januário, 2023). En el ámbito penitenciario, los llamados “smart prisons” integran sensores, big data e IA tanto para reforzar la seguridad de los centros como para personalizar programas de rehabilitación y facilitar la reinserción social de los internos (McKay, 2022; Zivanai & Mahlangu, 2022). Por su parte, las fuerzas policiales están adoptando sistemas de vigilancia inteligente y algoritmos de predicción del delito que identifican zonas y personas con mayor riesgo de involucrarse en conductas criminales, permitiendo una asignación más estratégica de los recursos policiales (Simmler et al., 2022).

De este modo, la inteligencia artificial no puede entenderse únicamente como un

recurso técnico orientado a la automatización de tareas de carácter administrativo, sino que debe concebirse como un verdadero agente de cambio que está redefiniendo las lógicas tradicionales de la seguridad, el enjuiciamiento y el castigo. Este proceso de transformación no se produce de manera homogénea, sino que responde a finalidades específicas que varían en función del ámbito de aplicación. Por ello, el recorrido que aquí se propone se estructura en torno a los tres ámbitos principales del sistema de justicia penal: el policial, el judicial y el penitenciario. En el terreno policial, la atención se centra en aquellas herramientas orientadas a la gestión y prevención del delito, como los sistemas de predicción criminal y la videovigilancia con capacidades de reconocimiento automático, junto con los instrumentos destinados a la investigación y al tratamiento de datos a gran escala. En el ámbito judicial, el foco recae sobre los sistemas de gestión de casos y sobre los modelos predictivos que buscan ofrecer apoyo en la toma de decisiones de jueces y magistrados, particularmente en relación con la emisión de sentencias y la evaluación de riesgos procesales. Por último, en el ámbito penitenciario destacan las aplicaciones vinculadas tanto a la clasificación de internos y al seguimiento electrónico de los mismos como a la gestión integral de los centros, en los que cobran especial relevancia las tecnologías de automatización y de detección temprana de riesgos relacionados con la seguridad institucional.

La finalidad última de este análisis es ofrecer una visión ordenada y comparativa de cómo la digitalización y la inteligencia artificial se despliegan en las distintas fases del ciclo penal, adaptándose a las funciones y necesidades específicas de cada ámbito. De este modo, mientras que en apartados anteriores se ha hecho referencia a las aplicaciones potenciales de estas tecnologías, el presente apartado se orienta a proporcionar una aproximación a su realidad actual. La intención es identificar las herramientas que se encuentran efectivamente en uso dentro del sistema de justicia penal, estableciendo una categorización basada en los ámbitos de aplicación que permita sistematizar el análisis y comprender con mayor precisión el alcance de su implementación.

3.1.1. Ámbito policial.

El ámbito policial se ha convertido en uno de los espacios donde la inteligencia artificial ha alcanzado un mayor desarrollo dentro del sistema de justicia penal actual. Su aplicación no se limita a la modernización administrativa o a la digitalización de procedimientos, sino que supone una transformación estructural en la forma en que se conciben la prevención, la investigación y la gestión de la seguridad pública. La posibilidad de procesar grandes volúmenes de datos procedentes de registros delictivos, sistemas de videovigilancia, sensores urbanos, bases de datos judiciales o redes sociales ha configurado un nuevo paradigma operativo: el de la policía algorítmica o *algorithmic policing* (Ferguson, 2017; Miró Llinares, 2018). En este modelo, la toma de decisiones policiales se apoya en el análisis predictivo y en modelos de aprendizaje automático, que buscan optimizar los recursos humanos, anticipar comportamientos delictivos y aumentar la eficiencia en la persecución penal.

En la literatura criminológica, la inteligencia artificial aplicada al ámbito policial se ha articulado en torno a dos ejes principales: la prevención del delito mediante modelos predictivos y la investigación criminal apoyada en sistemas analíticos y forenses automatizados**.** En ambos casos, los algoritmos pretenden traducir en datos los patrones empíricos de la criminalidad, generando inferencias estadísticas que orienten la actuación policial. Sin embargo, esta innovación tecnológica plantea al mismo tiempo desafíos epistemológicos, éticos y jurídicos, pues la lógica algorítmica, basada en correlaciones y probabilidades, entra en tensión con los principios de legalidad, proporcionalidad y presunción de inocencia que rigen el Estado de derecho.

El modelo de policía predictiva (*predictive policing*) representa uno de los ejemplos más paradigmáticos de la aplicación de la IA en la prevención del delito. Este enfoque parte de la premisa de que los fenómenos delictivos no son aleatorios, sino que presentan regularidades espaciotemporales susceptibles de modelización matemática (Perry, McInnis, Price, Smith & Hollywood, 2013). Mediante el análisis de datos históricos, los sistemas predictivos identifican zonas y franjas horarias con mayor probabilidad de comisión delictiva, orientando la distribución de patrullas y

la vigilancia preventiva. Esta lógica se sustenta en los principios de la criminología ambiental y en las teorías de la oportunidad delictiva, que destacan la interacción entre motivación, ocasión y ausencia de control como factores explicativos del delito (Clarke & Cornish, 2017). No obstante, el desarrollo y aplicación de estas tecnologías han suscitado un amplio debate ético y jurídico. Como señala Miró-Llinares (2020), las actitudes hacia la policía predictiva oscilan entre una visión utópica, que considera que los algoritmos pueden mejorar la objetividad y eficiencia policial, y una visión distópica, que advierte sobre los riesgos de vigilancia masiva, sesgos discriminatorios y erosión de derechos fundamentales. Frente a ambos extremos, el autor propone una postura realista, crítica e informada, que reconoce la ausencia de neutralidad tecnológica y aboga por un uso de la IA orientado por criterios éticos, evidencia empírica y respeto al Estado de derecho.

Uno de los casos más conocidos es PredPol, posteriormente comercializado como *Geolitica*, desarrollado en colaboración entre la Universidad de California en Los Ángeles (UCLA) y el Departamento de Policía de Los Ángeles (LAPD). Inspirado en modelos sismológicos, PredPol aplica algoritmos de procesos puntuales autoexcitantes (*self-exciting point processes*) que permiten identificar áreas urbanas donde el riesgo de delito aumenta tras un suceso inicial, generando "cajas predictivas" que delimitan los cuadrantes de mayor probabilidad delictiva en las horas siguientes (Mohler et al., 2015; Camacho-Collados & Liberatore, 2015). Este modelo operacionaliza hallazgos de la criminología ambiental, como la victimización repetitiva o la proximidad espacial del delito, y los traduce en predicciones probabilísticas que guían la actividad policial. Otros sistemas, como CompStat (*Computer Statistics*), surgido en Nueva York en los años noventa, introdujeron una gestión policial basada en datos cuantitativos y rendición de cuentas. Si bien CompStat no se concibió originalmente como una herramienta de IA, su integración con sistemas de análisis de datos geoespaciales y aprendizaje automático ha permitido su evolución hacia modelos de gestión táctica y predicción en tiempo real (Silverman, 2006). En el ámbito europeo, el proyecto VALCRI (*Visual Analytics for Sense-Making in Criminal Intelligence Analysis*), financiado por la Comisión Europea, constituye un referente en materia de análisis de inteligencia criminal. Esta herramienta combina minería de texto, análisis visual y aprendizaje automático

para asistir a los analistas en la detección de patrones ocultos y correlaciones complejas en grandes bases de datos policiales, incorporando mecanismos de supervisión humana activa (Islam, Anslow, Xu, Wong & Zhang, 2016).

En España, la incorporación de algoritmos predictivos e inteligencia artificial en el ámbito policial ha avanzado de forma más gradual y bajo una supervisión institucional más estricta. El ejemplo más relevante es el Sistema Integral de Seguimiento de los Casos de Violencia de Género (VioGén) constituye una herramienta desarrollada por el Ministerio del Interior de España en el año 2007 con el propósito de reforzar la protección de las víctimas y prevenir la reincidencia en los delitos de violencia de género, integrando en un único marco operativo los esfuerzos de las distintas fuerzas y cuerpos de seguridad del Estado y de las instituciones judiciales y sociales. Su creación se enmarca en la Ley Orgánica 1/2004 de Medidas de Protección Integral contra la Violencia de Género, que estableció la obligación de articular mecanismos coordinados para la valoración del riesgo y la adopción de medidas de protección adecuadas, de modo que VioGén surgió como una respuesta tecnológica e institucional a la necesidad de disponer de un sistema homogéneo y basado en criterios científicos para evaluar el peligro que enfrenta una víctima tras denunciar a su agresor.

El funcionamiento del sistema se sustenta en una lógica de evaluación automatizada del riesgo que combina el análisis estadístico y la ponderación algorítmica de múltiples variables con la valoración profesional de los agentes policiales. El núcleo operativo de VioGén es el Protocolo de Valoración Policial del Riesgo (VPR), un instrumento estructurado que los agentes deben cumplimentar en el momento de la denuncia y que recoge información detallada sobre la víctima, el agresor y el contexto en el que se produce la violencia. El VPR integra diversos factores de riesgo que abarcan aspectos como la existencia de violencia física o sexual previa, el uso de armas, las amenazas directas, los antecedentes del agresor, el consumo de alcohol o drogas, los celos obsesivos, los problemas económicos o laborales, los intentos de suicidio y los conflictos de pareja. Cada uno de estos factores se valora mediante una escala, y a partir de esos datos un algoritmo calcula automáticamente un nivel global de riesgo que puede ser “no apreciado”, “bajo”, “medio”, “alto” o “extremo”,

determinando con ello la intensidad de las medidas de protección policial y judicial que se activarán (Álvarez, 2016).

El sistema, sin embargo, no opera de manera puramente automática, ya que el agente policial mantiene la posibilidad de ajustar el nivel de riesgo resultante si considera que el algoritmo no refleja fielmente la realidad del caso, justificando su decisión conforme al principio del juicio profesional estructurado. Este equilibrio entre la objetividad de la herramienta y la experiencia humana constituye una de las claves de su eficacia, ya que permite conjugar la sistematización de la información con la sensibilidad necesaria en la valoración de cada situación individual. Además, VioGén se nutre de múltiples fuentes de información, incluyendo la declaración de la víctima, los datos del agresor, los testimonios de terceros y los informes policiales, judiciales o sociales existentes, lo que facilita una comprensión integral de cada caso.

Una vez determinado el nivel de riesgo, el sistema activa automáticamente un conjunto de medidas de protección proporcionales, reguladas por la Instrucción 5/2008 de la Secretaría de Estado de Seguridad, que pueden incluir la vigilancia policial, el acompañamiento de la víctima, la custodia domiciliaria, el seguimiento telemático del agresor o la derivación a programas de asistencia social y psicológica. Además, para garantizar la actualización permanente de la información y la adecuación de las medidas de protección, el sistema incluye la Valoración Policial de la Evolución del Riesgo (VPER), una reevaluación obligatoria que debe realizarse en un plazo máximo de tres meses y que permite ajustar la respuesta policial a la evolución de la situación de la víctima y del agresor.

La eficacia del sistema fue evaluada empíricamente por López-Ossorio et al. (2016), en su estudio analizaron 407 casos de mujeres denunciantes seguidas durante tres y seis meses. Los resultados demostraron una buena capacidad predictiva del protocolo, con un área bajo la curva (AUC) de 0.71 a los tres meses y una sensibilidad del 85 por ciento, lo que indica que el instrumento identifica correctamente la mayoría de los casos con riesgo real, mientras que la especificidad, del 53.7 por ciento, refleja un equilibrio razonable entre los aciertos y los falsos positivos. La probabilidad de reincidencia era 6.5 veces mayor en los casos clasificados como de riesgo, lo que valida su utilidad como herramienta de prevención a corto plazo. Los

autores concluyen que el VPR posee unas propiedades psicométricas adecuadas y que su eficacia depende, no solo de la calidad de los datos introducidos, sino también de la formación de los agentes y del uso complementario del juicio profesional.

Más recientemente, en enero de 2025, el Ministerio del Interior presentó una actualización integral del sistema, denominada VioGén 2 y acompañada del nuevo Protocolo 2025, con el fin de modernizar su arquitectura tecnológica, reforzar su interoperabilidad institucional y mejorar la precisión en la gestión del riesgo. Según la nota oficial de La Moncloa, esta versión incorpora una estructura de conexión entre bases de datos que permite integrar información procedente de distintos ámbitos administrativos y judiciales, incluyendo la Base de Datos de Señalamientos Nacionales, el Sistema Integrado de Gestión Operativa de la Guardia Civil, el Sistema Penitenciario y Social Penitenciario, el Sistema Judicial, así como los sistemas autonómicos de Cataluña y el País Vasco y programas de seguimiento como ALERTCOPS, COMETA, ATEMPRO y ONVIOS. Esta interconectividad amplía significativamente la cantidad y calidad de la información disponible para el cálculo del riesgo, generando valoraciones más precisas y actualizadas.

El Protocolo 2025 introduce además importantes modificaciones conceptuales y procedimentales, entre las que destaca la eliminación del nivel "no apreciado" de riesgo, de modo que las valoraciones se agrupan ahora en cuatro niveles: bajo, medio, alto y extremo. Dentro del nivel bajo se establece una subclasificación en función de la existencia de medidas judiciales, lo que permite una gestión más diferenciada y realista de los casos. Se reducen también los plazos para realizar la Valoración Policial de la Evolución del Riesgo y se simplifican los procedimientos para facilitar su actualización constante, asegurando que las nuevas denuncias o hechos relevantes se incorporen de manera inmediata al expediente. Asimismo, se incluyen directrices específicas para los casos que implican menores o víctimas con especial vulnerabilidad, así como para agresores reincidentes o con historial de violencia prolongada, y se incorpora el análisis del riesgo digital derivado del uso de tecnologías y redes sociales. Estas innovaciones se acompañan de un refuerzo en la seguridad informática, una mejora en los mecanismos de notificación automatizada y una optimización en la gestión de alertas y comunicación entre unidades policiales

y judiciales.

A continuación, y ya que han sido mencionadas en relación con la herramienta VioGen, el sistema policial español cuenta con otras herramientas como AlertCops, Cometa, ONVIOS o VeriPol.

La aplicación AlertCops, impulsada por el Ministerio del Interior de España, constituye un ejemplo paradigmático de la llamada seguridad participativa, al permitir que cualquier persona pueda convertirse en emisora directa de información relevante para las Fuerzas y Cuerpos de Seguridad del Estado. Su finalidad esencial es la de establecer un canal de comunicación inmediato, geolocalizado y bidireccional que acerque a los ciudadanos a la autoridad, posibilitando que alerten sobre la comisión de un delito, la existencia de un riesgo o una situación de emergencia. El diseño de AlertCops se apoya en tecnologías de geolocalización, transmisión cifrada y conectividad móvil que permiten una respuesta rápida y eficaz ante los avisos. La herramienta incluye, además, funciones como el botón SOS o la figura del Guardián, que habilitan al usuario a enviar alertas automáticas o compartir su ubicación con personas de confianza, ampliando así el alcance de la protección individual. Este modelo transforma la relación tradicional entre el ciudadano y las fuerzas policiales, al pasar de una comunicación reactiva y presencial a un sistema proactivo y digitalizado que convierte al usuario en un nodo de la red de seguridad pública. Sin embargo, no puede ignorarse que la efectividad de la aplicación se sustenta en la cesión constante de datos personales, lo que sitúa al ciudadano en una posición ambivalente, simultáneamente protegido y vigilado, partícipe de la seguridad colectiva pero también objeto de un control difuso. Desde una perspectiva criminológica, AlertCops encarna la tensión entre el ideal de cooperación ciudadana y el riesgo de una vigilancia ubicua propia de la era digital. La herramienta, en su vocación de servicio público, debe mantener un delicado equilibrio entre utilidad preventiva y respeto a los derechos fundamentales, garantizando que el incremento de la seguridad no se produzca a costa de la erosión de la privacidad.

El sistema COMETA, acrónimo de Control de Medidas Telemáticas, se ha consolidado como uno de los instrumentos más significativos en el marco de la política española

de lucha contra la violencia de género. Su creación responde a la necesidad de asegurar el cumplimiento de las órdenes de alejamiento dictadas por los órganos judiciales y de garantizar una protección efectiva a las víctimas. Este sistema introduce una dimensión tecnológica en el control penal, combinando supervisión electrónica y gestión de datos en tiempo real. En su operativa habitual, la persona sometida a una medida judicial lleva un dispositivo transmisor que registra su posición y la comunica de forma constante, mientras que la víctima dispone de un receptor que la alerta si el agresor se aproxima al perímetro de seguridad fijado. Toda esta información se centraliza en un centro de control operativo las veinticuatro horas del día, lo que permite activar una respuesta inmediata ante cualquier incumplimiento. Más allá de su eficiencia técnica, COMETA simboliza el compromiso del Estado con la protección de las víctimas y con la garantía de un entorno seguro, al tiempo que representa la aplicación práctica del principio de prevención en materia penal. No obstante, la vigilancia telemática introduce también dilemas éticos y jurídicos de considerable alcance. Convertir una medida judicial en un proceso de monitorización continua modifica la naturaleza del control penal y plantea interrogantes sobre los límites de la intervención tecnológica en la esfera privada. El agresor se convierte en objeto de un seguimiento permanente que, si no se administra con proporcionalidad, puede derivar en una forma de vigilancia total. Desde la perspectiva de la política criminal, COMETA revela la compleja búsqueda de equilibrio entre la necesidad de protección de las víctimas y el respeto a los derechos del penado, recordando que la tecnología, por sí sola, no debe desplazar la función garantista del Derecho.

Por otro lado, la Oficina Nacional contra las Violencias Sexuales y su sistema de información ONVIOS constituyen la respuesta más reciente del Estado español a la necesidad de abordar de manera integral y coordinada los delitos de naturaleza sexual. Este instrumento pretende unificar bases de datos dispersas, mejorar la comunicación entre los distintos cuerpos policiales y judiciales, y desarrollar capacidades analíticas que permitan detectar patrones de comportamiento y prevenir la reincidencia. ONVIOS no es solo un registro, sino una plataforma de que aspira a generar conocimiento útil a partir de la información acumulada por las diferentes administraciones. Su finalidad es ofrecer una visión global y coherente

del fenómeno de la violencia sexual, integrando dimensiones policiales, judiciales y penitenciarias. Al permitir el cruce de datos y la detección de coincidencias entre casos, ONVIOS se erige como un instrumento de prevención estratégica y de apoyo a la investigación. Sin embargo, la centralización de información de tan alta sensibilidad plantea desafíos éticos y jurídicos considerables. La gestión de datos personales de víctimas y agresores exige niveles máximos de protección y protocolos estrictos de confidencialidad. El riesgo de un uso indebido o de filtraciones obliga a mantener una vigilancia constante sobre los mecanismos de acceso y sobre la finalidad del tratamiento de los datos. Desde un punto de vista criminológico, ONVIOS refleja la tendencia contemporánea hacia la inteligencia criminal basada en la evidencia, que busca reemplazar la reacción fragmentada por una planificación informada. No obstante, la eficacia de la herramienta dependerá de la calidad de los datos introducidos, de la capacitación técnica de los profesionales que los analicen y de la capacidad del sistema para conjugar la racionalidad técnica con el respeto a los derechos fundamentales. En definitiva, ONVIOS simboliza la madurez de las políticas públicas en materia de violencia sexual: un intento de transformar la información en prevención, el dato en conocimiento y el conocimiento en protección, siempre bajo la premisa de que la tecnología, sin una ética sólida, corre el riesgo de convertirse en un fin en sí misma.

Finalmente, la experiencia española con VeriPol, herramienta desarrollada por la Policía Nacional para detectar denuncias falsas de robo mediante procesamiento del lenguaje natural, ilustra las dificultades que conlleva la implementación práctica de algoritmos policiales. A pesar de haber sido presentada en 2018 como un avance innovador, la herramienta ha sido recientemente retirada. Según un análisis independiente, la Policía no realizó la evaluación de impacto en protección de datos que sería exigible y tampoco publicó información sobre los tratamientos de datos y el perfilado asociados al sistema, pese a utilizar técnicas que implican decisiones automatizadas (Martínez Garay et al., 2024). Además, el propio cuerpo policial reconoció que el uso de VeriPol fue disminuyendo debido a la falta de actualización y a la preocupación de que su fiabilidad se redujera con el tiempo. El caso de VeriPol subraya la necesidad de garantizar la fiabilidad, la transparencia y la supervisión externa antes de incorporar algoritmos en procesos que puedan tener

consecuencias jurídicas directas.

Trasladándonos la fase de investigación criminal, las nuevas tecnologías se han incorporado principalmente a través de sistemas de reconocimiento facial, análisis de vídeo y minería de datos. El software Clearview AI, utilizado por agencias policiales de Estados Unidos y de algunos países europeos, permite comparar imágenes obtenidas de cámaras de seguridad con bases de datos masivas de fotografías en línea. En Europa, el despliegue de tecnologías similares ha sido más limitado por las restricciones derivadas del Reglamento General de Protección de Datos (RGPD), aunque programas como *Avigilon Appearance Search* o *BriefCam* se emplean en análisis forense de grabaciones, permitiendo identificar personas u objetos de interés en miles de horas de vídeo. En China, el sistema Skynet, que combina reconocimiento facial y seguimiento en tiempo real, constituye un ejemplo extremo de vigilancia automatizada, con implicaciones directas sobre la privacidad y la libertad individual. Asimismo, la IA se ha aplicado al análisis forense digital, particularmente en la detección de fraudes, ciberataques y delitos económicos, herramientas como las mencionadas en el ámbito judicial también se han utilizado en el ámbito policial para la investigación del crimen.

3.1.2. Ámbito judicial.

En el ámbito judicial, las tecnologías digitales y la inteligencia artificial pueden ser clasificadas de múltiples formas, en este caso lo dividiremos en dos funciones principales: por un lado, aquellas orientadas a la gestión de casos e investigaciones (*management*), y por otro, las que ofrecen apoyo en la toma de decisiones (*judgement*). Ambas dimensiones buscan incrementar la eficiencia, objetividad y rapidez de la administración de justicia, aunque no están exentas de riesgos y limitaciones.

En este sentido, la gestión judicial se orienta a optimizar los procesos internos del aparato judicial, mediante la incorporación de herramientas tecnológicas que permiten administrar de manera eficiente los casos, los recursos y la información probatoria. Su finalidad es reforzar la transparencia, la trazabilidad y la eficiencia institucional, facilitando la coordinación entre los distintos operadores jurídicos y

garantizando un tratamiento más ágil y equitativo de los expedientes.

En este contexto, los sistemas de gestión de casos han transformado de manera sustancial la organización y administración de los procesos judiciales, tanto en el ámbito público, mediante su implementación en tribunales y fiscalías, como en el ámbito privado, dentro de los despachos de abogados y firmas de litigio. Herramientas como Odyssey Case Manager y Case Center han revolucionado la gestión de causas en los tribunales al centralizar la información procesal, permitir la presentación electrónica de documentos y facilitar la coordinación interinstitucional entre juzgados, fiscalías y defensorías. Estas plataformas, ampliamente utilizadas en sistemas judiciales de Estados Unidos, Canadá y Reino Unido, han contribuido a reducir significativamente los tiempos de tramitación y a fortalecer la transparencia procedimental (Tyler Technologies, 2023; Thomson Reuters, 2022). En el sector privado, la adopción de herramientas digitales también ha avanzado significativamente, y Clio y CaseFleet representan ejemplos destacados de esta evolución tecnológica. Por un lado, Clio, orientada a despachos de abogados, permite centralizar en la nube las comunicaciones, calendarios, registros documentales y facturación, lo que facilita la coordinación entre equipos jurídicos y reduce considerablemente los tiempos administrativos (Clio, 2023). Por otro lado, CaseFleet incorpora funcionalidades avanzadas para la construcción de cronologías procesales y el análisis de pruebas documentales, aportando así una base más sólida para la formulación de estrategias procesales estructuradas y sustentadas en evidencia (CaseFleet, 2022). A estas se suma CaseMap+, desarrollada por LexisNexis, una herramienta analítica de gestión de casos que permite vincular hechos, pruebas, testigos y disposiciones normativas dentro de una base de datos estructurada. Su utilización favorece la visualización de relaciones causales y la detección de inconsistencias o lagunas probatorias, optimizando la preparación de litigios complejos mediante técnicas de análisis relacional y visualización de datos jurídicos (LexisNexis, 2023).

De manera complementaria, existen herramientas especializadas en la gestión, análisis y preservación de la evidencia digital con fines judiciales. Estos sistemas permiten recuperar, examinar y conservar grandes volúmenes de datos

provenientes de dispositivos electrónicos, garantizando la autenticidad, integridad y validez legal de la información analizada. Entre las soluciones más consolidadas, EnCase Forensic, desarrollada por OpenText, destaca por su capacidad para realizar adquisiciones completas de discos duros, recuperar archivos eliminados y preservar metadatos en formatos forenses estandarizados. Su arquitectura asegura el cumplimiento riguroso de la cadena de custodia y la generación de informes admisibles en sede judicial (OpenText, 2023).

En el ámbito del análisis de dispositivos móviles, XRY, creada por MSAB, posibilita la extracción y el examen de registros de llamadas, mensajes, ubicaciones y datos de aplicaciones, incluso en contextos de cifrado o eliminación intencional de información. La herramienta integra algoritmos de reconocimiento y clasificación automática de datos, lo que optimiza el tiempo de análisis en investigaciones complejas (MSAB, 2022).

Asimismo, Forensic Toolkit (FTK), desarrollada por AccessData, constituye una de las soluciones más avanzadas para el procesamiento de grandes volúmenes de evidencia digital. Su motor de indexación permite realizar búsquedas avanzadas en correos electrónicos, documentos y archivos multimedia, además de incorporar módulos de análisis de imágenes y detección de duplicidades. La generación de informes automatizados dota a los análisis de una alta trazabilidad y reproducibilidad (AccessData, 2021). En la misma línea, Cellebrite UFED se ha consolidado como una herramienta de referencia internacional en el ámbito del análisis forense móvil. Permite la extracción y decodificación avanzada de datos provenientes de dispositivos portátiles, almacenamientos externos y servicios en la nube. Su capacidad para recuperar conversaciones, archivos multimedia y metadatos ocultos la convierte en un recurso esencial en investigaciones penales donde los teléfonos inteligentes constituyen fuentes críticas de evidencia (Cellebrite, 2022).

Más recientemente, Magnet AXIOM, desarrollada por Magnet Forensics, ha ampliado el alcance del análisis digital al integrar en una sola plataforma la extracción y correlación de datos procedentes de dispositivos móviles, redes sociales, navegadores y servicios en la nube. Su arquitectura algorítmica permite reconstruir

de forma cronológica los comportamientos digitales y generar informes periciales estandarizados, fortaleciendo la consistencia, trazabilidad y validez de las investigaciones judiciales (Magnet Forensics, 2022). Otra innovación destacable dentro del management judicial corresponde a la reconstrucción algorítmica de secuencias de eventos. Tecnologías como SIMULIA, desarrollada por Dassault Systèmes, permiten realizar simulaciones físicas de colisiones o reconstruir dinámicas delictivas, ofreciendo a jueces y peritos representaciones visuales precisas de los hechos bajo examen (Dassault Systèmes, 2021). De forma análoga, CityStream, impulsada en Japón por Nexar, combina datos GPS y grabaciones vehiculares para recrear accidentes de tráfico en tiempo real, generando informes cronológicos de alto valor probatorio (Nexar, 2021).

En el ámbito de los delitos económicos y financieros, los sistemas de detección automatizada de fraude representan un componente esencial dentro del ecosistema tecnológico de apoyo a la gestión judicial y a la investigación penal. Estas herramientas emplean técnicas avanzadas de aprendizaje automático y detección de anomalías para identificar, en tiempo real, comportamientos atípicos en grandes volúmenes de transacciones financieras. IBM Safer Payments, desarrollada por IBM, permite analizar millones de operaciones en milisegundos, aplicando modelos adaptativos que aprenden continuamente de nuevos patrones de fraude y reducen los falsos positivos en la detección (IBM, 2022). Por su parte, SAS Fraud Management, de SAS Institute, combina reglas expertas y algoritmos predictivos basados en inteligencia artificial para supervisar de manera integral redes de pago, seguros o transferencias, identificando de forma temprana actividades sospechosas de estafa o blanqueo de capitales. La integración de estas soluciones en los procesos judiciales y en los sistemas de cumplimiento normativo ha fortalecido la capacidad institucional para responder con agilidad ante delitos económicos complejos, aportando evidencia digital verificable y auditables (SAS Institute, 2021).

Finalmente, el uso de plataformas de asesoramiento legal automatizado representa una tendencia emergente hacia la digitalización de la gestión jurídica. Sistemas como Neota Logic y LegalZoom demuestran cómo la inteligencia artificial contribuye a democratizar el acceso a la información legal y a automatizar procedimientos

recurrentes. Neota Logic permite transformar conocimiento jurídico especializado en aplicaciones interactivas sin requerir conocimientos de programación, mientras que LegalZoom combina formularios inteligentes con asistencia profesional para agilizar trámites como la constitución de sociedades o la elaboración de contratos (Neota Logic, 2021; LegalZoom, 2022).

El segundo gran ámbito, la toma de decisiones (judgement) en el ámbito judicial, abarca el conjunto de herramientas algorítmicas e inteligentes diseñadas para asistir en la labor decisoria de jueces, magistrados y demás operadores jurídicos. Estas tecnologías se enfocan en procesos de evaluación, predicción y apoyo a la decisión, a partir del análisis de datos judiciales, antecedentes penales y otra información contextual relevante. A diferencia de los sistemas de gestión, las herramientas incluidas en esta categoría intervienen directamente en el proceso valorativo del quehacer judicial, ofreciendo estimaciones o recomendaciones vinculadas con la valoración de riesgos, la predicción de resultados o la coherencia jurisprudencial. En este contexto, destaca el uso de la sentencia predictiva (*predictive sentencing*), la cual emplea algoritmos de evaluación de riesgo para estimar la probabilidad de reincidencia de un acusado y asistir al juez en la determinación de la pena. Si bien busca incrementar la consistencia y objetividad en las decisiones judiciales, su aplicación ha suscitado debate debido a los posibles sesgos y limitaciones inherentes a los modelos utilizados (Dressel & Farid, 2018). Cabe señalar que este tipo de herramientas también se utilizan en el contexto penitenciario, especialmente para valorar la progresión de grado o la concesión de beneficios, y que trazar una línea divisoria precisa entre ambos ámbitos resulta complejo; por ello, no debe adoptarse una postura excesivamente estricta respecto de la clasificación aquí propuesta.

En primer lugar, encontraríamos herramientas enfocadas a la evaluación de la reincidencia, estas herramientas utilizan algoritmos, modelos estadísticos o de aprendizaje automático para estimar la probabilidad de reincidencia delictiva o de incumplimiento de condiciones judiciales (Andrés-Pueyo & Echeburúa, 2010). Uno de los ejemplos más reconocidos a nivel internacional es el sistema Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), desarrollado

en Estados Unidos por la empresa Northpointe Inc., actualmente Equivant. Este modelo calcula la probabilidad de reincidencia o de incumplimiento de medidas judiciales a partir de cuestionarios estructurados que recogen información personal, social y criminológica del individuo, así como de bases de datos penitenciarias y judiciales que incluyen antecedentes, edad al primer delito o tipo de infracción cometida (Brennan et al., 2009). El sistema emplea métodos estadísticos y de aprendizaje automático, como la regresión logística, el análisis de supervivencia y técnicas de clasificación bootstrap, para generar una puntuación de riesgo entre 1 y 10 que indica la probabilidad de que una persona vuelva a delinquir en un período de dos años. Estas puntuaciones se agrupan en tres categorías: bajo riesgo (1–4), riesgo medio (5–7) y alto riesgo (8–10). El resultado se presenta en informes visuales que integran la valoración global del riesgo y los factores que más contribuyen a dicha estimación. COMPAS se utiliza habitualmente en fases decisorias del proceso penal, especialmente en evaluaciones de libertad condicional, fijación de fianzas y determinación de penas alternativas a la prisión, constituyendo uno de los sistemas de apoyo algorítmico más consolidados en la práctica judicial estadounidense (Vaccaro, 2019). Su propósito declarado es aumentar la objetividad y la eficiencia de las decisiones judiciales, reduciendo la carga de trabajo y la variabilidad entre evaluadores humanos.

En una línea similar, y aunque no utilizan inteligencia artificial para realizar la valoración del riesgo encontramos el SARA (Kropp, Hart, Webster y Eaves, 1995) y el SAVRY (Borum, Bartel y Forth, 2005).

El Spousal Assault Risk Assessment Guide (SARA) es una de las herramientas más utilizadas a nivel internacional para la evaluación del riesgo de violencia contra la pareja. Fue desarrollada en Canadá por Kropp, Hart, Webster y Eaves (1995) con el propósito de ofrecer un protocolo estructurado que permitiera estimar la probabilidad de que un agresor reincida en comportamientos violentos hacia su pareja o expareja. El instrumento se basa en el modelo de juicio profesional estructurado (Structured Professional Judgment, SPJ), que combina la evidencia empírica con la valoración clínica del evaluador. El SARA está compuesto por 20 ítems que abarcan factores de riesgo tanto históricos (como antecedentes de

violencia doméstica o delitos previos), clínicos (por ejemplo, consumo de sustancias, trastornos emocionales o actitudes de control y celos) y contextuales (como incumplimiento de órdenes judiciales o falta de apoyo social). A partir de la información recopilada mediante entrevistas, expedientes judiciales y policiales, informes psicológicos y otros registros oficiales, el evaluador emite una valoración del nivel de riesgo (bajo, medio o alto) sin que se derive de una fórmula matemática, sino de la integración razonada de los factores observados. El SARA se aplica habitualmente en contextos judiciales, especialmente en la valoración del riesgo de reincidencia en casos de violencia de género o doméstica, y sirve como guía para decidir la adopción de medidas de protección, órdenes de alejamiento o supervisión judicial. También se emplea en algunos entornos penitenciarios como apoyo para planificar la intervención y tratamiento de agresores de pareja, facilitando la coordinación entre servicios judiciales, policiales y sociales en la gestión integral del riesgo (Kropp et al., 1995).

Por su parte, el Structured Assessment of Violence Risk in Youth (SAVRY) es una herramienta diseñada para la evaluación del riesgo de violencia en adolescentes y jóvenes de entre 12 y 18 años, elaborada por Borum, Bartel y Forth (2005). Al igual que el SARA, se basa en el enfoque de juicio profesional estructurado, ofreciendo una guía sistemática para estimar la probabilidad de comportamientos violentos en población juvenil. El SAVRY está compuesto por 24 ítems distribuidos en tres grandes dominios: factores históricos (como antecedentes de violencia, problemas familiares o consumo de drogas), factores sociales/contextuales (por ejemplo, relaciones con iguales, actitudes hacia la autoridad o implicación escolar) y factores individuales o clínicos (como impulsividad, empatía o manejo de la ira). Además, incluye una sección adicional de factores protectores, que permite identificar elementos que podrían reducir la probabilidad de reincidencia, como el apoyo familiar o las metas prosociales. La información se obtiene mediante entrevistas con el menor, revisión de expedientes judiciales y escolares, así como informes de trabajadores sociales o psicólogos. A partir de esta información, el evaluador determina el nivel de riesgo global de violencia (bajo, moderado o alto), formulando además recomendaciones sobre las necesidades de intervención y supervisión. El SAVRY se utiliza principalmente en el ámbito judicial juvenil, tanto en fases pre-

sentenciales como durante la ejecución de medidas judiciales o programas de reinserción, con el objetivo de orientar decisiones sobre internamiento, libertad vigilada o derivación a programas de intervención socioeducativa (Borum, Bartel, & Forth, 2005).

Otro desarrollo destacado es el Public Safety Assessment, impulsado por la Arnold Foundation en Estados Unidos, que se emplea en las audiencias de medidas cautelares para estimar la probabilidad de reincidencia o de incomparecencia a juicio. A diferencia de los modelos tradicionales, este sistema se basa en un conjunto reducido de variables objetivas, tales como el tipo de delito, la edad o los antecedentes del acusado, generando un puntaje de riesgo que sirve de apoyo al juez en la decisión sobre la libertad provisional.

En el contexto británico, el Offender Assessment System, desarrollado por el Ministerio de Justicia, permite evaluar de manera sistemática los factores criminógenos y las necesidades de intervención de cada condenado, mientras que la herramienta denominada Harm Assessment Risk Tool, creada por la policía de Durham, aplica técnicas de aprendizaje automático para estimar la probabilidad de reincidencia a partir de registros policiales y judiciales.

Paralelamente al desarrollo de modelos centrados en la predicción del comportamiento delictivo, se han consolidado diversas iniciativas orientadas a la predicción de resultados judiciales y al análisis jurisprudencial. En este campo destaca el proyecto European Court of Human Rights Prediction Project, elaborado por Aletras y colaboradores (2016), que demostró la viabilidad de aplicar modelos de aprendizaje automático capaces de anticipar con un alto grado de precisión las decisiones del Tribunal Europeo de Derechos Humanos a partir del análisis textual de sentencias anteriores. En Francia, el laboratorio OpenJustice y la empresa Case Law Analytics han diseñado sistemas que analizan grandes bases de datos jurisprudenciales con el fin de calcular probabilidades de éxito de demandas o recursos en distintas materias, entre ellas el derecho laboral, civil y contencioso-administrativo. Estas herramientas se emplean principalmente por abogados y

despachos jurídicos, aunque en algunos casos sirven también de referencia para la magistratura al evaluar la coherencia de los precedentes aplicables.

En los contextos estadounidense y canadiense, el uso de la inteligencia artificial se ha orientado hacia plataformas de análisis predictivo de carácter comercial. Entre ellas se encuentran Lex Machina y Premonition, que ofrecen información estadística sobre tendencias judiciales, perfiles de jueces y tasas de éxito de litigios. También destaca Blue J Legal, un sistema desarrollado en Canadá que utiliza técnicas de aprendizaje automático para prever posibles desenlaces en casos tributarios y laborales, a partir del análisis de los hechos y de los criterios jurisprudenciales más relevantes. Este tipo de soluciones se basa en la extracción automatizada de patrones de decisión y constituye un ejemplo del modo en que la minería de datos jurídicos comienza a integrarse progresivamente en la práctica judicial y en la estrategia procesal.

Además de los sistemas de evaluación y predicción, se desarrollan actualmente herramientas destinadas a asistir el razonamiento jurídico mediante el procesamiento automático del lenguaje natural. Estas aplicaciones no emiten predicciones sobre el fondo del litigio, pero facilitan el análisis de textos legales, la identificación de precedentes y la elaboración de argumentaciones coherentes. Una de las experiencias más relevantes en este ámbito es ROSS Intelligence, plataforma basada en la tecnología Watson de IBM, capaz de responder a consultas jurídicas mediante el análisis semántico de amplias bases de datos normativos y jurisprudenciales. En Brasil, el sistema Jurimetria aplica modelos similares para ofrecer información estadística sobre la duración de los procesos, la frecuencia de determinados fallos y la probabilidad de éxito de una acción judicial, integrándose parcialmente en los tribunales superiores del país. En Europa, proyectos como BigData4Law y el European e-Justice Portal promueven el uso de la minería jurisprudencial para mejorar el acceso a la información judicial y favorecer la coherencia en la aplicación del derecho, consolidando un marco de desarrollo conjunto entre tecnología y práctica judicial.

El panorama actual muestra una diversidad creciente de enfoques y niveles de implementación en el uso de herramientas algorítmicas aplicadas a la dimensión de

judgement. En el ámbito penal, los sistemas de evaluación de riesgos se presentan como las aplicaciones más consolidadas, mientras que en las jurisdicciones civil y supranacional proliferan los proyectos de predicción jurisprudencial y de asistencia cognitiva. En conjunto, estas tecnologías configuran un proceso de transformación gradual del ecosistema judicial hacia un modelo en el que el análisis de datos, la modelización estadística y el aprendizaje automático se integran como instrumentos de apoyo al razonamiento y a la decisión, conformando un escenario en el que la justicia se apoya cada vez más en la información y en la capacidad analítica de los sistemas inteligentes.

3.1.3. Ámbito penitenciario.

Dentro del ámbito penitenciario, la incorporación de tecnologías digitales e inteligencia artificial en los centros penitenciarios aún se encuentra en una fase incipiente en comparación con su desarrollo en otros ámbitos del sistema de justicia, como la administración judicial o las fuerzas policiales. Las prisiones inteligentes se distinguen por el uso de tecnologías como la inteligencia artificial, el Internet de las cosas, la biometría y el análisis de datos masivos, con el propósito de crear entornos más seguros, transparentes y eficientes. En esta línea, McKay (2022) identifica dos vertientes principales en el desarrollo de las prisiones digitales: por un lado, las tecnologías orientadas al control, la seguridad y la eficiencia administrativa, y por otro, aquellas dirigidas a favorecer la rehabilitación y la reinserción de las personas privadas de libertad.

Integrando ambas vertientes, uno de los conceptos que ha adquirido mayor relevancia en los últimos años es el de *Smart Prison* o prisión inteligente, entendido como una propuesta innovadora que busca transformar la gestión carcelaria mediante la incorporación de tecnologías digitales y sistemas de inteligencia artificial (Kaun & Stiernstedt, 2020; Lindström & Puolakka, 2021; Puolakka, 2023; Puolakka & Nurmi, 2022). Este modelo surge como respuesta a la necesidad de modernizar las infraestructuras penitenciarias y optimizar los procesos de control, seguridad y rehabilitación, en un contexto donde la tecnología se ha consolidado como un instrumento central para la eficiencia institucional (Imandeka, Hidayanto & Mahmud, 2023). El concepto de prisión inteligente se fundamenta en la

integración de herramientas digitales avanzadas orientadas a mejorar la seguridad, la eficiencia operativa y la reinserción social de las personas privadas de libertad. A diferencia de las prisiones tradicionales, donde el control depende casi exclusivamente de la vigilancia humana, las *Smart Prisons* implementan plataformas digitales capaces de monitorear en tiempo real la ubicación, el comportamiento y las rutinas tanto de los internos como del personal penitenciario (Fedorczyk, 2024). Los objetivos principales de las *Smart Prisons* son incrementar la seguridad, reducir los incidentes internos, optimizar los recursos humanos y materiales, y promover una gestión penitenciaria más transparente y eficiente (Imandeka, Hidayanto & Mahmud, 2023).

En cuanto a las principales herramientas implantadas, uno de los casos más relevantes en Europa es el de Finlandia, donde el *Smart Prison Project* (2018–2022) marcó un hito en la digitalización penitenciaria (Puolakka, 2023). Este proyecto, desarrollado por el *Prison and Probation Service of Finland*, tuvo como finalidad convertir la prisión en un entorno de aprendizaje orientado a la vida sin delincuencia, promoviendo la inclusión digital de personas previamente marginadas y garantizando el acceso equitativo a servicios públicos. La primera prisión inteligente se inauguró en Hämeenlinna, un centro femenino equipado con dispositivos digitales individuales en cada celda que operan a través del sistema Doris, utilizado tanto por internas como por personal penitenciario y servicios de salud. Este sistema integra funcionalidades de comunicación interna, gestión de citas, educación virtual, videollamadas con familiares y profesionales, acceso restringido a Internet, y recursos digitales de apoyo rehabilitador. La experiencia demostró una alta aceptación por parte de las internas, quienes utilizan el sistema a diario, valorando especialmente la posibilidad de mantener contacto directo con familiares y acceder a programas formativos desde la celda (Puolakka, 2023). Asimismo, Finlandia ha extendido el modelo a otras prisiones e incorporado innovaciones como la formación en inteligencia artificial (*AI training*) como forma de trabajo penitenciario, cursos de alfabetización digital desarrollados por la Universidad de Helsinki (*Elements of AI*), y programas de realidad virtual orientados al bienestar psicológico y la reducción de la ansiedad. El país también experimenta con plataformas basadas en IA, como Aurora AI y RISE AI, destinadas a recomendar

servicios públicos o analizar perfiles de internos para planificar su ubicación y programas de rehabilitación. En Europa, varios países experimentan con la integración tecnológica orientada tanto al control como a la rehabilitación y educación digital, por ejemplo, en el Reino Unido, el *Her Majesty's Prison and Probation Service* (HMPPS) impulsa el acceso a tabletas seguras en celdas, videollamadas y programas de formación a distancia dentro de su Digital, Data and Technology Strategy (HMPPS, 2021). Este tipo de propuestas, en línea con lo que defiende Lopez Lorca (2023), permite ir más allá de la simple modernización de las administraciones públicas, reconfigurando el proceso de resocialización hacia un modelo más innovador y humano.

En relación con el control, la seguridad y la eficiencia, la inteligencia artificial y la digitalización se han convertido en herramientas clave para reforzar los mecanismos de vigilancia y control penitenciario. A nivel internacional, empresas como DAVANTIS han desarrollado sistemas de videovigilancia inteligente aplicados a prisiones españolas, con algoritmos de análisis en tiempo real que identifican comportamientos anómalos. En Asia, compañías como ACTI se han especializado en vigilancia penitenciaria mediante cámaras inteligentes (ACTI, 2023), mientras que en Canadá destacan soluciones integradas como *Senstar Sensor Fusion Engine*, que combina sensores, cámaras y análisis de datos en tiempo real. En el Reino Unido, plataformas como Synaedge ofrecen herramientas similares para reforzar la seguridad penitenciaria. En este contexto, diversos países han comenzado a implementar modelos basados en este paradigma tecnológico. En Hong Kong, el Departamento de Servicios Correccionales ha implementado un sistema de gestión inteligente que integra pulseras electrónicas y sensores para el seguimiento en tiempo real de las personas bajo custodia. Estas herramientas forman parte del denominado *Movement and Location Monitoring System,* diseñado para detectar movimientos anómalos o aglomeraciones y optimizar la supervisión diaria del personal penitenciario. Además, estos sistemas permiten la vigilancia continua de signos vitales mediante tecnología inalámbrica de banda ultraancha, lo que posibilita una respuesta médica inmediata ante emergencias (*Contactless Vital Signs Monitoring System*).

En relación con el seguimiento, pero en una etapa postpenitenciaria, los dispositivos electrónicos de monitoreo se consolidan como una alternativa a la privación de libertad. En Chile, durante la pandemia de COVID-19, se utilizó la aplicación GeoVictoria para supervisar en tiempo real a internos en arresto domiciliario. En Hong Kong, el *Operation Management System* combina inteligencia artificial con pulseras electrónicas para reforzar el control de liberados condicionales. En Estados Unidos, diversas compañías han desarrollado sistemas de geolocalización con pulseras inteligentes que emplean inteligencia artificial para anticipar patrones de incumplimiento. En España, estos dispositivos se aplican especialmente en casos de violencia de género, donde las pulseras electrónicas garantizan la efectividad de las órdenes de alejamiento, complementándose con aplicaciones móviles para las víctimas (Ministerio del Interior, 2022).

Un campo emergente dentro del control, la seguridad y la eficiencia penitenciaria es la prevención del suicidio y las autolesiones, donde la tecnología comienza a desempeñar un papel cada vez más relevante. La empresa Unilink ha desarrollado la aplicación AIM, que se integra en su software de gestión penitenciaria para analizar de manera continua datos sobre el comportamiento de los internos, como el uso del teléfono, las visitas, la participación en actividades educativas o laborales y los indicadores de aislamiento. A partir de estos patrones, el sistema identifica perfiles de riesgo y alerta al personal cuando detecta señales asociadas a posibles conductas autolesivas, favoreciendo así una intervención temprana (Unilink, 2023). Esta herramienta forma parte de un conjunto más amplio de soluciones tecnológicas desarrolladas por la empresa, entre las que destacan la plataforma U-Case, orientada a la gestión integral de casos desde la custodia hasta la reintegración en la comunidad; el sistema de autoservicio para personas internas, que permite realizar gestiones cotidianas de manera digital y reduce la carga administrativa del personal; y eMates, una aplicación de comunicación segura que facilita el contacto entre los internos y sus familiares o allegados. Por otro lado, en países como Australia y Hong Kong se han promovido o probado sistemas basados en videoanalítica e inteligencia artificial orientados a la vigilancia y la detección temprana de incidentes dentro de las prisiones, incluyendo potencialmente comportamientos autolesivos. Sin embargo, la evidencia sobre su eficacia o su aplicación sistemática aún es limitada y

requiere una evaluación más profunda (iOmni, 2022).

En el ámbito de la evaluación y la rehabilitación, herramientas actuariales como COMPAS, LSI-R, TVR y RisCanvi continúan siendo fundamentales en la clasificación de internos y en la determinación de programas de tratamiento (Andrews & Bonta, 1995; Andrés Pueyo, Arbach y Redondo, 2010). La herramienta COMPAS, ya explicada en apartados anteriores, constituye uno de los ejemplos paradigmáticos de evaluación automatizada del riesgo en el contexto judicial norteamericano.

El Level of Service Inventory Revised (LSI-R) es, por su parte, uno de los instrumentos actuariales más reconocidos y validados internacionalmente para la evaluación del riesgo de reincidencia y de las necesidades criminógenas en personas infractoras (Andrews & Bonta, 1995; 2010). Fue diseñado a partir del modelo teórico Risk-Need-Responsivity (RNR) (Andrews, Bonta & Wormith, 2008), que sostiene que la intervención penal debe ajustarse a tres principios fundamentales: el principio de riesgo, según el cual la intensidad del tratamiento debe corresponder al nivel de riesgo del individuo; el principio de necesidad, que orienta la intervención hacia los factores dinámicos relacionados con la conducta delictiva; y el principio de responsividad, que propone adaptar los métodos de intervención a las características personales del sujeto.

El LSI-R surge como una herramienta destinada a operacionalizar estos principios, integrando información objetiva y sistemática que permite a los profesionales orientar las decisiones relativas a la clasificación penitenciaria, la supervisión en libertad condicional y la planificación del tratamiento (Andrews & Bonta, 1995). A diferencia de las evaluaciones tradicionales basadas únicamente en el juicio clínico o en factores estáticos, este instrumento incorpora una amplia gama de variables dinámicas susceptibles de cambio, lo que lo convierte en una herramienta no solo predictiva, sino también útil para el seguimiento y la reevaluación del progreso individual (Andrews, Bonta & Wormith, 2008).

El cuestionario del LSI-R está compuesto por 54 ítems distribuidos en 10 dimensiones que abarcan tanto factores personales como contextuales: antecedentes delictivos, educación y empleo, situación económica, relaciones

familiares y conyugales, alojamiento, ocio y recreación, compañeros, consumo de alcohol y drogas, problemas emocionales y actitud u orientación delictiva. Cada ítem se puntúa de manera dicotómica y el total obtenido se asocia a diferentes categorías de riesgo (bajo, medio o alto). La combinación de estas dimensiones permite una visión integral del individuo, incorporando tanto elementos de vulnerabilidad estructural como indicadores de motivación y apoyo social. Desde una perspectiva aplicada, el LSI R no solo sirve como herramienta de diagnóstico, sino también como guía para la intervención y la gestión del riesgo. En el ámbito penitenciario y de la ejecución penal comunitaria, los resultados del instrumento orientan la asignación de programas específicos, como la intervención en drogodependencias o la mejora de habilidades sociales, y ayudan a determinar la intensidad y el tipo de supervisión necesarios. De esta forma, el LSI-R favorece una utilización más eficiente de los recursos y contribuye a reducir la reincidencia mediante la focalización del tratamiento en los factores de riesgo dinámicos (Bonta & Andrews, 2017).

En el contexto penitenciario español, la aplicación de herramientas de carácter actuarial se ha consolidado progresivamente como una práctica habitual dentro del sistema de individualización científica. Estas metodologías se utilizan con el propósito de predecir el comportamiento futuro de las personas privadas de libertad, especialmente en los procedimientos vinculados a la concesión de permisos de salida. Entre los instrumentos más representativos destacan la Tabla de Variables de Riesgo (TVR) (Fanega, Portillo & Camacho, 2024) y el informe de Concurrencia de Circunstancias Peculiares (CCP), ambos concebidos para objetivar la valoración del riesgo y reducir la discrecionalidad en la toma de decisiones.

La TVR integra un conjunto de indicadores empíricamente asociados al éxito o fracaso de los permisos penitenciarios, tales como la nacionalidad, la existencia de drogodependencia, la reincidencia delictiva o el historial de quebrantamientos previos. Cada uno de estos factores se pondera en función de su correlación estadística con la probabilidad de incumplimiento de las condiciones de la salida, lo que permite estimar el nivel de riesgo y orientar las decisiones de tratamiento y reinserción del interno. Por su parte, el informe CCP se centra en la valoración cualitativa de variables no estrictamente cuantificables, pero señaladas por el

artículo 156 del Reglamento Penitenciario como de especial relevancia. Entre ellas se incluyen la trayectoria delictiva singular, la presencia de rasgos de personalidad anómalos o la concurrencia de circunstancias personales desfavorables que puedan interferir en el proceso de reintegración. El funcionamiento de estas técnicas actuariales requiere comprender que tanto las resoluciones favorables como las denegaciones de permisos se sustentan en categorías previamente definidas. Así, la mayoría de las causas de denegación guardan relación directa con los factores de riesgo contemplados en la TVR y en el informe CCP, aunque también pueden existir motivos no vinculados expresamente con dichas variables. Estas herramientas, de uso extendido en los centros penitenciarios dependientes de la Administración General del Estado, han contribuido a dotar de mayor coherencia y trazabilidad a las decisiones adoptadas por las Juntas de Tratamiento.

Junto a estos instrumentos, merece especial atención una herramienta desarrollada en el ámbito autonómico catalán: RisCanvi. Este protocolo de valoración del riesgo de reincidencia violenta y de otros comportamientos problemáticos en prisión fue elaborado entre 2007 y 2009 por el Grupo de Estudios Avanzando en Violencia, a solicitud de los Servicios Penitenciarios de la Generalitat de Cataluña. Desde su incorporación en 2010 a los procedimientos ordinarios de gestión del tratamiento, el sistema opera de manera estable y plenamente integrado en la práctica profesional de los equipos técnicos. RisCanvi fue concebido para estimar la probabilidad de reincidencia y ofrecer apoyo técnico a las Juntas de Tratamiento en la toma de decisiones relativas a la progresión de grado o la adecuación de las intervenciones terapéuticas. Aunque inicialmente se diseñó para evaluar a personas condenadas por delitos violentos, su alcance se ha ampliado con el tiempo a otras tipologías delictivas presentes en el medio penitenciario. En su versión más reciente, Andrés Pueyo, Arbach y Redondo (2018) describen un total de 43 factores de riesgo que abarcan dimensiones como la historia delictiva y violenta, la conducta penitenciaria, los antecedentes personales y biográficos, las condiciones sociofamiliares y los elementos contextuales que pueden influir en la probabilidad de reincidencia. El resultado de la evaluación se expresa en una escala de riesgo de carácter tricotómico: alto, medio o bajo (Andrés Pueyo et al., 2010).

En cuanto a la capacidad predictiva del instrumento, Férez Mangas (2017a, 2017b)

señala que la escala de riesgo de quebrantamiento del RisCanvi alcanza un porcentaje de clasificaciones correctas del 69,9 %, con una sensibilidad del 74,8 %, especificidad del 67,2 % y valores predictivos del 56,1 % (positivo) y 82,6 % (negativo), obteniendo un AUC de 0,738 y un odds ratio de 6,065. Estos resultados confirman su fiabilidad como herramienta de apoyo para la toma de decisiones en el ámbito penitenciario. En los últimos años, el sistema ha sido objeto de actualizaciones que incluyen la revisión de sus algoritmos de cálculo y la incorporación de una nueva escala de valoración del riesgo de reincidencia común, lo que ha permitido ampliar su aplicabilidad a un espectro más amplio de perfiles delictivos. De este modo, RisCanvi se ha convertido en un instrumento central en la gestión del tratamiento penitenciario, contribuyendo a mejorar la objetividad de las evaluaciones, orientar los cambios de grado y facilitar la planificación postpenitenciaria (Andrés Pueyo, Arbach & Redondo, 2018).

Esta es solamente una muestra representativa de los múltiples usos y aplicaciones que la transformación tecnológica puede llegar a tener en los ámbitos policial, judicial y penitenciario. La clasificación presentada no pretende ser exhaustiva, sino más bien ilustrativa de las potencialidades que estas tecnologías ofrecen en la actualidad y de las diversas formas en que comienzan a integrarse en las estructuras institucionales de la justicia contemporánea. Cada una de las herramientas descritas a lo largo del apartado refleja una manifestación concreta del modo en que la inteligencia artificial, los modelos algorítmicos y el análisis masivo de datos están contribuyendo a redefinir los procesos de evaluación, decisión y gestión, introduciendo nuevas lógicas en la detección del riesgo, en la optimización de recursos y en la implementación de medidas de control, seguridad y rehabilitación. No obstante, más allá de su valor instrumental, estas tecnologías simbolizan una transformación de mayor profundidad, que afecta a la propia epistemología del sistema de justicia, al modo en que se produce, interpreta y aplica la información dentro de un marco cada vez más digitalizado.

Conviene subrayar que este panorama no es en absoluto estático, contrariamente la naturaleza de la innovación tecnológica implica un proceso de desarrollo constante, en el que incluso mientras se redacta este capítulo continúan gestándose nuevos

modelos, aplicaciones y sistemas que probablemente modificarán las prácticas actuales y reconfigurarán el futuro. Se trata de un cambio que avanza con una velocidad sin precedentes, impulsado por el perfeccionamiento de las técnicas de aprendizaje automático, la expansión del procesamiento del lenguaje natural, la generalización del big data y la creciente integración de infraestructuras digitales interconectadas. En este sentido, puede afirmarse que el proceso de automatización de la justicia se encuentra en una fase de expansión que, lejos de mostrar signos de desaceleración, continúa intensificándose. Asimismo, es previsible que las herramientas actualmente empleadas sean objeto de revisión, actualización y perfeccionamiento continuo. A medida que se incrementan las capacidades computacionales y se amplían las fuentes de información disponibles, surgirán sistemas más sofisticados, capaces de incorporar variables dinámicas, indicadores contextuales y datos procedentes de fuentes heterogéneas. Este tipo de desarrollos, que hoy pueden parecer hipotéticos, ilustran la tendencia hacia modelos predictivos más complejos, con una mayor capacidad de aprendizaje y adaptación, capaces de anticipar riesgos con niveles de precisión hasta ahora inéditos, pero también con un potencial ético y social que exige una regulación prudente y un escrutinio riguroso por parte de las instituciones y de la sociedad civil.

3.2. Las expectativas de la digitalización, la algoritmización y la inteligencia artificial: beneficios y riesgos esperados del uso de las nuevas tecnologías.

Como todo proceso de transformación, la incorporación de innovaciones en cualquier ámbito suele generar debates en torno a los beneficios y riesgos que dichos cambios pueden implicar, y el sistema de justicia penal no ha sido la excepción. Las reflexiones sobre la inclusión de la digitalización, algoritmización y la inteligencia artificial ha dado lugar a resultados tan diversos como las perspectivas de quienes han participado en esta discusión (Castro-Toledo, 2022; Cortina Orts, 2019; Miró Llinares, 2018). A pesar de dicha variedad, el debate tiende a organizarse en torno a dos grandes posturas: por un lado, quienes consideran que esta nueva revolución tecnológica augura un futuro más prometedor y objetivo, en el que la justicia funcionará de manera más eficiente; y por otro, quienes advierten

que esta etapa podría acarrear riesgos importantes en relación con la transparencia, las garantías procesales y la posibilidad de generar nuevas formas de discriminación.

La digitalización del sistema de justicia penal, así como la progresiva incorporación de la inteligencia artificial, constituyen uno de los fenómenos más significativos de transformación institucional de las últimas décadas. Como todo proceso de cambio estructural, este despliegue ha generado debates intensos que oscilan entre la exaltación de sus beneficios potenciales y la cautela frente a los riesgos éticos y jurídicos que entraña. En términos generales, el discurso académico y político se organiza en torno a dos grandes posiciones: quienes confían en que la digitalización permitirá una justicia más eficiente, accesible y objetiva, y quienes advierten que esta transición tecnológica puede poner en riesgo la transparencia, las garantías procesales y la igualdad de trato. La tensión entre innovación y garantías se ha convertido así en el núcleo del debate en la incorporación de las nuevas tecnologías en el sistema de justicia penal (Miró-Llinares, 2020; Završnik, 2020).

Desde una perspectiva optimista, la digitalización y el uso de sistemas algorítmicos ofrecen ventajas evidentes. Entre los beneficios que ofrece la digitalización del sistema de justicia penal, uno de los más evidentes se encuentra en la capacidad de gestionar con mayor eficiencia enormes volúmenes de información, pues la informatización de expedientes, registros y resoluciones permite reducir los tiempos de tramitación y liberar recursos que hasta ahora se destinaban a tareas repetitivas y mecánicas. La automatización de estos procesos administrativos no supone la sustitución de jueces, fiscales o funcionarios, sino su reorientación hacia funciones de mayor valor añadido, como la interpretación normativa, la ponderación de derechos o el control de legalidad, reforzando así el papel humano en la esfera donde realmente resulta insustituible (Wischmeyer & Rademacher, 2020). Esta redistribución de tareas representa una oportunidad para que el personal judicial dedique más atención a la fundamentación cualitativa de las decisiones, mientras los sistemas tecnológicos se encargan de los aspectos formales de gestión.

Otro beneficio relevante de la incorporación de la inteligencia artificial en el ámbito

judicial es la posibilidad de reducir la arbitrariedad en la toma de decisiones. El análisis automatizado de amplios conjuntos de resoluciones permite identificar patrones recurrentes y criterios compartidos, lo que contribuye a disminuir las disparidades injustificadas entre casos que, en el modelo tradicional, dependían en mayor medida de la interpretación individual. Este tipo de herramientas facilita la detección de consistencias en la argumentación jurídica y puede favorecer una mayor previsibilidad en los fallos, reforzando la seguridad jurídica sin sustituir el razonamiento humano (Ramos-Maqueda & Chen, 2025). Diversos estudios empíricos han mostrado que la aplicación de algoritmos de aprendizaje automático en contextos judiciales tiende a mejorar la coherencia entre decisiones, especialmente cuando se emplean con finalidades de apoyo y no de reemplazo (Rosili et al., 2021). Asimismo, la digitalización de los procesos judiciales genera un entorno propicio para la estandarización de prácticas y la reducción de contradicciones internas que afectan la confianza de los justiciables. Investigaciones recientes realizadas con magistrados europeos señalan que existe una disposición mayoritariamente favorable hacia el uso de sistemas de inteligencia artificial como instrumentos de asistencia, en la medida en que contribuyan a optimizar el trabajo cotidiano y a mantener la independencia decisional (Dhungel & Heine, 2024).

La digitalización también abre la puerta a un acceso más transparente a la información judicial, en tanto que permite la trazabilidad de los procesos y la consulta en línea de expedientes y resoluciones. Esta apertura puede reforzar la rendición de cuentas y la confianza ciudadana en la justicia, al facilitar que los ciudadanos comprendan mejor cómo y por qué se adoptan las decisiones (Hildebrandt, 2020). En este punto, la teoría de la justicia procedimental es esclarecedora, pues sostiene que la legitimidad de las instituciones depende no solo del resultado final, sino de la percepción de que el proceso ha sido claro, equitativo y participativo (van den Bos, 2001). En la medida en que la digitalización contribuye a explicar los procedimientos y a ponerlos al alcance de la ciudadanía, se incrementa la aceptación social de las decisiones judiciales. Sin embargo, este beneficio solo se materializa si los operadores humanos asumen el compromiso de diseñar interfaces comprensibles, de traducir los resultados técnicos a un lenguaje jurídico claro y de garantizar que la información sea realmente accesible para todos, evitando nuevas

brechas digitales.

Una ventaja adicional es la posibilidad de que los sistemas digitales evolucionen mediante aprendizaje continuo, de modo que, a diferencia de los procesos burocráticos tradicionales que permanecen rígidos en el tiempo, las soluciones tecnológicas pueden adaptarse y perfeccionarse conforme se incorporan nuevos datos. Este dinamismo representa una oportunidad para que la justicia penal sea más sensible a los cambios sociales y a las nuevas modalidades delictivas. Sin embargo, el aprendizaje automático no puede avanzar sin la retroalimentación del factor humano, que aporta la evaluación crítica de los resultados, la introducción de criterios normativos y la corrección de sesgos detectados en la práctica (Gohel, Singh & Mohanty, 2021). Lejos de ser prescindible, el ser humano se convierte en garante de que la mejora continua de los sistemas esté alineada con los valores constitucionales y con los principios del derecho penal.

Ahora bien, este panorama de beneficios no debe ocultar los riesgos que acompañan a la automatización de decisiones en materia penal. Uno de los más discutidos es la discriminación algorítmica. La evidencia empírica muestra que los sistemas entrenados con datos históricos tienden a reproducir y amplificar sesgos de género, raza o clase, perpetuando desigualdades estructurales bajo la apariencia de neutralidad tecnológica. Investigaciones pioneras como la de Angwin et al. (2016) sobre el sistema COMPAS en Estados Unidos revelaron que este sobreestimaba el riesgo de reincidencia de las personas afroamericanas en comparación con las blancas. Del mismo modo, Buolamwini y Gebru (2018) demostraron que sistemas comerciales de reconocimiento facial arrojaban tasas de error significativamente mayores en la identificación de mujeres racializadas. Estos hallazgos cuestionan la pretendida objetividad de la IA y evidencian que los algoritmos no son ajenos a los contextos sociales de desigualdad en los que se desarrollan (Eubanks, 2018; Leavy, O'Sullivan, & Siapera, 2020).

A ello se suma otro desafío crítico, la opacidad algorítmica, que atraviesa a numerosos modelos de predicción judicial concebidos como "cajas negras", en las que los procesos internos de cálculo no resultan comprensibles para los propios operadores que los aplican. Esta opacidad limita la posibilidad de auditar decisiones,

dificulta el ejercicio del derecho de defensa y debilita la rendición de cuentas institucional (Pasquale, 2015; Zerilli et al., 2019). En el ámbito penal, donde las decisiones afectan directamente a la libertad y dignidad de las personas, la falta de explicabilidad adquiere una gravedad particular. En este sentido, autores como Floridi et al. (2018) y Sartor (2020) han advertido sobre la necesidad de avanzar hacia una inteligencia artificial explicable (explainable AI) que permita comprender las razones detrás de cada recomendación o predicción, garantizando así el control humano significativo.. Esta exigencia cobra relevancia porque, como sostiene Adamson (2022), la mayoría de los sistemas de IA de alta complejidad permanecen como "cajas negras" cuya explicación solo puede formularse de manera post-hoc, lo que implica ofrecer inferencias aproximadas de su funcionamiento en lugar de descripciones veraces y completas. Ello plantea límites sustantivos a la fiabilidad de la explicabilidad, especialmente en contextos donde están en juego derechos fundamentales.

Desde una perspectiva más aplicada, Gohel, Singh y Mohanty (2021) destacan que la XAI se ha consolidado como un campo de investigación prioritario en sectores críticos como la justicia penal, la salud y la seguridad, precisamente porque la confianza pública en estas tecnologías depende de que los sistemas puedan responder a preguntas de tipo "por qué" y "cómo" respecto de cada decisión. En el ámbito judicial, esto resulta esencial para evitar errores o sesgos que afecten de manera desproporcionada a determinados grupos sociales y para dotar de legitimidad a decisiones que impactan directamente sobre la libertad de las personas.

Por otra parte, Leben (2023) enfatiza que las explicaciones no deben limitarse a la descripción técnica de los factores considerados por un modelo, sino que tienen un valor normativo adicional en tanto proporcionan evidencia de equidad. En su propuesta, los contraejemplos o explicaciones contrafactuales permiten mostrar tanto qué cambios plausibles habrían conducido a una decisión favorable (prueba positiva de equidad) como qué atributos irrelevantes no deberían haber afectado el resultado (prueba negativa de equidad). Esta dimensión conecta directamente la explicabilidad con la justicia procedimental y con el derecho de las personas a

comprender las razones de las decisiones automatizadas que las afectan.

En este punto, el análisis de Miró-Llinares (2020) en el ámbito policial resulta especialmente ilustrativo. El autor señala que la policía predictiva ha sido recibida con posiciones extremas: mientras que para algunos representa una utopía capaz de anticipar delitos, optimizar recursos y corregir sesgos humanos, para otros constituye una distopía marcada por la vigilancia masiva, la erosión de libertades y la reproducción de desigualdades estructurales. Frente a estas posturas opuestas, el autor defiende una visión crítica, informada y realista que evite tanto la tecnofilia ingenua como la tecnofobia paralizante, situando al ser humano y a la función policial legítima en el centro de la ecuación algorítmica. Este enfoque reconoce que las tecnologías no son neutras y que su implementación debe evaluarse a la luz de evidencias científicas, de su impacto en los derechos fundamentales y de su inserción en un marco democrático.

Estos problemas no son meramente técnicos, sino también normativos y epistemológicos. Tal como señalan Završnik (2020) y Brownsword y Harel (2019), el uso de algoritmos predictivos en la justicia penal no solo plantea la cuestión de su eficiencia, sino que transforma las lógicas tradicionales del derecho penal, desplazando el énfasis desde la culpabilidad individual hacia la predicción de comportamientos estadísticamente probables. Este cambio de paradigma implica riesgos de erosión del principio de culpabilidad y del principio de proporcionalidad en la imposición de penas. La justicia, concebida históricamente como un ejercicio de interpretación prudencial, se ve tensionada por la lógica probabilística de los modelos algorítmicos.

En respuesta a estos riesgos, se han formulado marcos regulatorios y principios éticos que buscan garantizar un uso responsable de la IA en el ámbito judicial. Entre ellos destacan la Carta Ética Europea sobre el uso de la Inteligencia Artificial en los Sistemas Judiciales de la CEPEJ (2018), que establece principios de transparencia, no discriminación, responsabilidad y control humano, y la propuesta de Reglamento Europeo de Inteligencia Artificial (AI Act), que clasifica las aplicaciones judiciales como de alto riesgo y exige requisitos estrictos de calidad de datos, supervisión humana y explicabilidad (European Commission, 2021). Estas iniciativas reflejan un

esfuerzo institucional por dar respuesta a los desafíos éticos y normativos derivados de la digitalización de la justicia. Tal como se ha analizado en estudios recientes, la respuesta europea frente a los sesgos discriminatorios de la IA revela un compromiso progresivo con el diseño de marcos normativos que integren principios de equidad y justicia social (Galdon Clavell & Frowd, 2015).

Este esfuerzo regulador se inserta en un debate más amplio sobre los retos epistemológicos de la justicia algorítmica. La creciente dependencia de datos para fundamentar decisiones judiciales plantea interrogantes sobre la validez del conocimiento producido por algoritmos y sobre la manera en que este conocimiento se articula con el razonamiento jurídico tradicional. Como señala Castro-Toledo (2022), los algoritmos funcionan como modelos de la realidad y no como su reflejo exacto, del mismo modo que los mapas representan el territorio sin llegar a reproducirlo plenamente. Esta naturaleza representacional implica que los algoritmos pueden contener errores o sesgos que afecten su precisión y su capacidad para describir fielmente los fenómenos sociales que pretenden modelar. Además, el autor subraya que toda solución algorítmica responde a fines humanos específicos, lo que introduce inevitablemente consideraciones éticas, jurídicas y políticas en su diseño y aplicación. Por ello, advierte que la transformación algorítmica del sistema de justicia penal debe entenderse como un proceso de modernización institucional que, si bien persigue la eficiencia y la mejora en la toma de decisiones, no puede desligarse de una reflexión crítica sobre sus límites, riesgos y consecuencias.

4. Marco normativo y ético de la utilización de la inteligencia artificial y la digitalización en el sistema de justicia penal.

Es precisamente en este contexto de avance y transformación de la inteligencia artificial en el Sistema de justicia penal donde han surgido profundas reflexiones acerca de su regulación y de los principios éticos que deben orientar su implementación. La progresiva incorporación de algoritmos en los distintos procesos ha abierto un espacio de debate en torno a los aspectos y principios que deben ser regulados para garantizar una adopción segura de esta tecnología (Binns, 2018; Cath, 2018; Miró-Llinares, 2025).

Si bien la inteligencia artificial ofrece oportunidades significativas para agilizar trámites, reducir la carga de trabajo de jueces y fiscales y minimizar ciertos sesgos humanos en la toma de decisiones, su uso también conlleva importantes desafíos. Entre los principales riesgos se encuentran la falta de transparencia en los algoritmos utilizados, la posible reproducción o amplificación de sesgos discriminatorios preexistentes en los datos y la dificultad de establecer mecanismos de rendición de cuentas cuando una decisión automatizada resulta perjudicial para una persona (Veale & Binns, 2017, Cath, 2018; Zedner, 2007).

Ante estos desafíos, se han desarrollado diversas normativas a nivel internacional para regular la inteligencia artificial en sectores críticos, incluido el ámbito judicial. La Unión Europea (UE) ha asumido un papel de liderazgo en la regulación de la inteligencia artificial, impulsado un marco normativo para regular el desarrollo y uso de la inteligencia artificial (Council of Europe, 2024), con el objetivo de garantizar su aplicación ética, segura y alineada con los valores fundamentales europeos. La creciente presencia de sistemas de inteligencia artificial en sectores críticos, como la justicia penal, la salud o la seguridad, ha generado la necesidad de establecer reglas claras que minimicen riesgos y protejan los derechos fundamentales de los ciudadanos (European Commission, 2021).

Desde un punto de vista legal, la regulación de la inteligencia artificial en Europa no solo se articula a través de normativas específicas, sino que también se integra con marcos jurídicos existentes, como el Reglamento General de Protección de Datos

(RGPD), que establece restricciones sobre el tratamiento automatizado de datos personales y refuerza los derechos de los ciudadanos ante decisiones basadas en algoritmos (European Parliament & Council of the European Union, 2016). Además, la Carta de los Derechos Fundamentales de la UE y el Convenio Europeo de Derechos Humanos juegan un papel clave en la delimitación del uso de tecnologías algorítmicas, especialmente en contextos que pueden afectar la privacidad, la libertad y la no discriminación (Council of Europe, 2021).

Este nuevo contexto requiere un sólido marco jurídico y ético que garantice el respeto a los derechos humanos y la integridad de los procesos. A continuación, se examinan los principales instrumentos normativos y orientaciones éticas relevantes, desde la regulación europea hasta principios internacionales, analizando su contenido, objetivos, principios y aportes clave para el uso de la IA en el sistema de justicia penal.

4.1. Reglamento Europeo sobre Inteligencia artificial (IA ACT).

En esta apuesta europea por regularizar la implementación de la inteligencia artificial, uno de los hitos más importantes es la Propuesta de Reglamento de inteligencia artificial, presentada por la Comisión Europea en abril de 2021, entrando en vigor el 1 de agosto de 2024, tras su publicación en el Diario Oficial de la Unión Europea el 12 de julio de 2024. Este reglamento, también conocido como Ley de Inteligencia artificial (o en inglés, *AI ACT*), establece un enfoque basado en el riesgo, clasificando los sistemas de Inteligencia artificial en diferentes categorías según su impacto potencial en la sociedad. Surge en un contexto de rápida expansión de sistemas de IA en múltiples sectores y de una conciencia creciente sobre sus riesgos, frente a una ausencia de reglas específicas suficientes en la legislación vigente. La Unión Europea busca armonizar las normas sobre IA en todo su territorio, evitando la fragmentación del mercado digital europeo y posicionándose como líder global en la regulación de una inteligencia artificial fiable y centrada en el ser humano. Los objetivos principales del IA Act incluyen fomentar el desarrollo y adopción de sistemas de IA seguros y confiables por parte de actores públicos y privados, garantizar el respeto de los derechos fundamentales en la UE, y a la vez impulsar la innovación y la competitividad tecnológica europea en el ámbito de la

inteligencia artificial. La normativa se aplica en ámbitos cubiertos por el Derecho de la UE (con exclusiones como usos militares), tiene efecto directo en todos los Estados miembros, y emplea un enfoque basado en el riesgo: a mayor riesgo que entrañe un sistema de inteligencia artificial, más estrictas son las obligaciones impuestas. De este modo, se pretende equilibrar los beneficios potenciales de la inteligencia artificial con la protección ante sus posibles perjuicios. En este sentido, el Reglamento de Inteligencia Artificial define cuatro niveles de riesgo para las aplicaciones que pueden tener estas herramientas, estableciendo para cada categoría distintos requisitos y restricciones proporcionales, pudiendo entenderse como una pirámide de riesgos (Figura 1). Esta representación gráfica permite visualizar la jerarquía normativa, situando en la cúspide el riesgo inaceptable y, en la base, el riesgo mínimo, correspondiente a los sistemas con menor impacto potencial, pasando por los niveles intermedios de riesgo alto y riesgo limitado.

Figura 1.

Clasificación de niveles de riesgo según la AI Act.

Fuente: Elaboración propia a partir del Reglamento de Inteligencia Artificial.

Esta representación es una síntesis que refleja la arquitectura normativa específica establecida en el Reglamento. Este enfoque escalonado de gestión del riesgo adapta las obligaciones legales a la intensidad y probabilidad del impacto que puede generar un sistema de IA, tal y como se establece en sus disposiciones generales y

en los artículos 5, 6 y 52, así como en el considerando 26.

En el nivel de riesgo inaceptable se encuentran aquellos sistemas de IA prohibidos por su incompatibilidad con los valores de la Unión y su potencial para producir daños graves e irreversibles. Entre las prácticas incluidas figuran la manipulación cognitiva subliminal, la explotación de vulnerabilidades derivadas de la edad, discapacidad u otras condiciones de especial protección, la clasificación social por parte de autoridades y la identificación biométrica remota en tiempo real en espacios de acceso público con fines policiales, salvo en supuestos excepcionales regulados de manera estricta. Estas prohibiciones responden a la necesidad de prevenir usos de la IA que comprometan la autonomía individual, la igualdad y la dignidad humana.

El nivel de riesgo alto comprende los sistemas de IA cuyo funcionamiento puede afectar de manera significativa a la seguridad o a los derechos fundamentales de las personas. Se incluyen, por un lado, los sistemas que forman parte de productos o componentes de seguridad regulados por normativa sectorial de la Unión, como maquinaria, vehículos o dispositivos médicos, y por otro, aquellos empleados en sectores de especial relevancia social, como la educación, el empleo, la gestión de infraestructuras críticas, la aplicación de la ley, el control de fronteras y la administración de justicia. Estos sistemas deben cumplir exigencias estrictas que abarcan la gestión y mitigación de riesgos, el uso de datos de entrenamiento de alta calidad, la elaboración de documentación técnica detallada, el mantenimiento de registros, la supervisión humana constante y la superación de procedimientos de evaluación de conformidad previos a su comercialización o puesta en servicio.

El nivel de riesgo limitado se refiere a sistemas que, sin presentar amenazas graves, requieren salvaguardias específicas en materia de transparencia para evitar que los usuarios sean inducidos a error. Ejemplos representativos incluyen asistentes conversacionales, generadores de contenido sintético como los denominados *deepfakes* y sistemas que imitan la interacción humana. En estos casos, la normativa exige informar de manera clara y accesible a las personas usuarias de que están interactuando con un sistema de IA o de que el contenido ha sido generado artificialmente, con el fin de preservar la capacidad de decisión informada.

El nivel de riesgo mínimo o nulo engloba la mayoría de los sistemas de uso general que no suponen riesgos relevantes para los derechos fundamentales ni para la seguridad. Para este grupo no se imponen requisitos jurídicos vinculantes adicionales, aunque el legislador europeo fomenta la adopción voluntaria de códigos de conducta, estándares técnicos y prácticas responsables que promuevan un desarrollo ético y seguro de la inteligencia artificial.

En el ámbito judicial, esta normativa reconoce tanto las oportunidades como los riesgos derivados de la implementación de la IA, catalogando como sistemas de alto riesgo aquellas herramientas que puedan incidir significativamente en los derechos fundamentales. En particular, el AI Act regula con especial atención las aplicaciones de IA en la predicción de sentencias, la evaluación de riesgos penales y la asistencia a la decisión judicial (Entcheva & Mazilescu, 2024).

En el contexto español, Cotino Hueso y Simón Castellano (2024) han publicado, con gran acierto, el *Tratado sobre el Reglamento de Inteligencia Artificial*, una guía exhaustiva, profunda y actualizada sobre el Reglamento (UE) 2024/1689 que combina análisis técnico jurídico, reflexiones éticas y orientación práctica para su implementación y cumplimiento. Esta obra se erige como una referencia esencial para comprender el alcance, las aplicaciones y los desafíos del marco regulador de la inteligencia artificial en Europa, prestando especial atención a sus implicaciones en el ámbito jurídico penal.

En este marco, diversos autores del volumen examinan de manera directa la incidencia del IA Act en contextos de justicia penal, identificando tanto sus fortalezas como sus limitaciones. Escajedo (2024) destaca que la prohibición del reconocimiento biométrico remoto "en tiempo real" en espacios públicos para fines de garantía del cumplimiento de la ley, prevista en el artículo 5.1.h, se ve matizada por un conjunto de excepciones tasadas incluidas en los apartados 2 a 8 del mismo precepto. Estas excepciones abarcan la búsqueda de víctimas, la prevención de amenazas graves y la persecución de los delitos enumerados en el Anexo II. Aunque esta técnica normativa pretende limitar el uso policial a supuestos de elevada

gravedad, la autora advierte que la amplitud y casuística de las excepciones puede diluir la efectividad de la prohibición, generando inseguridad jurídica y dificultades interpretativas (pp. 218–222).

Miró Llinares y Santisteban Galarza (2024) analizan la regulación de la denominada policía predictiva y distinguen entre usos prohibidos, que se producen cuando la evaluación de la peligrosidad criminal se fundamenta exclusivamente en el perfilado automatizado, y usos considerados de alto riesgo, que se caracterizan porque la herramienta desempeña una función auxiliar y no determinante, lo que obliga a cumplir las obligaciones reforzadas previstas en el Título III y el Anexo III. Estos autores defienden que este enfoque basado en el riesgo constituye una solución proporcional, aunque advierten sobre la dificultad práctica de clasificar un sistema en una u otra categoría. También llaman la atención sobre el riesgo de que se produzca un desplazamiento hacia prácticas que el Reglamento prohíbe, así como sobre la incidencia de sesgos de automatización en la toma de decisiones (pp. 319, 333).

En el plano procesal y contencioso, Ortega Giménez (2024) señala que los incumplimientos del IA Act pueden ser alegados ante los tribunales, incluso en materia de responsabilidad extracontractual derivada del uso defectuoso de un sistema de inteligencia artificial en contextos judiciales o policiales. Aunque este aspecto no está diseñado específicamente para el ámbito penal, constituye una vía relevante para ejercer el control jurisdiccional de la legalidad y de la conformidad de los usos de inteligencia artificial con el Reglamento (p. 133).

Sempere (2024) estudia el régimen sancionador y sostiene que el ejercicio de la potestad sancionadora debe respetar los principios propios del derecho penal, tales como legalidad, irretroactividad, tipicidad, culpabilidad, proporcionalidad, prescripción y prohibición de doble sanción, conforme a la Ley 40/2015 de Régimen Jurídico del Sector Público y a la doctrina constitucional. No obstante, advierte que el artículo 71 del Reglamento deja en manos de los Estados miembros la concreción de este régimen, incluida la decisión sobre si se sanciona a las administraciones públicas, lo que puede generar asimetrías, debilitar la eficacia coercitiva y retrasar la regulación de elementos esenciales como la prescripción y la tipificación (pp.

875–877).

Este debate académico identifica áreas de consenso, entre ellas la necesidad de establecer estándares reforzados en materia penal, la prohibición de evaluaciones basadas exclusivamente en procesos automatizados para determinar la peligrosidad criminal y la exigencia de controles estrictos en el reconocimiento biométrico remoto. También expone aspectos que pueden resultar problemáticos, como el alcance real de las prohibiciones frente a las excepciones, la complejidad de trazar límites claros en materia de policía predictiva y la fragmentación del régimen sancionador. Mientras algunos autores consideran que el enfoque basado en riesgos del IA Act representa un avance equilibrado, otros sostienen que su eficacia dependerá de interpretaciones restrictivas, del control judicial y del compromiso de los Estados miembros con la uniformidad y las garantías en su aplicación en el ámbito penal. Aunque el Reglamento (UE) 2024/1689 ya ha entrado en vigor, su aplicación plena todavía no se encuentra completamente instaurada, dado que el propio texto normativo prevé un calendario progresivo que se extiende hasta 2027 para la entrada en vigor de determinadas obligaciones. Este despliegue escalonado, si bien ofrece a los operadores jurídicos y tecnológicos un margen de adaptación, también evidencia ciertas incertidumbres y potenciales inconvenientes, entre los que se incluyen la necesidad de una interpretación uniforme en todos los Estados miembros y la capacidad de reacción frente a supuestos no previstos expresamente. Sin embargo, lo que sí parece que quedará instaurado es la incorporación del factor humano como elemento indispensable en la supervisión, validación y control de los sistemas de inteligencia artificial en aquellos riesgo más altos. En los próximos años resultará determinante analizar cómo esta normativa se acomoda al contexto jurídico penal europeo y cómo afronta la irrupción de nuevas herramientas y metodologías de inteligencia artificial, cuyo desarrollo acelerado podría tensionar o incluso superar los marcos regulatorios inicialmente diseñados.

4.2. Convención Marco sobre Inteligencia Artificial del Consejo de Europa.

De forma paralela al Reglamento de Inteligencia Artificial de la Unión Europea (AI Act), el Consejo de Europa adoptó en el año 2024 la Convención Marco sobre la Inteligencia artificial, los Derechos Humanos, la Democracia y el Estado de Derecho

(CETS No. 225), consolidándose como el primer tratado internacional jurídicamente vinculante en el ámbito de la inteligencia artificial . Esta convención fue adoptada en Vilna el 5 de septiembre de 2024 y está abierta tanto a los Estados miembros del Consejo de Europa como a terceros Estados que compartan sus principios y valores (Council of Europe, 2024).

El objetivo central de este instrumento jurídico consiste en garantizar que todas las actividades desarrolladas en el ciclo de vida de los sistemas de inteligencia artificial sean plenamente coherentes con los derechos humanos, los principios democráticos y el Estado de Derecho. Para ello, se exige a los Estados parte la adopción de medidas legislativas, administrativas o de otra índole que aseguren la implementación efectiva de los principios y obligaciones establecidos en la convención (Council of Europe, 2024, art. 1).

Entre los pilares fundamentales de la convención se destaca la adopción de un enfoque basado en el riesgo, que obliga a los Estados a identificar, evaluar, prevenir y mitigar los posibles efectos adversos que la implementación de sistemas de IA pueda generar en los derechos fundamentales y en el orden democrático (Council of Europe, 2024, art. 16). Asimismo, se refuerza la importancia de garantizar la transparencia, la supervisión y la rendición de cuentas, estableciendo que los sistemas de IA deben ser auditables y trazables, especialmente en aquellos contextos donde puedan afectar derechos esenciales como el acceso a la justicia o la presunción de inocencia (Council of Europe, 2024, arts. 8–9, 14–15).

La Convención también impone obligaciones específicas dirigidas a preservar la dignidad humana, la autonomía individual y la igualdad, haciendo énfasis en la protección de grupos vulnerables, como las personas con discapacidad o los menores de edad (Council of Europe, 2024, arts. 7, 10, 18). En esta línea, se promueve la adopción de mecanismos de reparación accesibles y efectivos, que incluyan el derecho de los individuos a recibir información comprensible sobre el funcionamiento de los sistemas automatizados que les afecten y a impugnar las decisiones derivadas de su uso (Council of Europe, 2024, arts. 14–15).

Un aspecto especialmente relevante de la convención es la disposición que habilita

a los Estados a establecer una moratoria, prohibición u otras medidas apropiadas respecto de determinados usos de la IA que puedan resultar incompatibles con los derechos humanos, la democracia o el Estado de Derecho (Council of Europe, 2024, art. 16.4). Esta cláusula refleja el principio de precaución y refuerza la legitimidad de los Estados para limitar o impedir desarrollos tecnológicos contrarios a los valores fundamentales del orden internacional democrático.

Por otra parte, la Convención establece un mecanismo de seguimiento institucional mediante la Conferencia de las Partes, órgano encargado de facilitar la aplicación efectiva del tratado, promover la cooperación internacional, evaluar el impacto de las reservas, formular recomendaciones interpretativas y fomentar el intercambio de información sobre desarrollos relevantes (Council of Europe, 2024, art. 23). También se prevé la designación de mecanismos nacionales de supervisión independientes e imparciales, dotados de recursos suficientes para garantizar el cumplimiento de las obligaciones del tratado (Council of Europe, 2024, art. 26).

En suma, la Convención Marco del Consejo de Europa representa un complemento fundamental al AI Act, al ofrecer un marco normativo con alcance paneuropeo e internacional centrado en la protección de los derechos fundamentales. Su adopción refuerza el liderazgo europeo en la gobernanza ética y jurídica de la inteligencia artificial , y sienta las bases para un modelo de regulación global basado en principios democráticos y en la protección de la dignidad humana.

4.3. Otros textos éticos y normativos internacionales relevantes.

4.3.1. Directrices Éticas para una Inteligencia Artificial Fiable del Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia artificial (HLEG).

El marco normativo de la Unión Europea también se fundamenta en principios previos desarrollados, como las Directrices Éticas para una Inteligencia Artificial Fiable, publicadas en 2019 por el Grupo de Expertos de Alto Nivel en Inteligencia artificial (High-Level Expert Group on AI, 2019). Su propósito es garantizar que la inteligencia artificial sea confiable y beneficie a la sociedad sin poner en compromiso los derechos fundamentales. En el documento se establecen tres

dimensiones clave para su correcta integración: la legalidad, la ética y la robustez técnica. Pero no solo se encarga de garantizar el marco regulatorio por el que debe garantizarse la legalidad de la inteligencia artificial dentro del ordenamiento jurídico europeo, sino que también pone un énfasis particular en la necesidad de una Inteligencia artificial ética y robusta.

En este sentido, su regulación se estructura en una serie de principios clave:

- El primero de ellos es el respeto por la autonomía humana, lo que implica que la Inteligencia artificial debe estar diseñada para complementar y potenciar las capacidades humanas sin sustituirlas ni coaccionar decisiones.
- El segundo principio es la prevención del daño, asegurando que los sistemas de IA operen de manera segura y que existan mecanismos efectivos para minimizar sesgos y riesgos. A esto se suma el principio de equidad, que exige que la IA no reproduzca desigualdades y que garantice un acceso equitativo a sus beneficios.
- Finalmente, la explicabilidad y la transparencia constituyen un pilar fundamental en la regulación, ya que permiten que los sistemas sean comprensibles, auditables y sujetos a mecanismos de rendición de cuentas.

Una vez establecidas las directrices principales que deben cumplirse, las cuales determinan una serie de requisitos, siete en concreto, es fundamental respetarlas para que los sistemas de inteligencia artificial sean considerados confiables.

En primer lugar, y en cuanto al factor humano, se considera que la supervisión humana es un elemento central, ya que garantiza la posibilidad de intervención en los procesos de toma de decisiones, especialmente en sectores críticos como la salud, la justicia y la seguridad. Además, la IA debe desarrollarse con solidez técnica y seguridad, lo que implica que los sistemas sean resistentes a ataques y fallos, asegurando su fiabilidad operativa. Otro requisito clave es la gestión de la privacidad y los datos, garantizando el cumplimiento del Reglamento General de Protección de Datos (RGPD) y protegiendo la información personal mediante medidas de anonimización y minimización de datos. Asimismo, la transparencia es un aspecto crucial dentro del marco regulatorio, exigiendo que los procesos de toma de

decisiones sean comprensibles y que se implementen mecanismos de trazabilidad y auditoría para evitar la opacidad algorítmica. También se establece la necesidad de promover la diversidad, la no discriminación y la equidad, evitando sesgos en los modelos de IA y asegurando su desarrollo y uso inclusivo. Además, se reconoce la importancia del impacto social y ambiental, fomentando la creación de sistemas sostenibles y evaluando su repercusión en la sociedad y la democracia. Finalmente, se subraya la responsabilidad y la rendición de cuentas, lo que implica la implementación de mecanismos de supervisión y la posibilidad de revisar o impugnar las decisiones de los sistemas de IA cuando sea necesario.

Este marco normativo ha tenido un impacto significativo en la regulación europea vigente, particularmente en el Reglamento Europeo de Inteligencia artificial (AI Act), el cual introduce una clasificación de los sistemas de IA basada en su nivel de riesgo. En este sentido, el AI Act impone restricciones estrictas a aquellas aplicaciones consideradas de alto riesgo, tales como los sistemas de reconocimiento facial en espacios públicos y los algoritmos utilizados en procesos de selección y contratación laboral. Asimismo, el impacto de estas directrices también se ha materializado en normativas como el Reglamento General de Protección de Datos (RGPD), que refuerza la necesidad de garantizar la privacidad en los sistemas de IA y establece derechos fundamentales, entre los cuales destaca la explicabilidad de las decisiones automatizadas.

En el ámbito judicial, las directrices del HLEG son relevantes para orientar un uso responsable de la IA. Al exigir supervisión humana y rendición de cuentas, se asegura que decisiones asistidas por algoritmos (por ejemplo, sistemas que ayudan a jueces o a funcionarios públicos) siempre puedan ser revisadas y atribuibles a una autoridad humana. Asimismo, el principio de no discriminación implica que las herramientas de IA utilizadas en ámbitos como la justicia penal o la administración (p. ej., algoritmos de evaluación de riesgo de reincidencia, sistemas de reparto de casos) no deben perpetuar sesgos ni violar la igualdad ante la ley. Así mismo, las directrices enfatizan el respeto a las leyes y derechos vigentes, lo cual exige que cualquier sistema automatizado empleado en decisiones públicas cumpla con garantías como el derecho a un juicio justo y la protección de datos personales

(Comisión Europea, 2019). Por último, la transparencia y explicabilidad son cruciales en contextos judiciales: si se emplea IA para apoyar decisiones, es imprescindible poder explicar cómo llegó a un resultado, de modo que las partes afectadas puedan entender y, si es necesario, impugnar la decisión. En suma, los principios de la IA fiable (especialmente legalidad, equidad, transparencia y supervisión humana) funcionan como salvaguardas para que la IA en la justicia no vulnere derechos fundamentales ni mine la confianza en el sistema jurídico.

4.3.2. Recomendación de la Organización para la Cooperación y el Desarrollo Económicos (OECD) sobre inteligencia artificial.

La Recomendación del Consejo de la Organización para la Cooperación y el Desarrollo Económicos (OECD) sobre la Inteligencia artificial (OECD/LEGAL/0449), adoptada inicialmente en 2019 y modificada por última vez en 2024, constituye el primer instrumento intergubernamental orientado a establecer principios para una gobernanza responsable, transparente y centrada en los derechos humanos del desarrollo y uso de sistemas de inteligencia artificial. A pesar de su carácter no vinculante, la Recomendación ha sido avalada por una amplia coalición de Estados miembros y no miembros, lo que refuerza su legitimidad normativa y su capacidad de influir en la formulación de políticas públicas (OECD, 2024).

La Recomendación se articula en torno a cinco principios fundamentales de gobernanza ética: (1) promoción del crecimiento inclusivo, el desarrollo sostenible y el bienestar; (2) respeto al Estado de derecho, los derechos humanos y los valores democráticos; (3) transparencia y explicabilidad; (4) solidez técnica, seguridad y protección; y (5) responsabilidad de los actores implicados (OECD, 2024, pp. 8–9). Estos principios han servido de base sobre la cual los Estados y otros actores relevantes han comenzado a desarrollar marcos regulatorios, políticas públicas, y mecanismos de supervisión para mitigar los riesgos y maximizar los beneficios de la IA.

Uno de los aspectos más relevantes de la Recomendación es la exigencia de que todo desarrollo y aplicación de IA respete de manera efectiva el Estado de derecho y los derechos fundamentales, con especial énfasis en la protección frente a la

discriminación, la salvaguarda de la autonomía individual y el derecho a un juicio justo. En el ámbito penal, donde se deciden cuestiones relativas a la libertad personal y la imposición de penas, este principio cobra especial trascendencia (OECD, 2024, p. 8). Asimismo, la Recomendación subraya la necesidad de garantizar la transparencia, explicabilidad y trazabilidad de los sistemas algorítmicos, especialmente cuando sus decisiones tienen consecuencias significativas sobre la vida de las personas. Esta exigencia tiene una aplicación directa en contextos judiciales y penitenciarios, donde resulta imperativo que los operadores jurídicos, las personas afectadas y las instancias de supervisión puedan entender y revisar el funcionamiento de los sistemas de IA utilizados (OECD, 2024, p. 8-9). La opacidad algorítmica es incompatible con los principios de publicidad procesal, derecho a la defensa y control jurisdiccional de la legalidad. Por otro lado, el principio de responsabilidad y rendición de cuentas exige que los actores públicos y privados que desarrollan implementan o supervisan sistemas de IA respondan por los efectos derivados de su uso. En el marco penal, esto implicaría que las decisiones asistidas por IA deben estar sometidas a revisión humana efectiva, y que los mecanismos institucionales deben garantizar la posibilidad de reparación en caso de error, sesgo o vulneración de derechos (OECD, 2024, pp. 9). Aunque no de manera explícita, la Recomendación incorpora además un enfoque normativo basado en el riesgo, particularmente dentro del principio de responsabilidad y rendición de cuentas, el cual establece que los actores de IA deben aplicar un "enfoque sistemático de gestión de riesgos en cada fase del ciclo de vida del sistema de IA, de forma continua y coherente con el estado del arte" (OECD, 2024, p. 9). Esta aproximación permite adaptar el nivel de exigencia regulatoria en función de la magnitud del impacto potencial de cada sistema. Este tipo de principios pueden resultar relevantes en el contexto del ámbito judicial, donde los riesgos derivados de una automatización mal calibrada, por ejemplo, la utilización de datos sesgados, la discriminación estructural o la descontextualización de la conducta individual, pueden derivar en violaciones graves de los derechos a la libertad, la igualdad ante la ley o la presunción de inocencia.

La Recomendación de la OECD, aunque no se formule como un instrumento jurídicamente vinculante, permite proporcionar un marco ético y político robusto

para orientar la gobernanza de la inteligencia artificial en contextos de alta sensibilidad institucional, como el sistema de justicia penal. Su enfoque basado en principios, su aplicabilidad transnacional y su énfasis en la compatibilidad con los derechos fundamentales la convierten en un referente indispensable para el diseño de políticas públicas que busquen conciliar la innovación tecnológica con la legitimidad democrática en la era digital.

4.3.3. Recomendación sobre Ética e Inteligencia artificial de la UNESCO.

La Recomendación sobre la Ética de la Inteligencia artificial , adoptada el 23 de noviembre de 2021 por la Conferencia General de la UNESCO, aborda las implicaciones éticas del desarrollo y uso de la inteligencia artificial. Su objetivo principal es orientar la gobernanza de la IA desde una perspectiva centrada en la dignidad humana, los derechos humanos y el desarrollo sostenible (UNESCO, 2022).

La Recomendación se estructura en torno a una serie de valores y principios éticos universales, entre los que destacan la promoción de los derechos humanos, la igualdad de género, la inclusión, la diversidad, la sostenibilidad y la equidad. Además, introduce principios operativos fundamentales para el diseño, implementación y supervisión de los sistemas de IA, como la proporcionalidad, la transparencia, la rendición de cuentas, la supervisión humana y la protección de datos personales (UNESCO, 2022, pp. 17–24). Uno de los aportes más relevantes del documento es la propuesta de evaluación del impacto ético de los sistemas de IA, entendida como una herramienta transversal para anticipar, prevenir y mitigar riesgos sobre derechos fundamentales, la democracia y el estado de derecho. Esta evaluación debe ser multidisciplinar, inclusiva y pública, y acompañar todas las fases del ciclo de vida de los sistemas automatizados, especialmente cuando se trata de su uso en servicios públicos sensibles (UNESCO, 2022, pp. 26–27).

Desde la perspectiva del ámbito penal, la Recomendación resulta especialmente pertinente por varias razones:

- Supervisión humana y decisiones sensibles: El texto establece que las decisiones que puedan afectar a derechos fundamentales, como la libertad o la vida, no deben delegarse exclusivamente a sistemas automatizados,

garantizando que siempre exista responsabilidad humana final (UNESCO, 2022, p. 22).

- No discriminación y equidad: La Recomendación advierte sobre el potencial de los sistemas de IA para reproducir y amplificar sesgos estructurales, lo cual resulta crítico en el contexto de justicia penal, donde decisiones algorítmicas sesgadas podrían perpetuar desigualdades raciales, sociales o de género (UNESCO, 2022, pp. 20–21).
- Transparencia y explicabilidad: Se promueve que toda persona afectada por una decisión automatizada tenga derecho a conocer si esta fue tomada (total o parcialmente) por un sistema algorítmico, así como a recibir explicaciones comprensibles y a recurrir la decisión ante una autoridad competente (UNESCO, 2022, p. 22).
- Responsabilidad y auditoría: Se insta a los Estados a desarrollar mecanismos jurídicos e institucionales que permitan la rendición de cuentas de todos los actores implicados en el ciclo de vida de la IA, incluyendo medidas como auditorías, trazabilidad y mecanismos de reparación (UNESCO, 2022, p. 23).
- Fortalecimiento del poder judicial: La Recomendación señala expresamente la necesidad de capacitar al sistema judicial para evaluar el uso de IA en sus propios procedimientos, manteniendo siempre los principios de independencia judicial y control humano sobre los sistemas (UNESCO, 2022, p. 27).

Esta recomendación, aunque no resulta ser una regulación obligatoria, representa un marco ético y político de referencia internacional para orientar el desarrollo de legislaciones y políticas públicas que aseguren una implementación responsable, justa y democrática de la inteligencia artificial. En el ámbito penal, su relevancia radica en proporcionar salvaguardas éticas esenciales para evitar vulneraciones de derechos fundamentales mediante el uso de tecnologías algorítmicas, contribuyendo así a una justicia digital más legítima y centrada en el ser humano.

4.3.4. Model Rules on Impact Assessment of Algorithmic Decision-Making Systems Used by Public Administration.

El documento *Model Rules on Impact Assessment of Algorithmic Decision-Making*

Systems Used by Public Administration, publicado por el European Law Institute (ELI) en 2022, constituyen una propuesta normativa de carácter no vinculante que ofrece un marco detallado para regular el uso de sistemas de toma de decisiones algorítmicas (ADMS) en el ámbito de la administración pública. Estas reglas se conciben como una herramienta jurídica orientada a garantizar que el uso de tecnologías basadas en inteligencia artificial por parte de las autoridades se desarrolle de forma conforme a los principios del Estado de derecho, la democracia y los derechos fundamentales (European Law Institute [ELI], 2022).

Aunque como ya se ha comentado, no constituyen legislación obligatoria, su valor radica en ofrecer una guía técnica y jurídica concreta para llevar a cabo evaluaciones de impacto algorítmico (*Algorithmic Impact Assessments*, AIA) antes de implementar un sistema automatizado en procesos administrativos que puedan afectar a los ciudadanos. Las reglas parten de un enfoque escalonado basado en el riesgo, estableciendo distintos niveles de exigencia dependiendo de la peligrosidad potencial del sistema algorítmico. Así, se prevé que los sistemas de alto riesgo, como aquellos que inciden directamente en derechos fundamentales, estén sujetos a una evaluación exhaustiva y a mecanismos de control adicionales, como la revisión técnica por parte de expertos independientes o la apertura de procesos de consulta pública (ELI, 2022, arts. 6, 10–11).

El proceso de evaluación contempla aspectos esenciales como la identificación de riesgos para los derechos fundamentales, la transparencia del sistema, la calidad de los datos utilizados, la explicabilidad de las decisiones y la posibilidad de revisión humana. Igualmente, se establece la necesidad de crear una autoridad supervisora independiente con competencias para controlar el cumplimiento de las evaluaciones y para intervenir en caso de uso indebido, falta de evaluación o riesgo sistémico (ELI, 2022, arts. 15–16).

Desde la perspectiva del ámbito penal, este documento contribuye a fortalecer la legitimidad del uso de IA en contextos penales al exigir que toda herramienta automatizada de alto impacto sea objeto de un análisis previo y documentado sobre su legalidad, su proporcionalidad, y su compatibilidad con los derechos fundamentales. Tal exigencia es coherente con los estándares internacionales en

materia de justicia algorítmica, y responde a preocupaciones ampliamente documentadas en la literatura sobre sesgos algorítmicos, opacidad tecnológica y falta de trazabilidad (Crawford, 2021).

Además, el principio de explicabilidad recogido en estas reglas es esencial en el ámbito penal, donde las decisiones deben estar debidamente motivadas y justificadas. Los sistemas algorítmicos que incidan en decisiones sobre libertad, detención o condena deben estar sometidos a estrictos controles que permitan auditar su funcionamiento y asegurar que sus resultados no sustituyen, sino que complementan, el juicio humano. Asimismo, la previsión de una supervisión institucional independiente y de mecanismos de participación pública constituye un avance significativo para el control democrático de las tecnologías en la administración penal, ya que permite incorporar puntos de vista diversos y garantizar mayor transparencia en el diseño e implementación de los sistemas (ELI, 2022, arts. 11, 15).

En reseñable que este tipo de documento, aunque no tienen carácter obligatorio, representan un modelo normativo útil para los legisladores europeos y nacionales que busquen garantizar un uso responsable de la inteligencia artificial en el ámbito público, y especialmente en el sistema de justicia penal. La introducción de evaluaciones de impacto como requisito previo a la utilización de tecnologías automatizadas puede ser una herramienta eficaz para preservar los valores fundamentales del Estado de derecho en el contexto de la justicia digital.

4.3.5. Carta de los Derechos Fundamentales de la Unión Europea.

La Carta de los Derechos Fundamentales de la Unión Europea (2000) constituye un instrumento jurídico de máximo rango dentro del ordenamiento de la Unión Europea, con la misma fuerza vinculante que los Tratados. Proclamada en el año 2000 y jurídicamente exigible desde 2009, la Carta sistematiza en un único documento los derechos civiles, políticos, económicos y sociales reconocidos por los Estados miembros, estructurados en seis grandes bloques temáticos: dignidad, libertades, igualdad, solidaridad, ciudadanía y justicia (Parlamento Europeo, Consejo y Comisión Europea, 2012).

Si bien la Carta no constituye una normativa específica sobre inteligencia artificial, su relevancia en la era digital es innegable, ya que proporciona el marco jurídico fundamental para garantizar que el uso de tecnologías emergentes, como los sistemas de IA, respete los derechos fundamentales. Esta función es especialmente crítica en el proceso de digitalización del sistema penal, donde las decisiones automatizadas pueden incidir directamente sobre la libertad, la presunción de inocencia o la igualdad ante la ley.

Entre las disposiciones más relevantes para el ámbito penal digitalizado destacan, en primer lugar, el derecho a la tutela judicial efectiva y a un juez imparcial (art. 47), que establece la necesidad de que toda persona pueda acceder a un recurso ante una autoridad judicial en condiciones de imparcialidad y con las debidas garantías procesales. Asimismo, el artículo 48 reconoce el principio de presunción de inocencia y el derecho de defensa, fundamentales en cualquier procedimiento penal, y que podrían verse comprometidos en escenarios donde sistemas automatizados influyan en la valoración probatoria o en la recomendación de medidas cautelares.

De especial importancia es también el artículo 49, que consagra los principios de legalidad y proporcionalidad penal, y que prohíbe la retroactividad desfavorable de la ley penal, así como la imposición de penas desproporcionadas. En este sentido, cualquier algoritmo empleado en el ámbito penal deberá operar dentro de los límites definidos por la legalidad vigente, respetando las garantías propias del derecho penal sustantivo y procesal.

Por otro lado, la protección de datos personales (art. 8) adquiere una dimensión crítica en la digitalización del sistema penal, dado que los sistemas de IA operan habitualmente sobre grandes volúmenes de información sensible. El uso de estos datos debe regirse por los principios de licitud, finalidad y minimización, así como estar sujeto a supervisión por una autoridad independiente. La igualdad ante la ley y la prohibición de discriminación (arts. 20 y 21) obligan a una evaluación rigurosa del impacto de los sistemas algorítmicos sobre colectivos históricamente vulnerables. En la práctica, esto implica que no pueden adoptarse herramientas que reproduzcan sesgos estructurales o que deriven en resultados discriminatorios, como ha ocurrido en algunos sistemas predictivos utilizados en contextos policiales

o penitenciarios.

Finalmente, es importante indicar que esta legislación influye de manera directa en el diseño y aplicación del marco regulador europeo sobre IA, como el Reglamento Europeo de Inteligencia artificial (AI Act), que establece que todos los sistemas de IA deben respetar los derechos fundamentales reconocidos por la Unión. De esta forma, la Carta opera como un referente obligatorio para evaluar la compatibilidad jurídica de cualquier herramienta algorítmica utilizada en procesos penales.

4.4. Principios comunes en las principales normativas sobre inteligencia artificial y su relevancia en el ámbito penal.

Como se ha podido constatar, la creciente incorporación de sistemas digitales, herramientas algorítmicas y de inteligencia artificial en sectores estratégicos ha impulsado el desarrollo de un conjunto de instrumentos normativos, éticos y políticos a nivel internacional. A pesar de que presentan diferencias en cuanto a su origen institucional, alcance jurídico o grado de vinculación, todos ellos convergen en un núcleo de principios normativos fundamentales que configuran un marco común para la gobernanza ética, legal y técnica de la IA. Dichos principios insisten en que el desarrollo y la implementación de estas tecnologías deben ser compatibles con el Estado de derecho, el respeto a los derechos humanos y los valores democráticos, así como cumplir con exigencias de transparencia, control y justicia distributiva. Esta convergencia resulta especialmente relevante en el ámbito de la digitalización del sistema penal, ya que en este contexto las decisiones automatizadas inciden de forma directa sobre garantías procesales y derechos fundamentales, lo que incrementa la necesidad de un marco sólido de regulación que limite los riesgos y oriente su utilización hacia fines legítimos.

Uno de los principios transversales que más se reitera en las distintas normativas es el de la supervisión humana significativa, que establece que los sistemas de IA deben complementar, y no sustituir, la autonomía humana, sobre todo en entornos sensibles como la justicia penal. De esta forma, se exige que toda decisión automatizada esté sujeta a revisión, intervención o aprobación por parte de una autoridad responsable, lo cual permite preservar la legitimidad democrática de las

decisiones y evita la delegación ciega en un sistema opaco o autónomo (European Commission, 2019; UNESCO, 2022; ELI, 2022). Este principio se vincula estrechamente con la idea de responsabilidad institucional, que obliga a desarrolladores, operadores y usuarios de IA a rendir cuentas por los efectos derivados de su aplicación, incluyendo la existencia de mecanismos eficaces de reclamación y reparación para las personas afectadas (OECD, 2024; ELI, 2022). De este modo, la gobernanza de la IA no se limita a su diseño tecnológico, sino que se extiende a la definición de marcos de rendición de cuentas que garanticen tanto la supervisión ex ante como la responsabilidad ex post.

La transparencia, la explicabilidad y la trazabilidad constituyen otro de los pilares que atraviesan de manera unánime estas regulaciones. El carácter comprensible, auditable y justificable de los procesos algorítmicos se considera esencial para garantizar la confianza ciudadana y la legitimidad institucional, más aún en el ámbito penal, donde la opacidad algorítmica puede comprometer derechos tan básicos como la defensa, el principio de contradicción o la obligación de motivar las resoluciones judiciales (OECD, 2024; HLEG, 2019; Parlamento Europeo et al., 2012). A diferencia de otros sectores en los que la falta de explicabilidad podría limitarse a un problema de eficiencia o de reputación, en la justicia penal se trata de un requisito indispensable para la vigencia del debido proceso, lo que justifica que los instrumentos internacionales subrayen este principio como condición de posibilidad de cualquier despliegue de IA en el ámbito judicial o penitenciario.

Otro eje de convergencia normativa es la obligación de prevenir la reproducción o amplificación de sesgos estructurales y discriminatorios. Las directrices internacionales sostienen que el desarrollo y uso de sistemas de IA debe incorporar medidas específicas de identificación, mitigación y corrección de sesgos con el fin de garantizar la igualdad ante la ley y la no discriminación. En el contexto penal, esta exigencia adquiere un carácter particularmente relevante, ya que el uso de bases de datos históricas puede perpetuar patrones de discriminación que afecten de manera desproporcionada a determinados colectivos sociales, reproduciendo desigualdades estructurales en lugar de corregirlas (UNESCO, 2022; OECD, 2024). En este sentido, tanto la Recomendación de la UNESCO como las Model Rules del ELI

introducen de manera explícita el principio de evaluación previa del impacto ético y legal, estableciendo que todo sistema algorítmico de alto riesgo debe someterse a un análisis multidisciplinar y público antes de su implementación. Esta evaluación debe contemplar, de manera exhaustiva, los efectos previsibles sobre los derechos fundamentales, la proporcionalidad de las medidas, la adecuación de los mecanismos de supervisión y la existencia de recursos efectivos para las personas potencialmente afectadas (UNESCO, 2022; ELI, 2022).

Este mapa normativo existente conforma un corpus convergente que gira en torno a principios como la supervisión humana, la transparencia, la equidad, la responsabilidad y la protección de los derechos fundamentales. Este consenso ético y jurídico constituye el punto de partida imprescindible para orientar las políticas públicas en materia de regulación de la IA aplicada al sistema penal. La existencia de una base común no solo facilita la armonización entre distintos instrumentos y contextos nacionales, sino que también garantiza que la digitalización no erosione las garantías propias del Estado de derecho. Al contrario, permite reforzarlas mediante un uso responsable, transparente y centrado en la persona, consolidando así un paradigma de justicia digital que no sacrifica derechos en nombre de la eficiencia, sino que los preserva como límite y condición de toda innovación tecnológica.

CAPÍTULO 2. EL PROCESO DE TRANSFORMACIÓN TECNOLÓGICA Y EL FACTOR HUMANO EN EL SISTEMA DE JUSTICIA PENAL.

1. El rol del factor humano en la transformación tecnológica de la justicia: ¿hacia una complementariedad o una sustitución por las nuevas herramientas?.

Como se ha podido constatar, la transformación tecnológica del sistema de justicia penal no puede interpretarse únicamente como un proceso técnico o administrativo, sino que debe entenderse como una transformación estructural con profundas implicaciones políticas, sociales y éticas (Guillén Burguillos & Serrano Robles, 2024). En este contexto, resulta imprescindible reconocer el papel del factor humano, que constituye un eje central por múltiples razones y cuya consideración, aunque ya abordada en apartados anteriores, conviene destacar de manera expresa para comprender en toda su complejidad los desafíos y oportunidades que plantea la transformación digital de la justicia. En primer lugar, tanto las normativas europeas como las directrices éticas internacionales subrayan que las herramientas algorítmicas y de inteligencia artificial deben estar sujetas a supervisión, garantizando que la responsabilidad última recaiga en las personas y no en los sistemas automatizados (Busuioc, 2021; Sterz et al., 2024; UNESCO, 2021). En segundo lugar, los operadores jurídicos son quienes deben interactuar directamente con estas tecnologías, aprender a utilizarlas y desarrollar competencias que les permitan integrarlas de manera crítica en su labor profesional (Consejo General del Poder Judicial, 2022). En tercer lugar, la literatura sobre interacción humano-máquina ha mostrado que el uso de sistemas algorítmicos puede influir significativamente en las decisiones, generando fenómenos como la delegación acrítica de responsabilidades o la confianza excesiva en los resultados (Vicente & Matute, 2023). Finalmente, el factor humano no se limita a los profesionales, sino que abarca también a la ciudadanía, destinataria última de las reformas tecnológicas, de modo que la legitimidad de una justicia digitalizada dependerá en gran medida de la aceptación social de estas herramientas y de la participación ciudadana en el proceso de implementación de estas.

Uno de los aspectos más relevantes en este debate es la supervisión humana,

concebida como salvaguarda fundamental para garantizar el respeto de los derechos fundamentales en contextos donde las decisiones son asistidas o automatizadas. Tanto el Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo sobre inteligencia artificial (AI Act), como las directrices de la Comisión Europea (2019), destacan la necesidad de que las decisiones con impacto significativo estén sujetas a un control humano significativo (*meaningful human oversight*), especialmente en ámbitos sensibles como la justicia, la seguridad y la administración pública. Este concepto no solo introduce límites técnicos y jurídicos al uso de sistemas de IA, sino que sitúa en el centro del debate el papel del factor humano en los procesos de digitalización.

Más allá de la normativa, múltiples investigaciones han evaluado cómo interactúan los operadores humanos con sistemas automatizados y cuáles son las implicaciones de dicha interacción. Ello ha generado un debate crucial sobre la función que debe desempeñar el factor humano en los procesos judiciales mediados por tecnologías digitales. El operador no puede reducirse a una figura legitimadora de la decisión algorítmica, sino que debe ejercer un papel activo en la interpretación, el control y, en su caso, la corrección de los resultados. Esta exigencia se vincula directamente con la idea de que la supervisión no es solo técnica, sino también ética y política, en tanto busca preservar la legitimidad institucional y la confianza de la sociedad en la administración de justicia.

La relación entre el factor humano y las nuevas tecnologías se manifiesta, además, en la influencia que estas ejercen sobre el comportamiento profesional, modulando percepciones, actitudes y conductas (Morales, Moya, Gaviria & Cuadrado, 2007). Ejemplo de ello es el estudio de Portela, Castillo, Tolan, Karimi-Haghighi y Andrés Pueyo (2025), que analizó la integración experimental de una versión algorítmica de RisCanvi en el sistema judicial catalán. Sus resultados mostraron que el soporte algorítmico aumentaba la precisión en la toma de decisiones, especialmente en perfiles especializados, aunque los participantes rechazaban el uso totalmente automatizado. De forma similar, Kleinberg et al. (2018) demostraron, a partir del análisis de más de 750.000 decisiones judiciales de libertad bajo fianza en Nueva York, que los algoritmos podían reducir la criminalidad sin aumentar la

encarcelación, siempre que se utilizaran como apoyo al juicio humano. Estas investigaciones subrayan que el factor humano es indispensable para orientar las herramientas hacia fines compatibles con valores democráticos y necesidades sociales.

Los estudios comparativos entre predicciones humanas y algorítmicas también han generado un intenso debate, Dressel y Farid (2018) señalaron que personas sin experiencia podían alcanzar niveles de precisión similares a los del software COMPAS. Sin embargo, trabajos posteriores como los de Bansak (2019) y Lin, Jung, Goel y Skeem (2020) matizaron estos hallazgos, mostrando que en escenarios más cercanos a la práctica real los algoritmos superaban sistemáticamente al juicio humano. Estos resultados alertan sobre dos riesgos contrapuestos: la confianza excesiva en las herramientas digitales o, en el extremo contrario, su uso acrítico sin comprensión de su lógica subyacente. En la misma dirección, Zerilli, Knott y Maclaurin (2021) sostienen que la supervisión solo puede considerarse auténtica si los profesionales comprenden y son capaces de cuestionar los sistemas, evitando convertirse en legitimadores pasivos de decisiones automatizadas. Desde esta perspectiva, la supervisión humana aparece como condición indispensable no solo por razones técnicas, sino también por su vínculo con la percepción ciudadana de legitimidad. Dado que el objetivo último de la justicia es garantizar un servicio público eficaz y justo, la participación de la ciudadanía en el debate sobre la implantación de tecnologías resulta esencial para asegurar su aceptación social y su adecuación a valores compartidos (Ferguson, 2017; Deeks, 2019; Wachter, Mittelstadt & Floridi, 2017; Završnik, 2021).

La cuestión que surge de manera inevitable es hasta qué punto la transformación digital de la justicia penal representa un riesgo real de sustitución del factor humano o, por el contrario, abre la posibilidad de modelos de complementariedad que refuercen su papel. En este contexto, el despliegue tecnológico y la incorporación de herramientas basadas en inteligencia artificial y algoritmos predictivos han colocado en el centro del debate la compleja interacción entre personas y máquinas en la toma de decisiones judiciales. Esta colaboración híbrida plantea desafíos éticos y jurídicos de gran envergadura, vinculados a la transparencia, la rendición de

cuentas y la preservación de la autonomía profesional. La opacidad de ciertos algoritmos y el riesgo de reproducir o amplificar sesgos presentes en los datos o en sus diseñadores han llevado a plantear si el factor humano se encamina hacia la sustitución o hacia una complementariedad con las herramientas tecnológicas.

El debate se acentúa en este marco, Zilberman (2016) advierte que la automatización puede redefinir o incluso eliminar puestos de trabajo en el corto plazo, mientras que Binns (2018) y Pasquale (2020) cuestionan el riesgo de una "automatización ciega" que limite la autonomía judicial. En paralelo, la regulación europea (European Commission, 2019; Beck & Burri, 2024) insiste en la necesidad de mantener la supervisión y el control humano en todo sistema de alto riesgo, aunque reconoce que existen contextos en los que dicha supervisión puede ser insuficiente o limitada. En consecuencia, el eje del debate no reside en la sustitución del factor humano, sino en la forma en que se articula su interacción con las herramientas tecnológicas, apostando por modelos de colaboración que potencien tanto las capacidades humanas como las digitales. Si bien la supervisión no garantiza por sí sola la eliminación de riesgos como la falta de explicabilidad o los sesgos de automatización (Goddard, Roudsari & Wyatt, 2012; Skitika, Mosier & Burdick, 2012), sí constituye un eje esencial en la construcción de un modelo híbrido que preserve la independencia judicial y la legitimidad institucional. En otros contextos, investigaciones como la de Lai y Tan (2019) han evidenciado que el grado de asistencia algorítmica condiciona tanto la precisión de las decisiones como la confianza depositada por los usuarios reforzando la idea de que la interacción entre el criterio humano y la capacidad analítica de la máquina obliga a replantear principios fundamentales como la transparencia, la rendición de cuentas y la responsabilidad.

La literatura especializada coincide en señalar que, en el escenario actual, no existe una sustitución plena del factor humano en el sistema de justicia penal. Más que anticipar el desplazamiento de jueces, fiscales u otros operadores jurídicos, los estudios empíricos y normativos comentados apuntan hacia un modelo de complementariedad, en el que las tecnologías digitales y algorítmicas se conciben como herramientas de apoyo destinadas a reforzar, estandarizar y agilizar la toma

de decisiones, pero sin reemplazar la capacidad crítica, la independencia y la responsabilidad inherentes a la función judicial. El debate se centra, por tanto, no en la posibilidad de una justicia automatizada, sino en las condiciones que permiten su integración legítima, asegurando que la autonomía profesional, la rendición de cuentas y la legitimidad institucional permanezcan garantizadas.

En este marco, resulta clave analizar el papel de los operadores jurídicos, quienes determinan en la práctica cómo se articula la interacción entre el juicio humano y el soporte algorítmico. Si bien parte de la literatura se ha ocupado de examinar la capacidad persuasiva de los algoritmos y su incidencia en la toma de decisiones (Green & Chen, 2019; Stevenson & Doleac, 2020), el estudio empírico de las actitudes profesionales ante estas tecnologías continúa siendo limitado. Esta carencia adquiere especial relevancia porque son precisamente las percepciones de los profesionales las que condicionan tanto el grado de aceptación como la modalidad concreta de incorporación de la inteligencia artificial a la práctica judicial. La evidencia disponible apunta hacia un modelo híbrido de colaboración en el que la tecnología refuerza la labor humana en lugar de desplazarla. No obstante, la existencia de marcos regulatorios o de resultados empíricos positivos no resuelve automáticamente los dilemas éticos y prácticos que suscita esta interacción. El verdadero desafío consiste en identificar los factores que permiten que la relación humano-máquina se traduzca en aceptación social y confianza institucional, sin comprometer los derechos fundamentales ni la legitimidad democrática.

El debate sobre la incorporación de la inteligencia artificial en la justicia no puede reducirse a la dicotomía entre sustitución y complementariedad, sino que requiere analizar los modelos de interacción que explican cómo se configuran las relaciones entre personas y sistemas algorítmicos en la toma de decisiones, así como las actitudes de la ciudadanía respecto a los diferentes modelos de interacción. Estos marcos permiten determinar cuestiones clave como el grado de autonomía tecnológica, el nivel de supervisión necesario y las implicaciones que cada configuración conlleva para la transparencia, la rendición de cuentas y la legitimidad democrática. En este sentido, la cuestión central radica en comprender cómo se articula en la práctica la interacción humano-algoritmo y bajo qué condiciones

puede consolidarse como un mecanismo funcional y, al mismo tiempo, legítimo y socialmente aceptado.

1.1. Modelos de interacción entre humanos y sistemas digitales en la justicia penal.

La transformación digital del sistema de justicia penal ha dado lugar a nuevas formas de interacción entre los operadores jurídicos y las herramientas tecnológicas. Esta relación no se limita al uso instrumental de dispositivos, sino que se configura a través de modelos cada vez más complejos de colaboración, supervisión y delegación de funciones a sistemas digitales, en particular aquellos basados en inteligencia artificial (Nieva Fenoll, 2018; Miró Llinares, 2018). Analizar estos modelos de interacción permite comprender no solo el grado de autonomía tecnológica involucrado, sino también los desafíos éticos, jurídicos y prácticos que plantea su implementación en decisiones que afectan derechos fundamentales (European Parliamentary Research Service [EPRS], 2020).

Para hacer efectivo este control, se han desarrollado diferentes modelos de supervisión (Harbers, Peeters & Neerincx, 2017):

- En el modelo *human in the loop*, el sistema solo ejecuta sus funciones tras la validación o intervención de un operador humano.
- En el *human on the loop*, la Inteligencia artificial puede operar de forma autónoma, aunque se mantiene la posibilidad de que un supervisor intervenga si es necesario, modificando o deteniendo su funcionamiento.
- En el *human out of the loop*, la supervisión se limita a una evaluación posterior del desempeño del sistema, sin intervención directa en su operación.

1.1.1. Human in the loop: supervisión activa de la IA en la justicia.

El avance de las herramientas de inteligencia artificial ha motivado a los investigadores a desarrollar modelos que redefinen las interacciones entre humanos y algoritmos de aprendizaje automático. Este enfoque, conocido como aprendizaje automático con intervención humana (*Human-in-the-Loop*, HITL),

busca optimizar la colaboración entre ambas partes para mejorar la toma de decisiones y la eficacia de los sistemas inteligentes. En los últimos años, el análisis de los sistemas human-in-the-loop (HITL) ha cobrado relevancia en el ámbito de la teoría de control y de la interacción hombre-máquina. Estos sistemas se caracterizan por la inclusión del operador humano como un elemento esencial en la toma de decisiones y en la dinámica de control, lo que plantea retos que no se presentan en los sistemas puramente automáticos.

En este sentido, Mabrok, Mohamed, Abdel-Aty y Alzahrani (2020) realizan una revisión exhaustiva de las principales líneas de investigación en torno a la modelización del comportamiento humano dentro de los sistemas HITL, destacando tanto las limitaciones físicas y cognitivas del operador como las posibilidades de integración de modelos humanos en metodologías formales de control. El estudio identifica cuatro direcciones prioritarias de investigación: la comprensión de los atributos y limitaciones del operador humano, la elaboración de una taxonomía de las aplicaciones HITL, la construcción de modelos realistas del comportamiento humano y la integración de estos modelos en la síntesis de control. Entre los hallazgos más relevantes se resalta que el comportamiento humano difiere radicalmente del de los controladores físicos debido a factores como los retrasos en la respuesta, la memoria limitada, la anticipación, la capacidad de aprendizaje y adaptación, así como la naturaleza no lineal y, en muchos casos, caótica de las conductas humanas. En el plano aplicado, los autores subrayan la importancia de diseñar sistemas personalizados que puedan adaptarse a las habilidades del operador, y de desarrollar marcos de asignación óptima de funciones entre humanos y máquinas. Asimismo, señalan que los modelos existentes son todavía demasiado generales o excesivamente específicos, lo cual limita su aplicabilidad a escenarios diversos. De ahí la propuesta de avanzar hacia una biblioteca de modelos mentales genéricos y configurables que permita abordar la complejidad de la interacción hombre-máquina en contextos variados.

Sin embargo, algunos autores como Natarajan, Mathur, Sidheekh, Stammer y Kersting (2025) defienden que gran parte de los sistemas denominados HITL realmente se ajustarían más a la definición de AI-in-the-Loop (AI2L), en el cual la

figura central de control sigue siendo el ser humano y la IA desempeña un papel de apoyo. La evolución de la analítica visual ha estado marcada por la necesidad de integrar las capacidades computacionales de los algoritmos con los procesos cognitivos y exploratorios de los analistas. Tradicionalmente, este campo se ha guiado por la filosofía del *"human in the loop"*, donde la intervención humana aparece como un elemento externo que valida o corrige las operaciones de los sistemas automatizados. Sin embargo, Endert et al. (2014) plantean un cambio de paradigma hacia la perspectiva *"human is the loop"*, que sitúa al analista como parte central del proceso analítico. Este enfoque reconoce que las interacciones humanas, tanto explícitas como implícitas, constituyen el núcleo del descubrimiento de conocimiento en contextos caracterizados por la incertidumbre y la complejidad de los datos.

La distinción entre ambos modelos no es meramente terminológica, sino que implica diferencias sustanciales en tres dimensiones: el locus de control, las fuentes de sesgo y los criterios de evaluación. En el marco HITL, la IA asume la dirección del proceso decisional y solicita intervención puntual de la persona, de modo que los sesgos más relevantes se vinculan con los datos históricos y con las aportaciones humanas. Por el contrario, en AI2L el ser humano conserva la autoridad sobre las decisiones, mientras que la IA actúa como asistente cognitivo, lo que desplaza la preocupación hacia sesgos algorítmicos y de interpretación. Asimismo, mientras los sistemas HITL se valoran prioritariamente con métricas algorítmicas como precisión, exhaustividad o recall, los sistemas AI2L exigen evaluaciones centradas en la experiencia y el impacto en los usuarios, incluyendo la confianza, la explicabilidad y la utilidad social (Natarajan et al., 2025). Debido a ello, es importante definir bien qué tipo de modelo se está utilizando ya que tratar erróneamente un sistema AI2L como si fuese HITL puede conducir a errores de diseño, evaluaciones inadecuadas y riesgos significativos en su implementación. De hecho, al considerar ejemplos prácticos en ámbitos como la medicina, la automoción o las finanzas, se aprecia que las tareas rutinarias y bien definidas son más aptas para enfoques HITL, mientras que aquellas más complejas y contextuales requieren necesariamente de una perspectiva AI2L. En este sentido, la transición conceptual hacia AI2L facilita concebir la IA como colaboradora que amplifica la agencia

humana, en lugar de como mera tecnología de automatización, favoreciendo sistemas más resilientes y alineados con fines sociales y éticos (Natarajan et al., 2025).

1.1.2. Human on the loop: supervisión pasiva y control de decisiones automatizadas.

En el modelo denominado *Human-on-the-Loop* (HOL), la persona no participa de manera activa en cada decisión, como ocurre en el enfoque *Human-in-the-Loop* (HITL), sino que se limita a una función de supervisión pasiva, con capacidad para validar o eventualmente corregir las recomendaciones del sistema (Cummings, 2014). En otras palabras, el operador humano permanece "en el circuito" de control, pero en una posición de segundo plano, donde su rol principal es garantizar que las salidas algorítmicas se ajusten a parámetros legales, éticos o técnicos previamente establecidos.

Este tipo de modelo se encuentra en expansión debido a la creciente sofisticación de los sistemas algorítmicos de predicción y clasificación. Rahwan et al. (2019) sugieren que los entornos *on the loop* responden a la necesidad de equilibrar la autonomía técnica de la máquina con un mecanismo mínimo de control humano, que opera más como salvaguarda que como instancia decisoria. A diferencia de HITL, donde los sesgos pueden introducirse tanto desde los datos como desde las intervenciones humanas, en HOL el riesgo principal radica en que los supervisores otorguen un nivel de confianza excesivo al algoritmo, fenómeno conocido como automation bias (Goddard, Roudsari & Wyatt, 2012).

El ámbito judicial ofrece algunos de los ejemplos más significativos de implementación del modelo HOL. Un ejemplo de este tipo de modelos es el Harm Assessment Risk Tool (HART), desarrollado en el Reino Unido, permite predecir la probabilidad de reincidencia de los acusados y, a partir de dicha estimación, apoyar decisiones relacionadas con la libertad condicional o la adopción de medidas preventivas (Oswald, Grace, Urwin, & Barnes, 2020). En este marco, el operador humano interviene únicamente al final del proceso, validando la recomendación algorítmica. El rol pasivo de la persona encarna el principio central del modelo

human on the loop: no se busca una interacción continua con el algoritmo, sino una verificación. No obstante, las experiencias en sistemas como HART o COMPAS en Estados Unidos han generado debates intensos sobre el alcance de dicha supervisión. La investigación de Angwin, Larson, Mattu y Kirchner (2016) en ProPublica evidenció que COMPAS tendía a clasificar de manera desproporcionada a personas afroamericanas como de alto riesgo, mientras que sobreestimaba la probabilidad de no reincidencia en personas blancas. Este hallazgo pone de relieve que, incluso en presencia de un humano en el circuito, la capacidad de corregir sesgos es limitada si la supervisión se da en una etapa final y con escaso acceso a la lógica interna del algoritmo.

Diversos autores advierten que la eficacia de HOL se encuentra condicionada por las limitaciones cognitivas de los supervisores. Cummings (2014) sostiene que cuando la intervención humana se reduce a la validación puntual de una decisión ya elaborada por la máquina, se genera un efecto de sobrecarga cognitiva: el operador debe evaluar decisiones complejas con información parcial y bajo presión de tiempo, lo que disminuye su capacidad crítica. Burrell (2016) complementa este argumento al señalar que la opacidad de los algoritmos de aprendizaje automático, como los modelos de caja negra, dificulta aún más que los humanos comprendan y evalúen adecuadamente los resultados. En este sentido, Floridi y Cowls (2019) plantean que los entornos HOL corren el riesgo de producir una ilusión de supervisión: aunque formalmente exista una figura humana en la cadena de decisión, en la práctica el poder efectivo recae en la máquina. Esto produce un fenómeno de responsabilidad difusa, en el cual resulta complejo atribuir culpas en caso de errores: ¿es responsable el programador, la institución que adopta el sistema o el supervisor humano que validó mecánicamente la recomendación?

Los marcos normativos internacionales han reconocido este problema, la UNESCO (2021), en su Recomendación sobre la ética de la inteligencia artificial, subraya que la supervisión humana debe ser "significativa", es decir, con información suficiente y capacidad real de intervención. De manera similar, el Reglamento de Inteligencia Artificial de la Unión Europea (European Commission, 2021) establece que en aplicaciones de alto riesgo, como la justicia penal, no basta con que exista un humano

formalmente presente en el proceso: es necesario que ese humano disponga de explicaciones claras, datos suficientes y autoridad para contradecir la salida algorítmica. Sin embargo, la implementación práctica de estas recomendaciones enfrenta obstáculos técnicos y organizativos. Mittelstadt (2019) critica que los principios éticos de la IA son, en muchos casos, meramente declarativos y no se traducen en mecanismos efectivos de gobernanza. En escenarios HOL, esto significa que la validación humana puede quedar reducida a un "checklist" burocrático, sin impacto real en la justicia sustantiva de las decisiones.

1.1.3. Human out of the loop: automatización total y sus riesgos.

El modelo denominado *Human-out-of-the-Loop* (HOTL) describe aquellos sistemas en los que la inteligencia artificial actúa de manera completamente autónoma en la toma de decisiones, prescindiendo por completo de la intervención o supervisión humana en cualquiera de sus fases. A diferencia de los modelos *Human-in-the-Loop* (HITL), caracterizados por una supervisión activa, o los modelos *Human-on-the-Loop* (HOL), basados en una supervisión pasiva, el HOTL traslada el control decisorio al algoritmo y excluye a los operadores humanos del proceso. Este enfoque, propio de la automatización plena, ha suscitado una intensa controversia académica y política, especialmente en el ámbito de la justicia penal, en el que las decisiones tienen efectos directos sobre derechos fundamentales como la libertad personal, la igualdad ante la ley y el derecho a un juicio justo (Završnik, 2020).

La principal característica de este modelo radica en la eliminación del juicio humano como instancia de corrección o deliberación. En este sentido, Yeung (2019) señala que los sistemas HOTL generan lo que denomina *accountability gaps*, es decir, vacíos de responsabilidad en los que resulta difícil identificar quién debe responder por las decisiones tomadas. Cuando un algoritmo emite una resolución judicial sin mediación humana, ¿es responsable el programador que diseñó el sistema, la institución que lo implementa o el propio tribunal que lo utiliza? Este desplazamiento de la agencia humana hacia una máquina plantea nuevos dilemas estructurales en torno a la rendición de cuentas y a la legitimidad democrática de las decisiones judiciales.

La incompatibilidad del HOTL con los marcos normativos internacionales ha sido señalada por diversos organismos. El Consejo de Europa (2021), en su Recomendación sobre el uso de la inteligencia artificial en los sistemas judiciales, subraya que el control humano significativo (*meaningful human control*) constituye una condición indispensable para preservar la independencia judicial y la legitimidad de los procesos. En una línea similar, la Unión Europea, en su propuesta de Reglamento de Inteligencia Artificial, establece la prohibición expresa de sistemas que eliminen la intervención humana en contextos de alto riesgo, entre ellos la justicia penal (European Commission, 2021). Estos documentos reflejan un consenso normativo en torno a la necesidad de mantener un papel central del ser humano en la toma de decisiones que afectan a los derechos fundamentales.

Este tipo de automatización total podría tener aplicaciones en tareas de carácter rutinario, como la clasificación documental, la gestión administrativa de expedientes o la verificación de requisitos formales (Goodman & Flaxman, 2017). En estas áreas, la delegación completa a sistemas algorítmicos podría liberar recursos humanos y reducir tiempos procesales. No obstante, incluso en tales escenarios, se advierte la necesidad de mantener mecanismos de auditoría y supervisión externa que garanticen el cumplimiento de estándares legales y éticos.

El modelo *Human-out-of-the-Loop* vendría a representar la forma más radical de delegación algorítmica y plantea riesgos de gran calado para los derechos fundamentales, la rendición de cuentas y la legitimidad de la justicia penal. Aunque pueda resultar útil en tareas administrativas o de bajo impacto, su aplicación en decisiones sustantivas contradice tanto los principios normativos internacionales como los valores democráticos que sustentan el Estado de derecho. Frente a los modelos colaborativos que buscan un equilibrio entre eficiencia tecnológica y control humano, el HOTL se perfila como un paradigma normativamente inviable y socialmente indeseable.

1.2. Más allá de la interacción instrumental: el factor humano en la arquitectura del sistema de justicia penal 4.0.

Más allá de los modelos de interacción entre humanos y sistemas algorítmicos, el denominado factor humano se proyecta en múltiples dimensiones que resultan decisivas para comprender la transformación digital de la justicia penal. Lejos de constituir un elemento accesorio, la intervención humana se revela como condición indispensable para garantizar que la digitalización de la justicia se mantenga alineada con los principios del Estado de derecho (UNESCO, 2021; European Commission, 2024, Art. 14). Este factor no se circunscribe a una sola dimensión, sino que atraviesa distintas fases en el uso de nuevas herramientas tecnológicas: la supervisión de los resultados que generan, el diseño y desarrollo de las herramientas, el uso cotidiano por parte de operadores del sistema de justicia penal, la regulación normativa que establece sus límites, y la experiencia de la ciudadanía como receptora de sus efectos. La identificación de estos planos resulta esencial para comprender no solo la configuración actual de la justicia penal digitalizada, sino también los desafíos que plantea su legitimación democrática.

Uno de los ámbitos en los que con mayor insistencia se ha subrayado la necesidad de intervención humana es la supervisión de los resultados algorítmicos. Organismos internacionales como la UNESCO (2021) y la Comisión Europea (2024) han establecido que, tratándose de sistemas de alto riesgo como los judiciales y penitenciarios, debe garantizarse siempre un control humano efectivo. Sin esa capacidad de supervisión crítica, el algoritmo corre el riesgo de convertirse en una "caja negra" que desplaza el juicio humano y consolida sesgos estructurales (Citron & Pasquale, 2014). Ejemplos como el caso estadounidense de COMPAS, cuestionado por sus sesgos raciales (Angwin et al., 2016), muestran cómo la ausencia de una intervención crítica por parte de jueces y operadores puede derivar en decisiones de enorme trascendencia tomadas sobre la base de predicciones opacas y discriminatorias.

El factor humano, sin embargo, no se limita a la fase final de validación de resultados, sino que se encuentra también en el diseño y desarrollo de las herramientas. Los algoritmos no son productos neutros: integran supuestos, criterios y valores que

dependen de quienes los construyen (Mittelstadt et al., 2016). Así, la participación de equipos multidisciplinares compuestos por juristas, criminólogos, ingenieros y expertos en ética resulta imprescindible para que las soluciones tecnológicas se ajusten no solo a parámetros de eficiencia, sino también a principios de proporcionalidad, transparencia y derechos fundamentales. La experiencia del sistema RisCanvi en Cataluña constituye un ejemplo paradigmático de este enfoque. Al haber sido desarrollado con la participación de criminólogos y profesionales penitenciarios, la herramienta no solo se adaptó mejor al contexto institucional, sino que también gozó de mayor legitimidad entre sus usuarios (Andrés-Pueyo & Echeburúa, 2010). Esta dimensión del factor humano demuestra que los algoritmos no se "descubren", sino que se diseñan, y que en ese proceso se definen los contornos de su impacto social.

Asimismo, el factor humano aparece en el uso cotidiano de las herramientas por parte de los operadores del sistema de justicia penal. No se trata únicamente de revisar los resultados, sino también de introducir datos, aportar información contextual y traducir los outputs algorítmicos a la práctica institucional. En el sistema VioGén, por ejemplo, los agentes policiales son responsables de completar formularios y valorar circunstancias que el algoritmo no puede procesar de manera autónoma, como elementos situacionales o percepciones de riesgo inmediatas. Estos datos alimentan el cálculo automatizado y, a su vez, el propio agente tiene la posibilidad de modular la valoración final en función de su criterio profesional (López-Ossorio et al., 2019). Este plano pone de relieve la importancia de la experiencia y sensibilidad de los operadores humanos, que se convierten en mediadores entre la frialdad estadística del algoritmo y la complejidad social de los casos concretos.

El diseño de marcos regulatorios constituye otra manifestación fundamental del factor humano. Las tecnologías algorítmicas no se insertan en un vacío normativo, sino en un entramado de regulaciones jurídicas que delimitan su alcance y condiciones de uso. La aprobación del Reglamento de Inteligencia Artificial por parte de la Unión Europea (European Commission, 2024) o la Recomendación del Consejo de Europa sobre inteligencia artificial y justicia penal (Council of Europe,

2024) son muestras de cómo la regulación busca establecer obligaciones de transparencia, rendición de cuentas y supervisión humana. Estos instrumentos reflejan que el control humano no solo se ejerce en la interacción con la herramienta, sino también en la definición de los marcos normativos que orientan su aplicación. Como advierte Yeung (2018), la regulación algorítmica debe ir más allá de aspectos técnicos e incorporar consideraciones sustantivas sobre equidad, proporcionalidad y justicia.

Finalmente, el factor humano se encuentra en la ciudadanía, en tanto destinataria última de las decisiones judiciales y penitenciarias apoyadas en herramientas algorítmicas. Las personas justiciables, las víctimas y la sociedad en su conjunto no son meros receptores pasivos, sino actores cuyas percepciones de legitimidad y confianza condicionan la eficacia social de estas tecnologías. Teorías como la de la justicia procedimental sostienen que la aceptación de las decisiones jurídicas depende, en gran medida, de la percepción de imparcialidad, respeto y transparencia durante el proceso, más allá del resultado concreto (Tyler, 2006). Así, incluso el algoritmo más preciso puede carecer de legitimidad si quienes se ven afectados por él lo perciben como opaco, deshumanizado o arbitrario.

Resulta evidente que el factor humano se despliega en todas las fases del ciclo de vida de las herramientas algorítmicas aplicadas al sistema de justicia penal: en la supervisión, en el diseño, en el uso práctico, en la regulación y en la recepción ciudadana. No obstante, la presente tesis doctoral centra su atención en dos de estos planos: los operadores del sistema de justicia penal y la ciudadanía. Esta delimitación responde a la convicción de que la legitimidad democrática de la digitalización judicial no depende únicamente de los marcos normativos o de la sofisticación técnica de los algoritmos, sino también de la forma en que quienes aplican y quienes reciben sus efectos perciben, aceptan o cuestionan su funcionamiento. Conviene señalar que, pese a su relevancia, esta doble dimensión ha sido la que menos atención ha recibido hasta el momento en la literatura especializada, lo que genera un vacío de conocimiento que resulta necesario abordar. Precisamente por ello, la reciente investigación doctoral sitúa en el centro de su análisis las actitudes de estos actores hacia el uso de herramientas

algorítmicas, partiendo de la premisa de que solo si se consideran sus percepciones y experiencias será posible introducir dichas innovaciones de manera adecuada y socialmente legítima. Analizar estas dos dimensiones permite comprender en qué medida la digitalización transforma no solo los procedimientos técnicos, sino también las percepciones sociales y las prácticas institucionales que constituyen la base de la confianza en la justicia.

2. Participación ciudadana en el uso de herramientas algorítmicas: actitudes, aceptación social y legitimidad democrática.

La revisión del papel del factor humano en la digitalización de la justicia penal ha puesto de relieve que la ciudadanía no puede entenderse únicamente como receptora pasiva de decisiones mediadas por algoritmos, sino también como un actor activo cuya percepción, confianza y capacidad de intervención resultan determinantes para la legitimidad del sistema. En este sentido, la cuestión trasciende el análisis de la supervisión técnica o de la intervención profesional, situando en el centro del debate la necesidad de abrir espacios de implicación directa de la ciudadanía en la definición de los fines, procedimientos y límites de estas herramientas. El debate sobre la aceptación social de herramientas algorítmicas en la toma de decisiones públicas ha venido acompañado de un interés creciente por la participación ciudadana como un componente esencial en la evaluación ética y social de tales tecnologías. La incorporación de innovaciones como la inteligencia artificial, los algoritmos predictivos o los sistemas automatizados de gestión de casos no solo transforma las prácticas de los operadores de justicia, sino que también interpela a la ciudadanía como sujeto último de las decisiones judiciales. Tal como sostiene Innerarity (2024), "la gobernanza algorítmica únicamente puede ser democrática cuando sus objetivos y procedimientos han sido expresamente autorizados por el pueblo en un acto de naturaleza política" (p. 21). Este planteamiento subraya que uno de los núcleos críticos de la participación ciudadana es precisamente la legitimidad democrática de estas herramientas algorítmicas.

Este razonamiento parte del reconocimiento de que las decisiones algorítmicas no son meramente técnicas, sino que conllevan consecuencias jurídicas, sociales y

políticas que afectan directamente los derechos ciudadanos. En este contexto, la participación emerge como un mecanismo de legitimación democrática en el diseño, la implementación y la vigilancia de algoritmos, especialmente en ámbitos sensibles como la justicia penal (Busuioc, 2020). Desde una perspectiva normativa, se argumenta que la participación ciudadana no solo mejora la transparencia y la rendición de cuentas, sino que también permite incorporar valores sociales, perspectivas plurales y preocupaciones contextuales ausentes en los enfoques puramente técnicos o centrados en expertos (Van de Poel, 2020). Esto implica que la participación no se limita a ser un procedimiento consultivo, sino que constituye un instrumento indispensable para garantizar que el desarrollo tecnológico se alinee con principios de justicia, equidad y respeto por los derechos fundamentales.

La importancia de estos aspectos se evidencia en estudios empíricos como el de Bruun (2024), quien, a partir de un trabajo etnográfico en la ciudad de Vejle (Dinamarca), analizó el proyecto I-REACT, diseñado para implicar a la ciudadanía en la gestión de inundaciones mediante una aplicación de datos participativos. Los resultados mostraron que la ciudadanía rechazó mecanismos de participación gamificados y percibidos como extractivos, defendiendo en su lugar formas de colaboración basadas en relaciones de confianza, cuidado mutuo y experiencias previas de interacción con las instituciones. Esta investigación demuestra que la legitimidad de los sistemas algorítmicos no puede construirse únicamente sobre la entrega de datos o la motivación individual, sino que depende de la trayectoria histórica de las relaciones entre ciudadanía y Estado, así como de expectativas morales de reconocimiento y reciprocidad.

En esta línea, el estudio de Lahdili, Önder y Nyadera (2024) refuerza que la inteligencia artificial aplicada a la participación ciudadana constituye una espada de doble filo. Por una parte, puede ampliar la transparencia, la rendición de cuentas y la capacidad de respuesta institucional mediante el uso de chatbots, algoritmos de aprendizaje automático y plataformas digitales. Por otra, plantea riesgos vinculados a la manipulación política, la vigilancia masiva, los sesgos algorítmicos y la erosión de derechos fundamentales. Los autores identifican tres escenarios posibles en la relación entre inteligencia artificial y democracia: el optimismo tecnológico, el uso

inclusivo y deliberativo de la tecnología y la amenaza de su instrumentalización para el control y la manipulación. Estas conclusiones muestran que el impacto de la inteligencia artificial en la participación ciudadana depende en última instancia de los marcos normativos, éticos y de gobernanza que acompañen su desarrollo y despliegue. Precisamente, el trabajo de Castellanos Claramunt (2019) evidencia este doble potencial de la inteligencia artificial en el terreno democrático. Desde una perspectiva negativa, el autor advierte sobre fenómenos como la personalización política, las técnicas de persuasión digital y el uso de algoritmos capaces de limitar la pluralidad informativa, generar burbujas de filtros y facilitar la manipulación electoral mediante la explotación de datos personales, con riesgos evidentes de control social y erosión de la libertad ciudadana. Desde una perspectiva positiva, señala que la robotización del trabajo podría liberar tiempo y recursos para la participación política y abrir espacios para recuperar formas de democracia más directa, semejantes a la ateniense. El artículo concluye que el reto consiste en diseñar marcos éticos y legales que garanticen que la inteligencia artificial refuerce la democracia en lugar de socavarla, planteando la necesidad de regulaciones sólidas y de una supervisión ciudadana efectiva como mecanismos de legitimación.

En el ámbito judicial, donde las decisiones impactan sobre la libertad, la dignidad y la reputación de las personas, esta exigencia adquiere una urgencia aún mayor. Sistemas como los de evaluación de riesgo, predicción de reincidencia o recomendación judicial deben someterse al escrutinio público, no solo por sus efectos, sino también por la legitimidad de su uso en un sistema tradicionalmente basado en la deliberación humana y la argumentación jurídica (Eubanks, 2018). Diversos estudios muestran que la participación ciudadana en el diseño e implementación de herramientas algorítmicas sigue siendo limitada y, cuando existe, suele ser simbólica o instrumental en lugar de auténticamente deliberativa o vinculante (Zuboff, 2019; Lepri, Oliver, Letouzé, Pentland & Vinckv, 2018). Esto plantea desafíos complejos para la gobernanza algorítmica. Si se aspira a una justicia digital verdaderamente inclusiva y ética, es necesario repensar los procesos institucionales de participación para asegurar que la ciudadanía tenga voz significativa en decisiones sobre qué tecnologías se adoptan, cómo se evalúan y bajo qué condiciones se retiran o corrigen. En esta misma línea, pero en el contexto

anglosajón, Donoghue (2017) advierte que la digitalización de los tribunales en Inglaterra y Gales ha priorizado criterios de eficiencia y reducción de costes sobre valores de equidad y participación, lo que evidencia que la incorporación de nuevas tecnologías judiciales no garantiza por sí misma un mayor acceso a la justicia ni una mejora en la calidad democrática de los procesos. Si se aspira a una justicia digital verdaderamente inclusiva y ética, es necesario repensar los procesos institucionales de participación para asegurar que la ciudadanía tenga voz significativa en decisiones sobre qué tecnologías se adoptan, cómo se evalúan y bajo qué condiciones se retiran o corrigen.

En respuesta a estas carencias, emergen marcos éticos recientes que incorporan explícitamente principios de participación, estableciendo procesos transparentes y mecanismos de supervisión con intervención pública, especialmente en casos de inteligencia artificial de alto riesgo (European Commission, 2021). Dichos marcos reconocen que la aceptación social no puede imponerse desde arriba, sino construirse mediante procesos de co-creación, deliberación y supervisión compartida. En esta línea, Shen, Cabrera, Perer y Hong (2020) plantean el enfoque public(s)-in-the-loop, que concibe la participación no como un acto individual ni simbólico, sino como un proceso colectivo y plural de deliberación. Este marco destaca tres elementos fundamentales: la necesidad de reconocer a los públicos como entidades políticas plurales con intereses en conflicto, la relevancia de la deliberación como espacio de intercambio y aprendizaje mutuo y la construcción de públicos en torno a los problemas algorítmicos. Aplicado a casos como el algoritmo COMPAS de predicción de reincidencia, este enfoque pone de relieve que la legitimidad de tales sistemas no reside en encontrar una métrica correcta de equidad, sino en exponer y debatir los intereses contrapuestos de los distintos actores sociales afectados, creando espacios deliberativos donde emerjan soluciones más aceptables socialmente.

En este debate, el aporte de Alnemr (2024) resulta clave al conceptualizar el denominado atajo algocrático, entendido como una forma de deferencia ciega hacia los algoritmos que reproduce y amplifica los déficits democráticos ya presentes en otros atajos identificados por Lafont (2019). La autora muestra que la gobernanza

algorítmica desplaza la deliberación ciudadana, reduce las oportunidades de participación cívica y genera desigualdades en influencia, voz y capacidad de exigir razones, configurando decisiones opacas, inapelables y poco justificables. Ejemplos como el caso Robodebt en Australia o la regulación europea del Reglamento General de Protección de Datos evidencian que las respuestas institucionales basadas en otros atajos epistocráticos o lottocráticos resultan insuficientes, al limitarse a ofrecer explicaciones técnicas o espacios de participación restringida. Frente a ello, Alnemr (2024) propone como alternativa la deliberación aspiracional, un enfoque orientado a garantizar procesos continuos de justificación pública e inclusión ciudadana en la definición de qué algoritmos se adoptan, cómo se diseñan y bajo qué condiciones son legítimos.

Desde una perspectiva teórica más reciente, Grimmelikhuijsen & Meijer (2022) identifica que la toma de decisiones algorítmicas pone en cuestión tres dimensiones de legitimidad: la de entrada, la del proceso y la del resultado. Para abordar estos desafíos, propone marcos institucionales que distinguen entre legitimidad de entrada, de proceso y de resultado en contextos de políticas públicas basadas en algoritmos. En esta misma línea, Ter Minassian (2025) desarrolla un análisis comparativo de modelos de gobernanza que combinan la experticia técnica con la participación ciudadana, destacando los casos de Francia y Brasil como ejemplos de praxis inclusiva y democrática. Trasladado al ámbito de las decisiones algorítmicas, Starke & Luenich (2020) examinan cómo los ciudadanos perciben la legitimidad de distintos modelos, ya sean humanos, algorítmicos o híbridos, en la toma de decisiones dentro de la Unión Europea. Sus hallazgos muestran que los modelos híbridos no son considerados menos legítimos en términos de proceso y resultado, aunque el modelo exclusivamente algorítmico continúa siendo evaluado como menos legítimo desde la perspectiva de la legitimidad de entrada.

A partir de todo lo anterior, debemos entender que el análisis del factor humano en la sistema de justicia penal 4.0 no puede limitarse a la dimensión instrumental de la interacción entre operadores jurídicos y herramientas algorítmicas, sino que debe incorporar también la perspectiva de la ciudadanía como usuaria final y principal destinataria de las decisiones derivadas de tales sistemas. Ello implica reconocer

que las actitudes sociales hacia el uso de estas tecnologías desempeñan un papel decisivo en la construcción de su aceptación social, dado que las decisiones algorítmicas afectan de manera directa a los derechos, expectativas y trayectorias vitales de las personas (Eubanks, 2018; Zuboff, 2019). La aceptación social debe ser entendida como un proceso dinámico en el que la participación ciudadana contribuye a garantizar que los algoritmos reflejen valores democráticos y se sometan a escrutinio público (Van de Poel, 2020; Alnemr, 2024). De este modo, el factor humano debe entenderse no solo como operador de la herramienta, sino también como sujeto político cuya voz y agencia influyen activamente en la definición de los marcos de uso, en la legitimación de los sistemas tecnológicos y en la determinación de las condiciones bajo las cuales estas tecnologías pueden ser implementadas de manera ética y legítima.

3. Aceptación social del uso de herramientas digitales en la justicia penal.

Tal como se ha expuesto en los apartados anteriores, la digitalización ha llegado para quedarse y transformar los procesos sociales tal y como los hemos entendido hasta ahora, especialmente en el ámbito de la justicia. No obstante, sin una comprensión profunda de los factores que influyen en la actitud y, por ende, en la aceptación de estas tecnologías, las inversiones significativas en innovación pueden no generar los beneficios esperados, lo que podría traducirse en un rendimiento deficiente o incluso en el fracaso total de los nuevos sistemas (Venkatesh et al., 2003; Davis, 1989). En este sentido, las actitudes hacia las herramientas tecnológicas representan el punto de partida para su aceptación. Es decir, la aceptación de nuevas soluciones digitales no se produce de forma automática con su introducción, sino que es el resultado de un proceso construido a partir de las actitudes previas que los usuarios ,en este caso, los operadores jurídicos y, en un sentido más amplio, la sociedad, desarrollan hacia estas herramientas. Dichas actitudes pueden definirse como representaciones mentales aprendidas que condensan evaluaciones favorables o desfavorables sobre objetos, personas o ideas, y que, a su vez, orientan la planificación y anticipación de comportamientos, en función tanto de experiencias previas como de la influencia del entorno social (Forgas, 2008).

El trabajo realizado desde la psicología social para comprender el comportamiento

humano ha situado el concepto de actitud en el centro del análisis de cómo las experiencias sociales se transforman en predisposiciones estables que guían la conducta. En 1928, Thomas y Znaniecki afirmaban que la psicología social podía definirse, en su esencia, como el análisis de las actitudes. En aquella época, el significado atribuido a dicho concepto abarcaba una perspectiva mucho más amplia que la vigente hoy en día. Para estos autores, una actitud no se entendía como un rasgo interno e individual del sujeto, sino como una manifestación socialmente construida, con un fuerte componente cultural y colectivo.

A lo largo del último siglo, el concepto de actitud ha sido objeto de múltiples transformaciones teóricas. En sus inicios, autores como Thurstone (1928) y Likert (1932) introdujeron metodologías que permitieron medir empíricamente las actitudes mediante escalas cuantitativas, lo que dio lugar a un auge en la investigación del área. En esa etapa, se concebía la actitud como una variable unidimensional, centrada especialmente en su valor evaluativo y afectivo. Esta perspectiva fue consolidada por Allport (1935), quien definió la actitud como "el concepto más distintivo e indispensable de la psicología social estadounidense contemporánea" (p. 198).

Posteriormente, surgieron modelos más integradores que describen la actitud como una estructura compuesta por componentes cognitivos, afectivos y conductuales. Este enfoque tridimensional reconoce que las creencias, emociones y tendencias conductuales pueden estar alineadas o entrar en conflicto, explicando así las inconsistencias observadas entre actitud y comportamiento (Rosenberg, 1965; Crano et al., 2010).

Durante la segunda mitad del siglo XX, se evidenció una creciente preocupación por la relación entre actitud y conducta. La aparición de la teoría de la disonancia cognitiva (Festinger, 1957) marcó un punto de inflexión al demostrar que las personas pueden modificar sus actitudes para reducir el malestar que surge cuando sus acciones no son coherentes con sus creencias. Esta teoría contribuyó a cuestionar la linealidad del modelo actitud y conducta, introduciendo la posibilidad de que las conductas también influyeran en la formación de actitudes.

En décadas recientes, los enfoques contemporáneos han enfatizado la naturaleza dual y dinámica de las actitudes. Se ha reconocido que estas pueden ser tanto explícitas como implícitas, es decir, conscientes o inconscientes, y que pueden activarse de forma automática o deliberada (Fazio & Olson, 2003; Greenwald et al., 1998). Además, se ha ampliado la mirada hacia procesos simbólicos y culturales, rescatando aportes clásicos como el de Weber (1947), quien entendía las actitudes como formas simbólicas que conectan la experiencia social con las estructuras institucionales y los comportamientos colectivos (Forgas, 2008).

Ahora que el concepto de actitud ha sido definido, es necesario comprender cómo se forman dichas actitudes. Desde la psicología social, se ha argumentado que las actitudes no emergen de forma espontánea, sino que son el resultado de procesos complejos de aprendizaje y socialización. Según el enfoque del aprendizaje social, las actitudes se desarrollan a partir de la observación de conductas y consecuencias en el entorno, así como de la interacción con agentes significativos como la familia, la escuela, los medios de comunicación o los pares (Bandura, 1977). De este modo, el contexto social y cultural cumple un rol fundamental en la construcción de actitudes individuales y colectivas hacia determinados objetos, como las tecnologías digitales.

Las teorías clásicas sobre la formación de actitudes han destacado tradicionalmente la importancia de los componentes afectivos y conductuales. Según Briñol, Falces y Becerra (2006), las actitudes se aprenden del mismo modo que otras respuestas conductuales, es decir, a través de procesos de condicionamiento clásico e instrumental. Estas aproximaciones parten de una visión del aprendizaje en la que el entorno y las consecuencias asociadas a una conducta o estímulo desempeñan un papel central.

Desde las teorías del aprendizaje, tanto el condicionamiento clásico como el condicionamiento instrumental han sido fundamentales para explicar la formación de actitudes. El condicionamiento clásico plantea que un estímulo inicialmente neutro puede adquirir carga emocional si se asocia repetidamente con otro que ya provoca una reacción determinada. En esta línea, Kunst-Wilson y Zajonc (1980) demostraron que la simple exposición reiterada a un estímulo puede generar una

evaluación positiva hacia él, incluso sin una asociación previa consciente. No obstante, esta teoría ha recibido críticas por su falta de consideración del contexto en el que se produce la evaluación del objeto, lo cual limita su aplicabilidad en situaciones sociales complejas. Por su parte, el condicionamiento instrumental sostiene que las actitudes pueden fortalecerse o debilitarse en función de las consecuencias que generen. Verplanck (1955) observó que los refuerzos verbales positivos fomentaban la repetición de opiniones, mientras que los negativos tendían a inhibirlas. En la misma línea, Insko (1965) aportó evidencia sobre cómo el refuerzo diferencial podía modificar actitudes en entornos experimentales con estudiantes. Sin embargo, este enfoque también ha sido cuestionado por centrarse exclusivamente en los efectos observables de la conducta, dejando de lado los procesos cognitivos que intervienen en la formación de actitudes.

La comprensión actual del fenómeno actitudinal se articula en torno a dos enfoques principales: por una parte, la perspectiva psicológica, que considera que las actitudes se originan a partir de la experiencia individual y del conocimiento directo del sujeto con respecto al objeto actitudinal; por otra, la perspectiva sociológica, que pone énfasis en la influencia del entorno social y en las relaciones interpersonales como factores clave en la construcción y modificación de las actitudes (Bolívar, 1994). Estas visiones han dado lugar a una diversidad de teorías que explican el fenómeno desde distintos ángulos. El enfoque cognoscitivo sostiene que la familiaridad con un objeto, derivada de la experiencia directa, facilita la elaboración de valoraciones, ya sean positivas o negativas, lo que convierte al conocimiento en un elemento clave en la estructuración de actitudes (Marín, 1976). En esta misma línea explicativa, el enfoque funcionalista, desarrollado por Katz y Scotland (1959), plantea que las actitudes cumplen funciones adaptativas para el individuo, ya sea respondiendo a necesidades inmediatas (función de proximidad), asociando el objeto actitudinal a la consecución de metas (función instrumental del objeto), o reforzando la autoestima mediante la identificación con creencias compartidas (función instrumental del ego).

Desde una perspectiva más conductual, el modelo del refuerzo propone que las actitudes se consolidan o modifican en función de las consecuencias asociadas a las

acciones del sujeto. Así, las experiencias pasadas, y particularmente las emociones generadas por dichas experiencias, cumplen un papel determinante en la formación de actitudes, lo cual se vincula estrechamente con los principios del condicionamiento emocional. Las respuestas afectivas actúan como reforzadores o inhibidores de determinadas disposiciones actitudinales, configurando patrones de comportamiento relativamente estables. Esta explicación se complementa con el enfoque del aprendizaje social o por observación, que introduce la dimensión social del aprendizaje y reconoce que las actitudes también pueden construirse a partir de la observación del comportamiento ajeno. En este sentido, Morris (1996) subraya que pensamientos, emociones y conductas pueden estar modelados por las acciones, reales o imaginadas, de otras personas, lo cual refuerza la idea de que las actitudes no se forman en el aislamiento, sino en el seno de contextos sociales significativos.

Finalmente, los enfoques contemporáneos tienden a integrar estos elementos y conciben las actitudes como estructuras mediadoras entre los estímulos sociales y las respuestas del individuo. En este marco, se reconoce que las actitudes son configuraciones complejas que incluyen componentes cognitivos, afectivos y conductuales, todos ellos derivados de la experiencia personal y de los procesos de interacción social. Esta visión estructural permite comprender las actitudes como entidades dinámicas, en constante construcción y susceptibles a la influencia del contexto. Desde esta perspectiva integradora, Morales (1999) define la actitud como el producto de múltiples interacciones con el objeto actitudinal, así como de los procesos psicológicos, emocionales y conductuales que acompañan dichas experiencias, lo que refuerza su relevancia en la comprensión del comportamiento humano.

En el caso concreto que nos compete en el marco de la presente tesis, el concepto de actitud resulta especialmente relevante en el mundo actual, donde las transformaciones tecnológicas han alterado profundamente las dinámicas sociales e institucionales. La digitalización del sistema de justicia penal representa un nuevo escenario en el que las actitudes de los operadores jurídicos, los ciudadanos y otros actores implicados cobran un papel central. La incorporación de tecnologías como la inteligencia artificial, los sistemas de gestión automatizada o las plataformas de

tramitación digital no solo modifica las prácticas judiciales, sino que también activa procesos de evaluación, aceptación o resistencia frente a dichos cambios. En este contexto, las actitudes se configuran como mediadoras clave entre la innovación tecnológica y su integración efectiva en el ámbito penal.

La forma en que los profesionales del sistema perciben la utilidad, la legitimidad o la imparcialidad de las herramientas digitales incide directamente en su disposición a utilizarlas, a confiar en sus resultados o a cuestionar sus implicaciones éticas y jurídicas. Así, el análisis contemporáneo de las actitudes debe considerar no solo sus componentes cognitivos, afectivos y conductuales, sino también los entornos sociotécnicos en los que se producen. Como advierte Sunstein (2018), en sociedades digitalizadas los algoritmos y plataformas pueden reforzar predisposiciones existentes, amplificando sesgos y generando cámaras de eco incluso en contextos institucionales. Por tanto, comprender las actitudes hacia la digitalización de la justicia penal requiere integrar perspectivas clásicas de la psicología social con enfoques críticos que reconozcan el papel de la tecnología en la producción y transformación de las predisposiciones humanas, comprendiendo cuales son los modelos que se relacionan con las actitudes y la aceptación social en la incorporación de nuevas tecnologías.

3.1. Modelos explicativos de las actitudes y la aceptación social en contextos de innovación tecnológica.

La importancia de comprender la aceptación de la tecnología por parte de los usuarios es un tema recurrente en la investigación de sistemas de información (Casiraghi et al., 2021; Gaede & Rowlands, 2018; González-Bravo & Valdivia-Peralta, 2015; Hossain & de Silva, 2009; Hueso et al., 2022; Torres Albero et al., 2017; Urquidi Martin et al., 2019; Varela, 2004; Yong Varela et al., 2010). La integración exitosa de nuevas tecnologías en diversos aspectos de la vida depende en gran medida de la voluntad de los individuos y las organizaciones para adoptarlas y utilizarlas de manera efectiva. Por lo tanto, los modelos que buscan explicar y predecir la aceptación de la tecnología no son meros ejercicios académicos, sino herramientas esenciales con implicaciones prácticas significativas para el desarrollo, la implementación y la gestión de la tecnología. La capacidad de anticipar

y abordar las preocupaciones y motivaciones de los usuarios es fundamental para maximizar el valor y el impacto de las innovaciones tecnológicas (Marikyan & Papagiannidis, 2024).

En este contexto, la literatura proveniente de la psicología social, la ciencia del comportamiento y los estudios sobre la adopción tecnológica ha desarrollado un marco teórico amplio y consolidado para explicar los factores que influyen en la disposición de los individuos a aceptar y utilizar nuevas tecnologías. Entre los enfoques fundacionales se encuentran la Teoría de la Acción Razonada (TRA) (Fishbein y Ajzen, 1975) y la Teoría del Comportamiento Planeado (TPB) (Ajzen, 1991), que ofrecen un marco desde la psicología social centrado en el análisis de las actitudes, las normas subjetivas y el control conductual percibido como predictores de la intención de comportamiento. Estas teorías han demostrado una alta capacidad explicativa en una variedad de contextos institucionales y han servido como base conceptual para el desarrollo de modelos más específicos aplicados a la adopción de tecnologías.

Sobre esta base se construyó el Modelo de Aceptación de la Tecnología (TAM), desarrollado por Davis (1989), que adapta los supuestos de la TRA al contexto tecnológico e introduce dos variables clave: la utilidad percibida y la facilidad de uso percibida, las cuales influyen directamente sobre la actitud hacia el uso de una tecnología y, por ende, sobre la intención y la conducta efectiva. Posteriormente, Venkatesh, Morris, Davis y Davis (2003) propusieron la Teoría Unificada de Aceptación y Uso de la Tecnología (UTAUT), que sintetiza elementos de ocho modelos previos y plantea cuatro constructos fundamentales: la expectativa de rendimiento, la expectativa de esfuerzo, la influencia social y las condiciones facilitadoras. Adicionalmente, algunos autores han incorporado perspectivas complementarias como el Modelo de Difusión de Innovaciones de Rogers (2003), que enfatiza la importancia de características de la innovación (ventaja relativa, compatibilidad, complejidad, posibilidad de prueba y observabilidad) y del contexto social en el proceso de adopción, así como el modelo Technology Readiness Index (TRI) (Parasuraman, 2000), que explora la predisposición individual hacia las tecnologías en función de factores como el optimismo, la innovación, la incomodidad

y la inseguridad.

Para comprender con mayor profundidad el alcance explicativo de estos modelos y su aplicabilidad al contexto de la digitalización en el sistema de justicia penal, a continuación se procederá a una exposición pormenorizada de cada uno de ellos. Esta revisión detallada permitirá identificar los factores específicos que influyen en la aceptación tecnológica por parte de los distintos actores implicados, así como las posibles limitaciones y oportunidades de su implementación en entornos jurídicos

2.1.1. Teoría de la Acción Razonada (TRA).

La Teoría de la Acción Razonada (Theory of Reasoned Action, TRA) constituye uno de los modelos más influyentes en la psicología social para la comprensión y predicción del comportamiento humano (Fishbein & Ajzen, 1975). Desarrollada por Martin Fishbein e Icek Ajzen en su obra *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*, esta teoría propone que la intención conductual aparece como el predictor inmediato del comportamiento, mediando la influencia de la actitud hacia la conducta y de las normas subjetivas percibidas. Así, sus autores establecieron los principios básicos de la TRA, clarificando la definición del constructo actitud y proponiendo un modelo relacional entre creencias, intenciones y comportamiento. Uno de los principales aportes de esta teoría fue abordar de manera explícita la conocida "brecha actitud-comportamiento", observando que las actitudes positivas hacia una acción no necesariamente conllevan a su ejecución. La solución propuesta consistió en considerar la intención como variable mediadora, modulada a su vez por la presión social percibida y las valoraciones individuales.

Conforme a lo establecido en el apartado anterior, la actitud hacia la conducta se define como la evaluación general, positiva o negativa, que un individuo realiza respecto a la ejecución de una acción concreta (Ajzen, 1991). Esta evaluación se construye a partir de las creencias conductuales (percepciones sobre los posibles resultados del comportamiento) y la valoración atribuida a dichos resultados, conforme al modelo de expectativa-valor. Por ejemplo, en el contexto concreto de la digitalización, si un sujeto considera que el uso de una herramienta tecnológica

aumentará su productividad, y otorga valor positivo a dicha mejora, es probable que desarrolle una actitud favorable hacia el uso de esa herramienta.

Por su parte, las normas subjetivas aluden a la percepción que tiene el individuo sobre las expectativas de personas significativas en su entorno respecto a la ejecución o no de un determinado comportamiento (Fishbein & Ajzen, 1975). Estas normas se componen de las creencias normativas y de la motivación para cumplir con ellas. Así, una persona que cree que sus colegas esperan que adopte una práctica profesional determinada y valora su opinión, tenderá a desarrollar una intención positiva hacia esa conducta.

La TRA establece que la intención conductual es el factor más inmediato en la cadena causal que lleva al comportamiento. Cuanto más favorable sea la actitud hacia la acción y mayor sea la presión social percibida para ejecutarla, más fuerte será la intención del sujeto de llevarla a cabo (Ajzen, 1991). Sin embargo, el peso relativo de estos dos determinantes puede variar según el contexto, el tipo de comportamiento y las características individuales. Por ejemplo, en culturas colectivistas, la opinión del grupo suele influir más que lo que piensa cada persona por separado.

A pesar de su solidez teórica, la TRA ha recibido diversas críticas. Una de las principales objeciones es su supuesto de racionalidad, al asumir que las personas actúan de forma deliberada y planificada, lo cual ignora la influencia de emociones, hábitos o procesos automáticos (Sheeran, 2002). Además, la teoría presupone un alto grado de control voluntario sobre el comportamiento, lo que no siempre es realista. A partir de esta serie de limitaciones, surgen otros modelos más complejos, como la Teoría del Comportamiento Planificado (Ajzen, 1991)

2.1.2. Teoría del Comportamiento Planeado (TPB).

Como ya sabemos, la Teoría del Comportamiento Planeado desarrollada por Ajzen (1985) surge como una extensión de la Teoría de la Acción Razonada, la TPB introduce el concepto de control conductual percibido como un factor clave en la predicción del comportamiento, lo cual permite mejorar el poder explicativo de su predecesora en contextos donde la conducta no está completamente bajo control

volitivo (Ajzen, 1991; Ajzen, 2002). Dicha intención está determinada por los dos componentes anteriores, la actitud hacia la conducta y la norma subjetiva, pero se incluye el control conductual percibido (Ajzen, 1991).

Según la TRA, la actitud hace referencia a la evaluación favorable o desfavorable que un individuo tiene respecto a ejecutar una conducta concreta, y la norma subjetiva alude a la percepción de presión social para realizarla o evitarla. Como elemento novedoso, el control conductual percibido se refiere a la percepción individual sobre la facilidad o dificultad para llevarla a cabo (Fishbein & Ajzen, 2010).

De manera más detallada, el control conductual percibido alude a la percepción del individuo sobre su capacidad para realizar una conducta, considerando factores externos que pueden facilitarla o dificultarla (Ajzen, 2002). Está estrechamente relacionado con el concepto de autoeficacia (Bandura, 1997), y representa una de las principales innovaciones respecto a la TRA. Además de influir en la intención conductual, el control percibido puede tener un efecto directo sobre la conducta, especialmente cuando existe una alta correspondencia entre percepción y realidad (Ajzen, 2002; Chen & Slade, 2024). De esta forma, la TPB ofrece un nuevo marco analítico para explicar por qué una persona decide actuar de una manera determinada, en función de sus creencias, valores y contexto social.

2.1.3. Modelo de Aceptación de la Tecnología (TAM).

Mas concretamente en el contexto de la digitalización, Davis (1989), adapta la TRA al ámbito de los sistemas de información incluyendo la utilidad y facilidad de uso percibidas. Fred Davis en la década de 1980 en respuesta a las preocupaciones sobre la resistencia a la tecnología y las frecuentes fallas de los sistemas, desarrolla el Modelo de Aceptación de la Tecnología o más comúnmente conocido en inglés como TAM por sus siglas en inglés (*Technology Acceptance Model*).

Se trata de un marco teórico diseñado para predecir la probabilidad de que individuos u organizaciones adopten nuevos sistemas tecnológicos. Este modelo tiene como principal objetivo poder explicar y predecir la aceptación de la tecnología y proporcionar una explicación teórica para la implementación exitosa, centrándose en las percepciones de los usuarios sobre la tecnología.

Este modelo se erige mediante dos constructos centrales, la Utilidad Percibida (PU) y la Facilidad de Uso Percibida (PEOU). Respecto a la variable de La Utilidad Percibida, se define como el grado en que una persona cree que el uso de un sistema en particular mejoraría su desempeño laboral. Esta creencia de que la tecnología será beneficiosa y útil para realizar tareas predice de cerca la probabilidad de que los usuarios acepten y adopten el nuevo sistema, y está influenciada por factores externos como las normas sociales y el conocimiento previo. Por otro lado, respecto a la Facilidad de Uso Percibida, es el grado en que una persona cree que el uso de un sistema en particular estaría libre de esfuerzo, es decir, el grado en que un individuo espera que la tecnología sea fácil de aprender y operar. Si la tecnología es fácil de usar, se supera la barrera de adopción, y además, influye en la percepción de utilidad; los sistemas más fáciles de usar tienen más probabilidades de ser vistos como útiles.

2.1.4. Teoría Unificada de Aceptación y Uso de la Tecnología (Unified Theory of Acceptance and Use of Technology - UTAUT).

Del mismo modo que la Teoría de la Acción Racionada fue revisada y actualizada a nuevos contextos y variables explicativos, la Teoría Unificada de Aceptación y Uso de la Tecnología (Venkatesh, 2003), surge como un esfuerzo integrador que consolida ocho modelos previos de aceptación tecnológica, entre ellos el Modelo de Aceptación de la Tecnología (TAM), la Teoría del Comportamiento Planificado (TPB) y la Teoría de Difusión de Innovaciones (IDT). Esta teoría tiene como objetivo explicar las intenciones de los usuarios de utilizar un sistema de información y el comportamiento de uso posterior en un contexto organizacional, buscando proporcionar una comprensión más holística de la aceptación de la tecnología (Venkatesh et al., 2003). Se ha demostrado que UTAUT supera a los ocho modelos individuales en los que se basó (Dwivedi, Rana, Chen. & Williams, 2011; Williams, Rana & Dwivedi, 2015; Venkatesh, Thong & Xu, 2016; Momani, 2020).

Los cuatro constructos principales que explican la intención y el uso de la tecnología en UTAUT son: (1) la expectativa de desempeño, (2) la expectativa de esfuerzo, (3) la influencia social y (4) las condiciones facilitadoras (Venkatesh et al., 2003; Dwivedi et al., 2019). La expectativa de desempeño se refiere al grado en que el

usuario cree que el uso del sistema mejorará su rendimiento laboral. Este constructo es análogo a la utilidad percibida en TAM y es considerado uno de los predictores más fuertes de la intención de uso (Dwivedi et al., 2020). La expectativa de esfuerzo representa la facilidad percibida en el uso de la tecnología, y su influencia suele disminuir con el tiempo conforme los usuarios adquieren experiencia (Venkatesh et al., 2003). Por su parte, la influencia social hace referencia al grado en que los individuos perciben que personas importantes para ellos creen que deben utilizar el sistema, siendo particularmente relevante en contextos donde el uso de la tecnología es obligatorio (Venkatesh et al., 2003). Finalmente, las condiciones facilitadoras incluyen la percepción del usuario sobre la existencia de una infraestructura técnica y organizacional que respalde el uso del sistema, como acceso a internet, soporte técnico o formación (Williams et al., 2015).

El modelo UTAUT también incorpora variables moderadoras como la edad, el género, la experiencia previa con la tecnología y la voluntariedad de uso, las cuales pueden influir en la fuerza de relación entre los constructos principales y la intención o el uso de la tecnología (Venkatesh et al., 2003; Dwivedi et al., 2020). A lo largo de los años, UTAUT ha sido actualizada y revisada. En su segunda versión, se incorporaron nuevos constructos como el valor hedónico, el hábito y el valor del precio, y se adapta mejor a contextos de consumo individual (Venkatesh et al., 2012). Además, se han propuesto otras extensiones que incluyen variables como la confianza, la privacidad, el apoyo social, la autogestión del aprendizaje o la ansiedad tecnológica (Dwivedi et al., 2019; Alalwan et al., 2017).

El modelo UTAUT ha sido ampliamente utilizado para estudiar la adopción de tecnologías móviles, plataformas de comercio electrónico, servicios de gobierno digital, tecnologías en el ámbito de la salud, e incluso herramientas emergentes como ChatGPT (Tam & Oliveira, 2017; Alzahrani & Alzahrani, 2025).

A pesar de su robustez y poder explicativo superior respecto a modelos previos, UTAUT también presenta limitaciones. Se ha criticado su complejidad y la necesidad de grandes cantidades de datos para validarlo empíricamente (Venkatesh et al., 2012; Bagozzi, 2007). Otros autores han señalado que la teoría asume relaciones lineales entre variables que pueden no reflejar adecuadamente la realidad (Jeyaraj

et al., 2006).

2.1.5. Modelo de Difusión de Innovaciones de Rogers (2003).

El Modelo de Difusión de Innovaciones, desarrollado por Everett M. Rogers, constituye uno de los marcos teóricos para explicar cómo, por qué y a qué ritmo se difunden las innovaciones dentro de un sistema social (Rogers, 1995). La difusión se conceptualiza como el proceso mediante el cual una innovación es comunicada a través de ciertos canales durante un período de tiempo entre los miembros de dicho sistema (Rogers, 2003). Este modelo se fundamenta en cuatro elementos clave: la innovación, los canales de comunicación, el tiempo y el sistema social (Rogers, 2003; Sahin, 2006).

El proceso de adopción se despliega a lo largo de cinco etapas, caracterizadas por diferentes perfiles de adoptadores. Estos incluyen: innovadores (2.5%), adoptadores tempranos (13.5%), mayoría temprana (34%), mayoría tardía (34%) y rezagados (16%) (Rogers, 2003). Por ende, lo que Rogers hipotetiza es que no todos los individuos adoptan las innovaciones al mismo tiempo, ya que sus decisiones están influidas por factores como la apertura al cambio, el liderazgo de opinión o el escepticismo frente a nuevas tecnologías (Rogers, 2003). Así mismo, identifica cinco atributos de la innovación que influyen directamente en su tasa de adopción: (1) ventaja relativa, entendida como el grado en que la innovación es percibida como superior a las alternativas existentes; (2) compatibilidad con los valores, experiencias y necesidades del usuario; (3) complejidad, o dificultad percibida de uso; (4) posibilidad de prueba previa a su adopción completa; y (5) observabilidad, es decir, el grado en que los resultados de la innovación son visibles para otros.

No obstante su influencia, el modelo de difusión de innovaciones también ha sido objeto de diversas críticas. Una de las más recurrentes es su supuesto pro-innovación, que parte de la premisa de que toda innovación es inherentemente positiva y debe ser adoptada, sin considerar que algunas pueden ser rechazadas de forma legítima por los usuarios (MacVaugh & Schiavone, 2010; Shah, 2025). Además, se ha señalado una tendencia a responsabilizar al individuo por la no

adopción, omitiendo factores estructurales como barreras tecnológicas, sociales o de aprendizaje, que pueden limitar la capacidad de adopción más allá del control del usuario (Sahin, 2006; Selwyn, 2003). Otro conjunto de críticas apunta a limitaciones metodológicas, como el uso de estudios retrospectivos con sesgo de recuerdo o la dificultad para analizar empíricamente procesos complejos de difusión que involucran múltiples actores y niveles (MacVaugh & Schiavone, 2010). Finalmente, desde una perspectiva sistémica, se ha indicado que el modelo no integra de forma adecuada la superposición entre los diferentes dominios contextuales (individuo, comunidad e industria), lo que limita su capacidad explicativa en entornos complejos y cambiantes. En este sentido, se considera que el modelo es útil como marco secuencial, pero insuficiente para capturar las interacciones dinámicas entre condiciones y actores (MacVaugh & Schiavone, 2010).

2.1.6. Technology Readiness Index (TRI).

Además de los modelos centrados en la intención de uso o en factores contextuales, resulta pertinente incorporar enfoques que profundicen en las disposiciones individuales previas al contacto con la tecnología. En este sentido, se destaca el Technology Readiness Index (TRI)**,** el cual ofrece una perspectiva complementaria orientada a comprender las diferencias individuales en la predisposición hacia la adopción tecnológica.

El TRI, desarrollado por A. Parasuraman en el año 2000, constituye un modelo psicométrico diseñado para medir el grado de preparación psicológica de los individuos hacia el uso y adopción de tecnologías nuevas. A diferencia de otros modelos centrados en el contexto organizacional o en el análisis del comportamiento intencional (como el TAM o el UTAUT), el TRI se enfoca en los rasgos de personalidad y actitudes generales de los usuarios hacia la tecnología (Parasuraman, 2000).

El índice identifica cuatro dimensiones fundamentales que conforman la predisposición tecnológica del individuo: (1) optimismo, (2) innovatividad, (3) disconformidad e (4) inseguridad. Las dos primeras son consideradas impulsores (drivers), mientras que las dos últimas actúan como inhibidores (barriers) de la

aceptación tecnológica. La variable optimismo se refiere a la creencia positiva de que la tecnología puede contribuir de manera significativa a mejorar la vida cotidiana, al aumentar el control personal, ofrecer mayor flexibilidad y facilitar la eficiencia en diversas actividades. Este rasgo refleja una actitud general de confianza hacia los beneficios potenciales que ofrecen las innovaciones tecnológicas. Por su parte, la innovatividad representa la tendencia del individuo a ser pionero o líder en la adopción de nuevas tecnologías. Las personas con altos niveles de innovatividad suelen mostrar entusiasmo por experimentar con herramientas emergentes y, con frecuencia, actúan como influenciadores dentro de sus entornos sociales o profesionales. En contraste, la dimensión de disconformidad se relaciona con la percepción de que los sistemas tecnológicos resultan complejos, poco intuitivos o difíciles de controlar. Este sentimiento puede generar frustración o rechazo hacia el uso de tecnologías que no se perciben como accesibles o amigables para el usuario. Finalmente, la variable inseguridad hace alusión a la desconfianza o al temor frente a los posibles efectos negativos del uso de la tecnología, en particular en aspectos relacionados con la privacidad, la fiabilidad de los sistemas o la dependencia excesiva de dispositivos tecnológicos (Parasuraman & Colby, 2001). Estas dimensiones permiten segmentar a los usuarios según su nivel de disposición tecnológica, lo que resulta útil para diseñar estrategias de implementación y comunicación más efectivas. Por ejemplo, un usuario con alta innovatividad pero también con elevada inseguridad puede necesitar una mayor garantía de confianza antes de adoptar una tecnología específica (Walczuch et al., 2007).

Sin embargo, el modelo también ha recibido críticas por su enfoque relativamente estable y disposicional, ya que las actitudes generales pueden no reflejar adecuadamente la variabilidad del comportamiento en contextos específicos de adopción. También se ha cuestionado su validez transcultural, dado que las percepciones de tecnología pueden diferir significativamente según factores socioculturales y económicos (Walczuch et al., 2007; Lin & Hsieh, 2007).

3.2. Factores asociados a la aceptación social tecnológica en el sistema de justicia penal.

La incorporación de tecnologías emergentes en ámbitos tradicionalmente conservadores como el sistema judicial exige un análisis que vaya más allá de las implicaciones técnicas y normativas, e incluya las actitudes y la aceptación social y profesional de estas innovaciones. En este escenario, donde confluyen el factor humano y el progreso tecnológico, tanto la ciudadanía como los operadores del sistema penal muestran actitudes ambivalentes frente al proceso de transformación tecnológica (Miró Llinares, 2020), lo que fuerza a valorar el grado de aceptación del proceso no solo desde la óptica de jueces, fiscales, abogados y funcionarios, sino también desde la experiencia de quienes interactúan con las instituciones judiciales.

La transformación digital de la justicia penal no puede entenderse como la mera incorporación de dispositivos o aplicaciones, pues constituye un proceso complejo que afecta a la organización institucional, a los procedimientos y a los principios fundamentales que sostienen la función jurisdiccional. La introducción de sistemas de inteligencia artificial y algoritmos predictivos supone, en este sentido, un cambio de paradigma que trasciende la modernización administrativa y plantea retos sustantivos vinculados con la legitimidad, la transparencia y la protección de los derechos fundamentales (de Fine Licht & de Fine Licht, 2020; He & Zhang, 2025; Levy, Chasalow & Riley, 2021). El actual proceso de modernización de la justicia requiere de la identificación de los factores que determinan el éxito o el fracaso en la incorporación de nuevas tecnologías, pues la mera disponibilidad de recursos no asegura por sí sola una implementación eficaz ni una aceptación plena entre los distintos operadores jurídicos y la ciudadanía. Para que la transformación digital pueda integrarse de manera adecuada en los ámbitos judicial, policial y penitenciario, es necesario que se desarrolle bajo condiciones que garanticen tanto su funcionalidad como su compatibilidad con los principios esenciales del Estado de Derecho. Estas exigencias constituyen el fundamento de un ecosistema digital que no debe limitarse a incrementar la eficiencia operativa, sino que ha de contribuir también a preservar la calidad de las decisiones y a fortalecer la confianza pública en el sistema penal.

Los marcos teóricos y la literatura especializada ofrecen instrumentos valiosos para analizar las dinámicas de adopción tecnológica en el ámbito judicial, tomando en consideración tanto experiencias internacionales como el contexto español. La relevancia de estos enfoques se aprecia en la forma en que la digitalización del sistema de justicia penal ha dejado de ser un planteamiento abstracto para convertirse en una realidad con aplicaciones concretas: herramientas algorítmicas sustentadas en el análisis de datos, el aprendizaje automático y la modelización estadística se emplean hoy en distintas fases del proceso penal con el propósito de optimizar la gestión de recursos, apoyar la toma de decisiones y elaborar predicciones sobre riesgos o patrones de comportamiento (Berk, 2019; Kleinberg et al., 2018). Este desarrollo confirma que la aceptación de tales innovaciones no depende únicamente de su precisión matemática o capacidad de análisis, sino de un entramado más amplio de condicionantes técnicos, institucionales, éticos, culturales y sociales que determinan, en última instancia, su legitimidad y sostenibilidad. En este marco, la literatura sobre adopción tecnológica constituye un primer punto de referencia para comprender esta complejidad, como muestran modelos clásicos de la disciplina, como la Teoría de la Acción Razonada (Fishbein y Ajzen, 1975), la Teoría del Comportamiento Planificado (Ajzen, 1991) o el Modelo de Aceptación Tecnológica (Davis, 1989), que destacan la importancia de variables como la utilidad percibida, la facilidad de uso, la actitud hacia la innovación y la influencia de las normas sociales en la disposición a emplear nuevas herramientas. Trasladados al ámbito judicial, estos enfoques han permitido examinar cómo los profesionales valoran las oportunidades y limitaciones de la digitalización, así como los obstáculos que condicionan su incorporación (Wirtz et al., 2023; Oswald, 2020). No obstante, aunque tales modelos resultan útiles, el campo penal introduce exigencias adicionales vinculadas a la legitimidad democrática de las decisiones, el respeto a los principios jurídicos y la protección de la dignidad y los derechos fundamentales, lo que obliga a superar una visión meramente instrumental e integrar dimensiones éticas e institucionales propias del derecho penal y procesal.

En este marco, diversos estudios coinciden en que la aceptación social de los algoritmos judiciales depende en gran medida de la percepción de imparcialidad y transparencia de los sistemas (Lee, 2018). Un diseño opaco o de difícil acceso tiende

a generar desconfianza, mientras que la posibilidad de comprender y verificar su funcionamiento, en coherencia con los principios jurídicos, favorece actitudes más positivas hacia su uso. La confianza en las instituciones que desarrollan y supervisan estas herramientas constituye, asimismo, un elemento decisivo: cuando la ciudadanía percibe que los organismos responsables actúan con legitimidad y responsabilidad, aumenta la disposición a aceptar la mediación algorítmica en los procesos de decisión (van den Bos et al., 1998). A ello se suma la importancia de la supervisión humana, pues la investigación muestra que las decisiones en las que un juez valida o revisa la recomendación del sistema se consideran más legítimas que aquellas totalmente automatizadas, lo que refuerza la idea de que los modelos *human in the loop* representan un punto de equilibrio entre eficiencia y responsabilidad (Rahwan et al., 2019; Starke et al., 2022; Green & Chen, 2019). La percepción de imparcialidad constituye otro de los ejes centrales en este debate. Si los algoritmos se entienden como justos, objetivos y libres de sesgos, la aceptación se incrementa (Binns, 2018). Sin embargo, la evidencia empírica demuestra que los sistemas algorítmicos no están exentos de reproducir desigualdades estructurales y, en algunos casos, de amplificarlas. El conocido estudio de Angwin et al. (2016) sobre el sistema COMPAS en Estados Unidos mostró que los algoritmos de predicción del riesgo de reincidencia replicaban sesgos raciales, atribuyendo mayor peligrosidad a acusados afroamericanos en comparación con acusados blancos en condiciones similares. Estos hallazgos reflejan que la percepción de justicia no depende únicamente de la precisión técnica, sino de la garantía de que los sistemas no refuercen prejuicios históricos.

Otro aspecto señalado por la literatura es el nivel de familiaridad y el conocimiento de los usuarios con la tecnología. Horowitz, Kahn, Macdonald y Schneider (2024) muestran que la familiaridad con la inteligencia artificial incrementa el apoyo a su adopción en ámbitos civiles como los vehículos autónomos, la cirugía y la ciberdefensa, aunque no en el caso de las armas autónomas, donde prevalece el rechazo. Asimismo, identifican una brecha entre el respaldo a nivel de política pública y la disposición al uso personal, lo que revela que la confianza generada por la experiencia puede coexistir con recelos frente a los riesgos individuales. Por su parte, Schiavo, Businaro y Zancanaro (2024) evidencian que la alfabetización en

inteligencia artificial influye de manera significativa y positiva en la aceptación de estas tecnologías, tanto de forma directa como a través de un aumento en la percepción de su utilidad y facilidad de uso. Asimismo, muestran que la ansiedad vinculada a la IA ejerce un efecto negativo, aunque limitado, sobre la aceptación, especialmente en sus dimensiones de aprendizaje y ceguera sociotécnica. En este sentido, los autores concluyen que la ansiedad actúa como un mediador parcial complementario: una parte del impacto positivo de la alfabetización sobre la aceptación se canaliza mediante la reducción de la ansiedad, si bien la alfabetización mantiene un efecto independiente sobre la disposición a aceptar la IA. En un sentido complementario, Lin y Hsieh (2007) argumentan que la preparación tecnológica de los usuarios constituye un factor determinante que incide de forma directa tanto en su nivel de satisfacción como en su intención de uso continuado de las tecnologías de autoservicio. Este hallazgo revela que la aceptación de dichas herramientas no depende únicamente de las características técnicas del sistema, sino también de la actitud y disposición previa de los individuos que interactúan con ellas. Por su parte, Kelly, Kaye y Oviedo-Trespalacios (2023) sostienen que variables como la utilidad percibida, la facilidad de uso, la expectativa de rendimiento y la confianza depositada en la tecnología constituyen factores determinantes en los procesos de aceptación de la inteligencia artificial. Sin embargo, advierten que en aquellos contextos donde el contacto humano resulta especialmente valorado, la incorporación de sistemas algorítmicos tiende a generar resistencias significativas. Esta evidencia pone de relieve que la aceptación de la IA no responde a un patrón homogéneo, sino que se configura a partir de la interacción dinámica entre las propiedades técnicas de la herramienta y las expectativas culturales y sociales del entorno en el que se implementa.

La confianza en la tecnología aparece, además, como un componente transversal que articula múltiples dimensiones de este proceso. Según Choung, David y Ross (2022), la confianza no ejerce un efecto directo sobre la aceptación de la inteligencia artificial, sino que opera de forma mediada, al mejorar tanto la percepción de utilidad como la actitud hacia la tecnología. Asimismo, los autores distinguen entre dos dimensiones de la confianza en IA: una asociada a los aspectos humanizados de la tecnología y otra vinculada a su funcionalidad técnica, basada en su fiabilidad y

rendimiento. Los resultados de su investigación muestran que, aunque ambas dimensiones contribuyen a la aceptación, la confianza en la funcionalidad tiene un peso relativamente mayor en la disposición de los usuarios a adoptar estas tecnologías. En el ámbito judicial, Yalcin et al. (2023) ofrecen evidencia adicional: aunque los jueces algorítmicos son valorados por su rapidez y bajo coste, los usuarios expresan una clara preferencia por jueces humanos, especialmente en casos de alta carga emocional. Ello refleja que la confianza funcional, asociada al rendimiento técnico, puede ser insuficiente cuando las decisiones involucran dimensiones humanas y éticas complejas. En esta misma línea, el metaanálisis de Kuen, Westmattelmann, Bruckes y Schewe (2023) muestra que la confianza en la tecnología y la confianza en el proveedor constituyen entidades diferenciadas pero interrelacionadas, ambas relevantes para explicar la intención de uso en entornos digitales. Si bien los dos tipos de confianza favorecen la adopción, la evidencia empírica indica que, cuando se consideran de manera conjunta, la confianza en la tecnología adquiere un peso mayor que la confianza en el proveedor, especialmente en lo relativo a la percepción de fiabilidad y rendimiento del sistema.

Otro de los factores clave en la aceptación social es el modo en que se combinan la intervención humana y la automatización, Hermstrüwer y Langenbach (2023) muestran, mediante un experimento en los ámbitos de la policía predictiva, las admisiones escolares y la asignación de refugiados, que los procedimientos híbridos con alta supervisión humana son percibidos como los más justos, mientras que las decisiones puramente algorítmicas reciben las valoraciones más bajas de equidad procedimental. Las decisiones exclusivamente humanas y las híbridas con baja participación se sitúan en posiciones intermedias, lo que indica que la ciudadanía acepta cierta delegación siempre que no se elimine la agencia humana.

Más allá de los aspectos técnicos y actitudinales, los factores psicológicos y motivacionales también resultan determinantes. El estudio de Bergdahl et al. (2023) evidencia que la satisfacción de las necesidades psicológicas básicas, entendidas como los requerimientos universales de autonomía, competencia y relación formulados por la Self-Determination Theory, se vincula con actitudes más positivas hacia la inteligencia artificial. En particular, la competencia y la relación mostraron

asociaciones consistentes con una mayor valoración positiva de la IA, mientras que la autonomía presentó un efecto favorable solo en determinados contextos, como en Finlandia, y en el análisis longitudinal se relacionó con un incremento de la positividad y una disminución de la negatividad hacia la IA.

A esta complejidad se añaden las diferencias culturales y sociodemográficas, que condicionan de forma significativa las percepciones hacia la inteligencia artificial. La evidencia empírica muestra que variables como la nacionalidad, el género, el nivel educativo o la religiosidad inciden de forma significativa en los niveles de confianza depositados en la inteligencia artificial (González-Anleo, Delbello, Martínez-Gonzálo & Gómez, 2024; Kim & Lee, 2024; Kozak & Fel, 2024). Esto significa que los modelos de implementación no pueden ser uniformes, sino que deben adaptarse a los contextos sociales en los que se despliegan.

La investigación existente evidencia que la aceptación de herramientas algorítmicas en la justicia penal constituye un fenómeno complejo, de naturaleza multidimensional y dinámica. En este sentido, resulta necesario atender a factores como la transparencia y la explicabilidad de los sistemas, la imparcialidad percibida en sus resultados, el grado de supervisión humana que incorporan, el nivel de familiaridad tecnológica de los usuarios, su disposición psicológica ante la innovación y, finalmente, las diferencias culturales y sociodemográficas que condicionan las actitudes hacia estas tecnologías.

PARTE II. ESTUDIOS SOBRE EL FACTOR HUMANO EN LA TRANSFORMACIÓN TECNOLÓGICA DEL SISTEMA DE JUSTICIA PENAL.

CAPÍTULO 3. JUSTIFICACIÓN, OBJETIVOS Y ENFOQUE CIENTÍFICO.

1. Justificación.

La presente tesis doctoral se justifica en la necesidad de analizar de manera crítica y sistemática el papel del factor humano en la transformación tecnológica del sistema de justicia penal. La irrupción de la inteligencia artificial y de herramientas algorítmicas en este ámbito ha abierto la posibilidad de optimizar procesos, reducir tiempos y ofrecer una aparente objetividad en la toma de decisiones (Hildebrandt, 2020; Wischmeyer & Rademacher, 2020; Ramos-Maqueda & Chen, 2025; Rosili et al., 2021; van den Bos, 2001; Wischmeyer & Rademacher, 2020). Sin embargo, estas promesas tecnológicas no pueden evaluarse de forma aislada, ya que su éxito y legitimidad dependen en gran medida de la interacción con quienes las utilizan y con quienes reciben sus efectos. En este sentido, el factor humano constituye el núcleo de esta tesis, entendido tanto desde la perspectiva de los operadores jurídicos (jueces, fiscales, policías y profesionales penitenciarios), como desde la ciudadanía, destinataria final de las decisiones judiciales.

La literatura especializada ha centrado gran parte de su atención en los marcos normativos, en los riesgos de sesgos algorítmicos o en los principios éticos que deben guiar la justicia digital. No obstante, se ha relegado a un segundo plano el análisis empírico de cómo las personas perciben, aceptan y utilizan estas herramientas en su práctica cotidiana. Esta omisión resulta especialmente significativa, ya que son las actitudes, creencias y experiencias de los actores humanos las que determinan en última instancia si la incorporación de estas tecnologías será efectiva, legítima y sostenible. Aunque se han desarrollado investigaciones centradas en la implementación de sistemas algorítmicos en los ámbitos judicial y penitenciario, persisten vacíos notables en torno a la comprensión de cómo los operadores jurídicos perciben, interpretan y utilizan estas

herramientas, así como respecto al modo en que su aceptación y las implicaciones prácticas derivadas de su uso condicionan la efectividad de su implementación.

La psicología social ofrece un marco valioso para comprender esta cuestión. Las actitudes, concebidas como representaciones mentales aprendidas que condensan evaluaciones favorables o desfavorables sobre objetos, ideas o personas, orientan el comportamiento y pueden condicionar la adopción o rechazo de innovaciones tecnológicas. Estas no surgen espontáneamente, sino que se construyen socialmente mediante procesos de aprendizaje, socialización y experiencia directa, y pueden ser tanto explícitas como implícitas, lo que implica que los profesionales del derecho pueden albergar percepciones conscientes e inconscientes hacia los sistemas algorítmicos que influyen en la forma en que interpretan y aplican los resultados generados por estas herramientas. En consecuencia, las actitudes hacia las tecnologías representan un punto de partida determinante para su aceptación o rechazo.

En el contexto español, los estudios que abordan de manera integral el impacto de la digitalización en la justicia penal desde una perspectiva centrada en el factor humano siguen siendo escasos, lo que genera un vacío de conocimiento que esta tesis busca cubrir. La interacción entre los operadores jurídicos y las herramientas algorítmicas presenta una complejidad particular, modulada por variables como la confianza, la percepción de imparcialidad, el nivel de comprensión tecnológica o la atribución de responsabilidad. Estas dinámicas no solo condicionan la aceptación de las tecnologías, sino también su legitimidad democrática, en tanto que las decisiones adoptadas inciden directamente en derechos fundamentales. En este sentido, la tesis adopta un enfoque multidimensional orientado a analizar la relación entre el factor humano y el proceso de digitalización, atendiendo a cómo los distintos actores perciben estas transformaciones, qué actitudes desarrollan ante ellas y de qué modo las incorporan en su labor cotidiana dentro del sistema de justicia penal.

Mediante metodologías cualitativas y cuantitativas, la investigación abordará las percepciones, experiencias y actitudes de los operadores jurídicos y de la ciudadanía respecto al uso de herramientas algorítmicas y de inteligencia artificial en el sistema de justicia penal. El análisis parte de la premisa de que para que la tecnología sea

incorporada en la práctica profesional requiere ser comprendida, supervisada y aceptada por quienes la emplean y por la sociedad que experimenta sus efectos. Además, el carácter aplicado de la tesis, desarrollada en colaboración con la empresa Plus Ethics, permite situar el análisis en la práctica real del uso de algoritmos, generando un conocimiento transferible que contribuye tanto al ámbito académico como a la formulación de políticas públicas y a la promoción de una gobernanza responsable de la justicia digital.

En consecuencia, la justificación de esta tesis se articula en tres dimensiones interrelacionadas que, en conjunto, buscan ofrecer una comprensión integral del fenómeno. Desde una perspectiva teórica, propone un marco interdisciplinar que vincula los avances tecnológicos con el estudio del factor humano como condición esencial para la aceptación profesional y social de la justicia digital, mientras que, en el plano empírico, aporta evidencia sobre las percepciones y actitudes de operadores jurídicos y ciudadanía en España, un ámbito aún poco explorado en la literatura. Finalmente, desde una dimensión profesional, orienta la elaboración de políticas regulatorias, estrategias institucionales y procesos de diseño tecnológico que reconozcan al factor humano como el eje sobre el cual puede sostenerse una transformación responsable del sistema de justicia penal.

En suma, esta investigación se justifica porque busca responder a una pregunta crucial: ¿cómo se reconfigura el sistema de justicia penal cuando la tecnología, no solo modifica procedimientos, sino que redefine el papel de las personas en la práctica profesional? Lejos de ser un elemento secundario, el factor humano es el verdadero mediador entre la promesa de la innovación tecnológica y la necesidad de preservar los principios fundamentales del Estado de derecho.

2. Objetivos.

El presente trabajo tiene como fin contribuir al conocimiento científico sobre la transformación del sistema de justicia penal en el contexto de la digitalización y el uso de nuevas tecnologías como la inteligencia artificial. Para ello, se plantean los siguientes objetivos generales:

I. Analizar los posibles sesgos existentes en el uso algoritmos predictivos en la toma de decisiones de diferentes operadores del sistema de justicia penal.
II. Explorar las actitudes y las implicaciones de la inteligencia artificial y la digitalización en la ciudadanía y los profesionales del sistema de justicia penal.

Con el propósito de alcanzar los objetivos generales previamente expuestos, se desarrollarán los siguientes objetivos específicos:

I. Analizar la incidencia del sesgo humano en la toma de decisiones judiciales en el sistema de justicia penal, así como en el contexto de la digitalización.
II. Explorar las actitudes sociales hacia el uso de herramientas algorítmicas en el sistema de justicia entre la población española y los profesionales.
III. Detectar los desafíos y el impacto del uso de algoritmos predictivos en operadores judiciales y policiales del sistema de justicia penal.
IV. Diseñar una herramienta que permita evaluar la aceptación social y los riesgos éticos del uso de la inteligencia artificial en la justicia penal.

El marco teórico desarrollado en los capítulos 1 y 2 proporciona el sustento necesario para abordar este primer objetivo específico, el cual tiene como finalidad realizar una revisión exhaustiva de la literatura científica y técnica en torno a la implementación de herramientas algorítmicas en el sistema de justicia penal.

Para cumplir con este objetivo, se llevó a cabo una revisión sistemática que permite mapear las principales líneas de investigación, tipos de herramientas implementadas, áreas del sistema penal en las que se han aplicado, y los marcos regulatorios que las acompañan. Este análisis, expuesto en el capítulo 4, también permitió identificar los países con mayor grado de desarrollo en esta materia, así

como destacar los debates éticos, técnicos y jurídicos que rodean el fenómeno de la justicia algorítmica. El trabajo realizado permite situar la investigación dentro de un contexto global, ofreciendo una base comparativa sobre la cual analizar los siguientes objetivos de la tesis.

En lo que respecta al objetivo específico II, este se sustenta en el marco conceptual desarrollado en el Capítulo 2, que integra el análisis de las teorías de aceptación tecnológica y la literatura sobre las actitudes ciudadanas ante el uso de la inteligencia artificial en contextos institucionales. Asimismo, incorpora las aportaciones de la psicología y del derecho, orientadas a comprender los factores que pueden influir o distorsionar la neutralidad de los procesos de toma de decisiones.

Para su desarrollo, se diseñó una serie de estudios cuantitativos a nivel nacional, cuyos resultados se presentan en el capítulo 6, con el fin de identificar los niveles de aceptación, rechazo o desconocimiento de la ciudadanía respecto al uso de algoritmos en el ámbito judicial. Además, el estudio indaga en la relación entre variables sociodemográficas y actitudes hacia la automatización del sistema penal. El análisis realizado permite delinear los perfiles y variables asociadas a la aceptación de este tipo de herramientas, permitiendo a partir de ahí, explorar cuales son las mejores estrategias para una implementación acorde a las demandas sociales.

El objetivo específico III se fundamenta en los marcos teóricos desarrollados en los Capítulos 1 y 2, donde se analizan los debates éticos y normativos derivados de la incorporación de tecnologías predictivas en el sistema de justicia penal. Para su abordaje, se llevaron a cabo diferentes estudios con operadores jurídicos, policiales y penitenciarios empleando metodologías cualitativas que permitieron explorar las demandas, inquietudes y retos que estos actores identifican ante el uso de algoritmos en su práctica profesional.

Asimismo, se examinaron los efectos subjetivos y organizacionales asociados a la introducción de estas herramientas, tales como la percepción de pérdida de autonomía, el nivel de confianza depositado en los sistemas automatizados y las

transformaciones en los roles y dinámicas profesionales. Los resultados obtenidos constituyen una base empírica para la formulación de recomendaciones orientadas a una implementación ética y responsable de la digitalización en el ámbito judicial y penitenciario.

Finalmente, y dado el carácter industrial de la presente tesis, el último objetivo se orienta a la explotación de los resultados en el ámbito empresarial, mediante la transferencia y aplicación práctica del conocimiento generado. Con este fin, se ha diseñado la herramienta, un instrumento metodológico estructurado que permite evaluar la aceptación del uso de sistemas de inteligencia artificial por las administraciones públicas. La herramienta facilita a empresas, instituciones públicas y entidades de innovación identificar y gestionar los factores que condicionan la implementación de tecnologías basadas en inteligencia artificial. De este modo, este objetivo industrial no solo busca la validación práctica de los resultados científicos, sino también su transferencia efectiva al mercado, reforzando las capacidades de consultoría ética y social de la empresa colaboradora y contribuyendo al desarrollo de soluciones que permitan avanzar a la integración de nuevas tecnologías.

3. Enfoque científico.

A continuación, se presenta una síntesis general de la metodología empleada en el desarrollo de la presente investigación. Dado que el estudio integra diversos enfoques, diseños y técnicas de análisis, este apartado tiene como propósito ofrecer una visión global de los procedimientos metodológicos aplicados. Los aspectos específicos de cada técnica, junto con los instrumentos utilizados y los criterios establecidos para la recopilación y el análisis de datos, se describen de forma detallada en los capítulos correspondientes.

3.1. Marco metodológico de la investigación.

El marco metodológico de esta investigación expone el itinerario seguido para abordar el objeto de estudio, integrando distintas fases y enfoques metodológicos. Se adoptó un enfoque mixto, combinando metodologías cualitativas y cuantitativas para analizar de forma integral los sesgos en la toma de decisiones en el sistema de

justicia penal y el impacto derivado del uso de algoritmos predictivos.

El recorrido comenzó con una revisión sistemática de la literatura, orientada a identificar y analizar los sesgos presentes en operadores de justicia, tanto en contextos sin digitalización como en escenarios condicionados por la transformación digital. Esta fase permitió construir un marco teórico y conceptual sólido, detectar vacíos de conocimiento y guiar el diseño de los estudios empíricos posteriores. De forma complementaria, se desarrollaron dos líneas principales: dos estudios de metodología cualitativa (Capítulo 5) mediante grupos nominales, destinado a explorar percepciones, experiencias y valoraciones de expertos y operadores sobre el uso de herramientas algorítmicas; y tres estudios cuantitativos (Capítulo 6) basado en encuestas y casos escenario, diseñado para evaluar las actitudes ciudadanas y profesionales en el uso de herramientas algorítmicas e IA en el sistema de justicia penal.

3.2. Diseño y técnicas de investigación.

En coherencia con lo expuesto en el apartado anterior, la presente investigación emplea diversos enfoques metodológicos adaptados a las particularidades de cada uno de sus objetivos, en concreto, se han empleado las siguientes técnicas de investigación:

Para el Capítulo 4, se llevó a cabo una revisión sistemática de literatura, siguiendo las directrices PRISMA (Page et al., 2021). Este estudio tuvo como finalidad examinar la producción científica existente sobre los sesgos en la toma de decisiones dentro del sistema de justicia penal y evaluar el impacto del uso de algoritmos en este contexto. Se siguieron todas las directrices necesarias para su consideración como revisión sistemática PRISMA.

En el Capítulo 5, se desarrolló un diseño cualitativo basado en la técnica de grupos nominales (Tashakkori & Creswell, 2007; Zanón, 1990), con el objetivo de explorar el impacto de la digitalización y algoritmización en la administración de justicia y la seguridad ciudadana. En el Estudio 1, se reunieron expertos del ámbito de la seguridad para identificar y priorizar los principales desafíos que plantea la incorporación de herramientas digitales en este sector. En el Estudio 2, se aplicó la

misma metodología con profesionales del ámbito judicial, analizando sus percepciones sobre la transformación digital y la irrupción de la inteligencia artificial en la práctica profesional dentro del sistema judicial.

En el Capítulo 6, se adoptó un diseño de métodos mixtos con una fase experimental y una fase empírica no experimental para analizar las actitudes sociales hacia la justicia algorítmica. En el Estudio 1 y el Estudio 2, se aplicaron encuestas estructuradas para identificar barreras percibidas en la implementación de tecnologías de inteligencia artificial en el sistema judicial y explorar las concepciones sociales sobre el uso de algoritmos y sistemas autónomos en la toma de decisiones legales. En el Estudio 3, se diseñó un experimento en el que se expuso a los participantes a escenarios de toma de decisiones judiciales, diferenciando entre decisiones humanas y algorítmicas, con el objetivo de evaluar la aceptación social de estas herramientas.

La combinación de estos enfoques metodológicos ha permitido obtener una visión integral del fenómeno en estudio, en línea con los objetivos propuestos. En particular, esta estrategia ha facilitado tanto el análisis de los posibles sesgos asociados al uso de algoritmos predictivos en la toma de decisiones por parte de distintos operadores del sistema de justicia penal (Objetivo I) como la exploración de las actitudes, percepciones e implicaciones que la digitalización y la inteligencia artificial suscitan en la ciudadanía y en los profesionales del ámbito penal (Objetivo II).

3.3. Instrumentos.

Los instrumentos de recopilación de datos utilizados en cada una de las diferentes investigaciones se describen detalladamente en los capítulos correspondientes. Para facilitar su consulta y garantizar la transparencia metodológica, se incluyen en la sección de anexos, concretamente en el Anexo 2, los diferentes instrumentos utilizados. A modo resumen, indicar:

- El primer capítulo empírico, el Capítulo 4, aborda el Objetivo Específico 2 (OBE.2) mediante una metodología empírica, donde se aplican los criterios PRISMA, utilizando el cheacklist de PRISMA así como el CASP (Critical

Appraisal Skills Programme)

- Posteriormente, en el Capítulo 5 se presentan dos estudios cualitativos orientados al cumplimiento del Objetivo Específico 4 (OBE.4). Ambos estudios emplearon la técnica de grupos nominales, para recoger información se utilizó cuestionarios de Google y una plantilla de Excel elaborada ad-hoc.
- Finalmente, el Capítulo 6 incluye tres estudios cuantitativos dirigidos a alcanzar el Objetivo 3 (OB3), enfocado en analizar el impacto de diversas variables en la aceptación de herramientas algorítmicas. El primer estudio se basa en casos experimentales ad-hoc, mientras que los estudios segundo y tercero recurren a encuestas diseñadas ad-hoc para recoger datos estructurados de una muestra representativa de participantes. Ambos instrumentos se recogen en el Anexo 2.

3.4. Análisis de datos

Respecto al análisis de datos, y con la finalidad de comprender de manera general los diferentes análisis desarrollados en la tesis, se han realizado los siguientes análisis:

En cuanto a los resultados del capítulo 4, se empleó VOSviewer, un software especializado en la construcción y visualización de redes bibliométricas. Esta herramienta permitió realizar un análisis de coocurrencias y de relaciones entre términos clave extraídos de la literatura científica, lo que facilitó la identificación de las principales líneas temáticas.

Con el fin de sistematizar el tratamiento de la información obtenida en los Grupos Nominales realizados en el capítulo 5, se desarrolló un archivo Excel creado ad hoc, que permite registrar las puntuaciones asignadas por cada participante, calcular automáticamente los totales y porcentajes relativos, y generar gráficos que facilitan la visualización y priorización de las propuestas.

Para el análisis de los datos cuantitativos recogidos en el Capítulo VI, se utilizaron dos herramientas estadísticas complementarias: SPSS y RStudio. SPSS (Statistical Package for the Social Sciences) es un software ampliamente utilizado en el ámbito

de las ciencias sociales, que permite realizar análisis estadísticos descriptivos e inferenciales a través de una interfaz gráfica intuitiva, facilitando la gestión y exploración de bases de datos. En este estudio, se empleó principalmente para el cálculo de frecuencias, medidas de tendencia central, análisis bivariados y contrastes de hipótesis. Por su parte, RStudio es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, orientado al análisis estadístico avanzado y la visualización de datos. Su estructura basada en código ofrece una mayor flexibilidad para la construcción y evaluación de modelos estadísticos complejos. En esta investigación, RStudio se utilizó específicamente para la realización del análisis de regresión logística, dada su capacidad para ajustar modelos multivariantes y personalizar los criterios de evaluación del ajuste. La combinación de ambas herramientas permitió un abordaje riguroso y complementario del análisis estadístico, en coherencia con los objetivos planteados.

Finalmente, con el propósito de ofrecer una síntesis clara y visual, se presenta una tabla (Tabla 1) que resume las diferentes metodologías empleadas, estableciendo su relación con los objetivos y capítulos de la presente tesis doctoral. Así mismo, La elección de los diferentes análisis estadísticos se realizó teniendo en cuenta los criterios establecidos en el árbol de decisión desarrollado por Gerwien (2014) expuesto en el Anexo 4.

Tabla 1.

Resumen de las metodologías empleadas en la tesis.

Capitulo	Estudio	Obj.	Metodología	Análisis de datos	Herramientas
IV	-	OBE.1	Revisión sistemática	• Extracción de datos estructurada • Análisis de términos clave y relaciones • Análisis temporal de evolución temática • Evaluación de calidad de los estudios	VOSviewer CASP
V	Estudio 1	OBE.3	Grupos nominales		
	Estudio 2	OBE.3	Grupos nominales	• Recogida de datos. • Análisis descriptivo (frecuencias, media, etc).	Google Forms Excel
VI	Estudio 1	OBE.2	Encuesta ad-hoc	• Análisis descriptivo (frecuencias, media, etc.). • Validación de supuestos (normalidad, homogeneidad, independencia). • Pruebas no paramétricas. • Modelo de regresión ordinal (exploratorio, no implementado).	SPSS v.29
	Estudio 2	OBE.2	Encuesta ad-hoc	• Análisis descriptivo (frecuencias, media, etc). • Validación de supuestos (normalidad, homogeneidad, independencia). • P. no paramétricas. • Construcción de índices compuestos + fiabilidad (alfa de Cronbach). • Correlaciones entre creencias y aceptación. • Regresiones logísticas binarias. • Árboles de decisión (CHAID)	SPSS v.29 RStudio
	Estudio 3	OBE.2	Casos experimentales ad-hoc	• Análisis descriptivo (frecuencias, media, etc.). • Validación de supuestos (normalidad, homogeneidad, independencia). • Pruebas no paramétricas. Kruskal-Wallis. • Pruebas post-hoc: Dunn con corrección de Bonferroni • Tamaño del efecto	SPSS v.29

CAPÍTULO 4. SESGOS EN LA JUSTICIA HUMANA Y ALGORÍTMICA: UNA REVISIÓN SISTEMÁTICA.

1. Justificación.

Como ya se ha podido evidenciar en los apartados anteriores, la progresiva incorporación de sistemas basados en inteligencia artificial en el ámbito judicial ha reconfigurado el debate en torno a la imparcialidad, la equidad y la legitimidad de las decisiones penales. Aunque la IA se presenta como una herramienta capaz de mejorar la consistencia y reducir los sesgos propios del juicio humano, su uso ha generado importantes controversias desde los planos académico, ético y político (Malek, 2022; Saavedra-Vera, Jáuregui-Bustamante & Arista-Bustamante, 2023). En concreto, la automatización de decisiones judiciales en procesos como la evaluación del riesgo, la determinación de penas o la concesión de libertad condicional, ha reactivado antiguas preocupaciones sobre los mecanismos de discriminación y control en el sistema penal (Martin, 2019; Wisser, 2019; Hueso & Reilly, 2022).

La finalidad de este estudio es analizar, de forma sistemática y crítica, la literatura científica existente sobre los sesgos en la toma de decisiones judiciales, tanto aquellos que derivan del juicio humano como los que pueden surgir del uso de herramientas algorítmicas, con el objetivo de ofrecer una visión actualizada, transversal y rigurosa del fenómeno. Esta necesidad surge no solo por la creciente implementación de la IA en el ámbito penal, sino también por la falta de trabajos que integren, desde una perspectiva unificada, los distintos enfoques disciplinares, metodológicos y teóricos que han abordado la cuestión. En este sentido, el presente trabajo busca llenar un vacío relevante en la literatura: la ausencia de una revisión sistemática que articule el conocimiento previo sobre los sesgos judiciales con las nuevas dinámicas emergentes en la era digital.

En el debate actual sobre la IA en justicia penal predominan dos posturas opuestas. Por un lado, las perspectivas tecno-utópicas plantean que los algoritmos pueden contribuir a eliminar prejuicios humanos gracias a su capacidad de procesamiento de grandes volúmenes de datos y su aparente neutralidad estadística (Hayward &

Maas, 2020; Farfán Intriago et al., 2023; Segura, 2023). Por otro, las visiones tecno-distópicas alertan sobre los riesgos de replicar o intensificar desigualdades estructurales, al operar sobre datos históricos sesgados o emplear variables que inducen a discriminación indirecta (Angwin et al., 2016; O'Neil, 2016; Eubanks, 2018; Lepri et al., 2018). Estas posturas polarizadas reflejan no solo un conflicto de expectativas tecnológicas, sino una disputa sobre el significado mismo de justicia en contextos automatizados.

Además, es importante destacar que la preocupación por los sesgos judiciales no es exclusiva de la inteligencia artificial. La literatura previa, especialmente desde la psicología cognitiva y social, ha demostrado ampliamente la existencia de sesgos sistemáticos en la toma de decisiones judiciales humanas ,como el sesgo de anclaje, disponibilidad, retrospectivo o de representatividad,, así como la influencia de factores emocionales, de género, raciales y socioeconómicos en la evaluación de pruebas y en la severidad de las sentencias (Tversky & Kahneman, 1974; Steffensmeier & Demuth, 2000; Maroney, 2011; Forza, Menegon & Rumiati, 2024). La digitalización del sistema judicial no elimina estas influencias; por el contrario, las transforma y reconfigura, a menudo en formas menos visibles y más difíciles de controlar (Barona, 2019b; Ferrara, 2024).

Así, este estudio responde a una necesidad científica y social apremiante: recopilar, sintetizar y analizar de manera estructurada el conocimiento existente sobre los sesgos en la justicia penal, con el fin de comprender cómo estas dinámicas se ven afectadas, mitigadas o amplificadas, por el uso de tecnologías predictivas. A través de una revisión sistemática de la literatura, se busca no solo clarificar el estado actual de la cuestión, sino también generar una base sólida para futuras investigaciones empíricas, evaluaciones de impacto y propuestas regulatorias que permitan orientar un uso ético, transparente y garantista de la inteligencia artificial en contextos judiciales (Cerezo-Martínez et al. 2024; Slobogin, 2021; Castro-Toledo, 2022).

En definitiva, esta investigación se justifica por su capacidad para aportar una visión holística y crítica sobre uno de los desafíos más relevantes de la justicia contemporánea: la coexistencia, y posible tensión, entre el juicio humano, con sus

limitaciones cognitivas, y el juicio algorítmico, con sus riesgos de reproducción automatizada de desigualdades. Entender esta interacción es esencial para avanzar hacia un sistema penal más justo, responsable y adaptado a las exigencias éticas de la era digital.

2. Objetivos.

El propósito de la presente investigación es analizar la presencia de diversos sesgos en los procesos de toma de decisiones judiciales y su repercusión en la imparcialidad del sistema de justicia. Posteriormente, se observó que un número significativo de fuentes consultadas se centraban en la digitalización. Por consiguiente, se tomó en consideración tanto el juicio humano como el empleo de herramientas algorítmicas en el contexto de la digitalización de la justicia. En consecuencia, se plantean los siguientes objetivos específicos:

OE1. Identificar y categorizar los sesgos que influyen en la toma de decisiones judiciales, evaluando su efecto en la objetividad y equidad del sistema judicial.

OE2. Examinar cómo el debate sobre los sesgos en la justicia se ha trasladado al contexto de la digitalización, con especial atención a la incorporación de herramientas algorítmicas y sus implicaciones en la imparcialidad de las decisiones judiciales.

3. Metodología.

3.1. Procedimiento

Las revisiones sistemáticas son un tipo de estudio de investigación que permite recopilar, examinar y sintetizar de manera rigurosa la información existente sobre una pregunta o tema de investigación, siguiendo un protocolo previamente establecido (Moher, 2009). Su propósito principal es ofrecer una perspectiva objetiva y completa de la evidencia científica disponible en relación con un problema específico.

Con el propósito de examinar la influencia de los sesgos en la toma de decisiones judiciales, se ha llevado a cabo una revisión sistemática siguiendo un enfoque

metodológico riguroso. Para garantizar la transparencia y la calidad del proceso, se ha seguido la metodología establecida en el protocolo PRISMA 2020 (Page et al., 2021), asegurando el cumplimiento de los criterios esenciales para la elaboración de revisiones sistemáticas. Además, se ha utilizado la herramienta en línea *PRISMA Checklist* (PRISMA, n.d.)[6] para verificar la exhaustividad y precisión en la selección y análisis de los estudios incluidos en la revisión. Igualmente, se tuvieron en consideración las etapas recomendadas en otras revisiones sistemáticas (por ejemplo, Dehghanniri y Borrion, 2019; Gough, Oliver y Thomas, 2012; Pérez Domínguez, 2024; Pérez Domínguez, Castro-Toledo y Miró-Llinares, 2019; Nicolás-Sánchez y Castro-Toledo, 2024; Wright, Brand, Dunn y Spindler, 2007). Finalmente, para la ejecución de la revisión sistemática, se siguieron cuatro pasos fundamentales: 1) criterios de elegibilidad, 2) identificación de estudios relevantes, 3) selección de los estudios y 4) análisis de datos y presentación de los resultados.

a) Criterios de elegibilidad.

El propósito de este estudio es examinar cómo los sesgos afectan la objetividad del proceso judicial y de qué manera la automatización puede reforzarlos o, por el contrario, ayudar a mitigarlos o minimizarlos. Para ello, se incluirán estudios que aborden distintos aspectos clave relacionados con este fenómeno, siempre que su enfoque principal esté vinculado a la identificación y evaluación de sesgos en la toma de decisiones judiciales. Se excluirán aquellos estudios que no tengan una relación directa con el análisis de los sesgos en el sistema judicial o que se centren exclusivamente en la automatización sin considerar la equidad y la imparcialidad en el ámbito de la justicia. Además, la recopilación de investigaciones a nivel internacional permitirá obtener una perspectiva global sobre la manifestación y abordaje de los sesgos judiciales. Los hallazgos finales incluirán una síntesis detallada y un análisis temático de los estudios seleccionados. Los criterios de

[6] La lista de verificación en formato web puede cumplimentarse a través del siguiente enlace: https://prisma.shinyapps.io/checklist/. Si quiere comprobarse el cheacklist se debe acceder al artículo publicado en el indicio de calidad 1.

inclusión y exclusión adoptados para la selección de las publicaciones identificadas fueron los siguientes:

Criterio 1: las fuentes deben abordar los sesgos en la toma de decisiones judiciales.

Este criterio se ha establecido con el propósito de garantizar que los estudios incluidos en la revisión analicen de manera directa la influencia de los sesgos en la equidad y objetividad del sistema judicial, considerando tanto la intervención humana como el uso de herramientas algorítmicas. Incluir únicamente estudios que aborden esta temática permite centrar el análisis en la imparcialidad judicial y evitar la inclusión de investigaciones que, aunque relacionadas con los sesgos, no se enfoquen en el ámbito judicial.

Criterio 2: las fuentes deben pertenecer a revistas científicas indexadas.

Para asegurar la calidad metodológica y el rigor de los estudios seleccionados, solo se han considerado publicaciones en revistas científicas revisadas por pares. Esto excluye fuentes de menor fiabilidad, como artículos de opinión, informes no revisados o documentos sin base empírica.

Criterio 3: las fuentes deben estar redactados en español o inglés.

Este criterio se estableció debido a la falta de conocimiento de otros idiomas por parte de los investigadores. Además, la exclusión de publicaciones en otros idiomas no tuvo un impacto significativo en la revisión, ya que del total de estudios revisados únicamente 11 fuentes fueron redactadas en idiomas diferente.

Criterio 4: La población del estudio debe estar compuesta por profesionales del ámbito judicial.

Para garantizar que los hallazgos sean directamente aplicables al contexto judicial, se han seleccionado estudios cuya muestra esté formada por profesionales del derecho, como jueces, fiscales, abogados y otros actores del sistema judicial. Se han excluido estudios centrados en la percepción de la población general, a menos que incluyeran una comparación directa con la perspectiva de profesionales del sector.

b) Identificación de estudios relevantes.

Como primer paso, se llevó a cabo una revisión exploratoria de la literatura (*scoping review*)**,** con el fin de examinar la extensión y variedad de estudios previos sobre el tema. Esta etapa resulta interesante realizarla antes de proceder con una revisión sistemática completa, ya que facilita la identificación del alcance de la investigación y ayuda a definir con precisión los objetivos del estudio. Además, permite reconocer los términos clave más empleados en la literatura académica, optimizando así las estrategias de búsqueda en las bases de datos y asegurando una recopilación exhaustiva de información relevante. Gracias a este análisis preliminar, fue posible determinar qué palabras clave eran más recurrentes en los artículos vinculados al objeto de estudio, lo que mejoró la precisión de la búsqueda posterior.

La recopilación de artículos se realizó en mayo 2024 mediante dos bases de datos ampliamente reconocidas en la comunidad científica: ProQuest y Web of Science**.** Estas plataformas fueron seleccionadas por su alto nivel de confiabilidad y su extenso alcance en diversas disciplinas académicas.

- **ProQuest** destaca por la diversidad de fuentes que alberga, incluyendo revistas científicas, tesis y documentos especializados en áreas como el derecho, la criminología, o la psicología¡. Entre sus bases de datos más relevantes se encuentran *ProQuest Criminal* que ofrecen acceso a investigaciones especializadas en estos campos.
- **Web of Science** proporciona acceso a más de 182 millones de registros en diversas áreas del conocimiento, como biomedicina, ciencias sociales, ingeniería y humanidades. Su infraestructura incluye colecciones de alto impacto, tales como *Web of Science Core Collection*, *BIOSIS*, *SciELO* y *Data Citation Index*, lo que permite una cobertura interdisciplinaria amplia y relevante.

Estrategia de Búsqueda

Para garantizar la recuperación de artículos pertinentes, se emplearon términos y operadores de búsqueda específicos en cada una de las bases de datos seleccionadas. Dado que cada base de datos utiliza criterios diferentes de búsqueda

fue necesario adaptarla a las particularidades de cada uno de ellos, resultado finalmente la siguiente búsqueda:

- **ProQuest:** NOFT(judicial decision-making) AND NOFT(bias).
- **Web of Science:** "judicial decision-making" (Topic) AND bias (Topic).

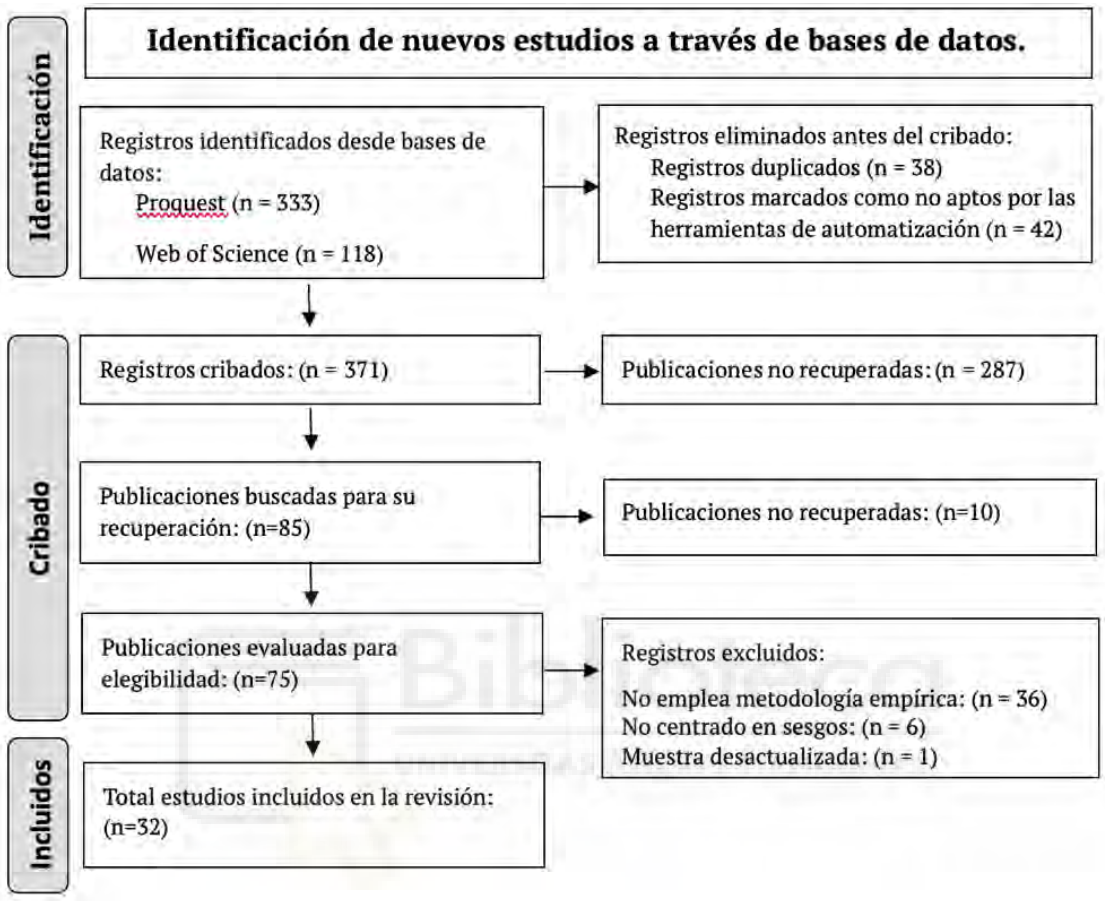
Esta estrategia permitió maximizar la recuperación de estudios relevantes y garantizar que la revisión sistemática se realizara sobre una base de datos sólida y representativa. En ese sentido, y aunque en un primer momento se consideró incluir la variable de digitalización, esta reducía considerablemente el número de trabajos a evaluar y no permitía analizar el traslado del debate, por lo que se optó por una visión más amplia del fenómeno.

c) Selección de los estudios.

Una vez realizada la búsqueda en las bases de datos y aplicados los criterios de inclusión y exclusión automatizados (criterios 1 y 2), así como eliminados los artículos duplicados mediante el gestor bibliográfico Zotero, se recopilaron 371 artículos potenciales. A continuación, se procedió a la lectura de títulos y resúmenes (abstracts) para determinar su alineación con los objetivos de la revisión, lo que permitió descartar 198 estudios que no los abordaban adecuadamente. Posteriormente, se realizó una lectura detallada de los artículos restantes, lo que dio lugar a una selección final de 32 trabajos que cumplían con los criterios de elegibilidad establecidos. La Figura 2 muestra de forma esquemática el procedimiento seguido para la búsqueda y selección de las fuentes de datos.

Figura 2.

Diagrama de la selección de los artículos.



d) Análisis de datos y presentación de los resultados.

La extracción de datos se organizó en una tabla resumen (Tabla 3), estructurada según las variables necesarias para el cumplimiento de los objetivos, con la finalidad de facilitar tanto el análisis cualitativo como cuantitativo. En concreto, se recopilaron las siguientes variables: autor/es, año de publicación, objetivos del estudio, muestra, diseño metodológico, tipo de sesgo y principales resultados.

En relación con la variable "tipo de sesgo", se optó por una doble codificación con el objetivo de garantizar un análisis más detallado y comparativo de los resultados. Por un lado, se estableció una categoría específica que recoge de forma literal el sesgo particular identificado y analizado en cada uno de los estudios incluidos. Esta codificación se realizó respetando la terminología empleada por los propios autores,

con el fin de preservar la fidelidad conceptual y evitar interpretaciones que pudieran alterar el contenido original. Por otro lado, una vez analizadas todas las fuentes, se desarrolló una segunda codificación basada en categorías generales, construidas a partir de un proceso iterativo de revisión interna. Este proceso permitió identificar patrones comunes entre los distintos tipos de sesgo, garantizando la coherencia, eliminando redundancias y maximizando la relevancia práctica de las categorías resultantes. Así, se agruparon los sesgos específicos en categorías temáticas más amplias. Concretamente, la categoría de sesgos demográficos agrupa aquellos estudios en los que se evidencian prejuicios o tratamientos diferenciados en función de características personales como la edad, el género, la etnia o la raza. Asimismo, se identificaron otras categorías relevantes, como los sesgos cognitivos, que se relaciona con heurísticos o atajos mentales según lo define la literatura en esta materia (Tversky y Kahneman, 1974). Estas categorías permiten analizar de manera agregada los efectos que estas variables pueden tener en la toma de decisiones judiciales o en el funcionamiento de herramientas algorítmicas aplicadas al ámbito, facilitando el análisis comparado y la identificación de patrones comunes o divergentes en el abordaje de los sesgos en el ámbito de la justicia y la inteligencia artificial.

Con el objetivo de complementar el análisis de contenido temático realizado sobre los estudios incluidos en esta revisión sistemática (Tabla 2 y Tabla 3), se incorporó un análisis de redes de términos (Figura 3) utilizando el software VOSviewer. Esta herramienta permite visualizar mapas de coocurrencia léxica que revelan relaciones semánticas entre los principales términos del corpus documental, facilitando la identificación de patrones y la estructura conceptuales subyacente en la literatura sobre sesgos en el sistema de justicia penal. El análisis de redes ofrece una representación gráfica clara de la centralidad, frecuencia e interrelación de los conceptos, enriqueciendo la interpretación de los resultados (Van Eck & Waltman, 2010; Arruda, Silva, Lessa, Proença & Bartholo, 2022). El análisis se desarrolló en tres fases:

- Recopilación de los datos de sesgos (generales y específicos), indicando su frecuencia y estableciendo sus relaciones categóricas.

- Generación, mediante VOSviewer, de un mapa de coocurrencia basado en la frecuencia de aparición conjunta de los términos.
- Interpretación de los clústeres generados automáticamente a partir del análisis previo de contenido, lo que permitió vincular grupos de términos con dimensiones temáticas emergentes. En relación con la interpretación: a) el tamaño de los nodos indica la frecuencia de cada término en el corpus, b) la anchura de las líneas representa la intensidad de coocurrencia entre términos y c) los colores agrupan términos asociados, revelando subtemas o áreas conceptuales.

Paralelamente, se realizó un análisis temporal de las variables en función del año de publicación, con el objetivo de explorar la evolución del discurso científico sobre los sesgos, especialmente tras la irrupción de tecnologías basadas en inteligencia artificial. Esta perspectiva longitudinal permite identificar cambios en prioridades investigativas y la aparición de nuevos enfoques. Para ello, se clasificaron los sesgos específicos por año y se incluyeron también 12 estudios teóricos no considerados en el análisis principal, alcanzando un total de 44 artículos. La inclusión de trabajos empíricos y teóricos permite una visión más completa sobre los desplazamientos discursivos vinculados a la transformación tecnológica. La evolución temática se representa gráficamente en la Figura 4.

Finalmente, para asegurar que los estudios incluidos en esta revisión sistemática cumplieran unos estándares mínimos de calidad, se aplicó el método de evaluación de calidad desarrollado por el *Critical Appraisal Skills Program (CASP).* Esta herramienta permite a los evaluadores examinar el grado de rigor metodológico, la fiabilidad y la pertinencia de cada estudio (Long et al., 2020). Los hallazgos obtenidos en la evaluación de las publicaciones seleccionadas se presentan en la Tabla 32 incluida en el Anexo 1. En la presente tabla se representa con un ✓(sí) aquellos estudios que cumplen completamente con el criterio, ± (parcialmente) se cumplen algunos elementos, pero con deficiencias y X (no) cuando el criterio no se cumple. Es importante señalar que todos los estudios fueron incluidos en la revisión. En consecuencia, ningún artículo que había superado la fase de cribado fue descartado en esta etapa del análisis.

4. Resultados.

4.1. Resultados de los artículos incluidos en la revisión.

Los estudios revisados pueden agruparse en cinco categorías principales. La mayoría de los artículos analizados exploran la presencia de sesgos demográficos en la toma de decisiones judiciales, representando el 72% (n = 23) del total. Estos estudios se centran en variables como el género, la raza, la etnicidad, la religión y la ideología influyen en las resoluciones judiciales, observándose que el género (n = 9) y la raza (n = 9) son los factores más estudiados dentro de este grupo. En segundo lugar, el 19% (n = 6) de los estudios abordan sesgos cognitivos, explorando cómo heurísticos y errores sistemáticos en el procesamiento de la información afectan la toma de decisiones en el ámbito judicial. Entre estos sesgos, destaca especialmente el sesgo de anclaje (n = 3), pero también se analizan otros sesgos como el de confirmación, encuadre, representatividad, retrospección y sesgos egocéntricos (n = 1 cada uno). Un tercer grupo, representando el 9% (n = 3), analiza el impacto de la digitalización y la inteligencia artificial en el sistema judicial. Estos estudios examinan cómo la incorporación de herramientas algorítmicas puede mejorar la eficiencia y la transparencia, pero también introducir sesgos y errores en la toma de decisiones. En este sentido, uno de los estudios revisados se centra en el uso del procesamiento de lenguaje natural (NLP) para predecir decisiones judiciales a partir del contenido textual de las sentencias. Resulta relevante destacar, que se encontraron un total de 17 artículos relativos a la digitalización que formaban aparte de los 56 artículos para evaluación completa, sin embargo, la gran mayoría de ellos la abordaban desde un prisma puramente teórico y por ello no han sido incluidos finalmente en la revisión. Por otro lado, el 3% (n = 1) de los artículos se enfoca en el contexto social, explorando cómo el trasfondo socioeconómico de los jueces puede influir en sus decisiones, mientras que otro 3% (n = 1) investiga el uso del lenguaje en la argumentación judicial y su impacto en las resoluciones.

Desde el punto de vista metodológico, la mayoría de los estudios emplea enfoques cuantitativos, basados en bases de datos de decisiones judiciales y modelos estadísticos para identificar patrones de sesgo en la toma de decisiones. También se incluyen enfoques experimentales y de simulación, que evalúan cómo distintos

factores pueden afectar la objetividad de los jueces en entornos controlados. Un porcentaje menor de estudios emplea métodos cualitativos, como análisis de discurso y entrevistas, para examinar la influencia del contexto social y los valores individuales en la toma de decisiones. Los resultados se han dividido en resultado cualitativos (Tabla 2), donde se exponen los datos de cada una de las fuentes analizadas en profundidad y los resultados cuantitativos (Tabla 3).

Tabla 2.

Resultados cualitativos de las 32 fuentes resultantes de la revisión sistemática.

Autores	**Objetivo**	**Muestra**	**Tipo de sesgo**		**IA**	**Resultados**
			General	**Específico**		
Hochstedler Webb, Riley, Wells (2024)	Evaluar disparidades raciales en decisiones de detención preventiva mediante modelos de clasificación errónea.	1,990 personas admitidas en el sistema de justicia preventiva de Virginia	Demográfico/ digitalización	Raza H. algorítmica	Sí	Mayor tasa de detención y errores para acusados negros; el VPRAI y los jueces muestran sesgos raciales.
Chatziathanasiou (2022)	Evaluar la validez del efecto 'juez hambriento' y su uso en la justificación de IA en derecho.	Análisis de estudio previo sobre el efecto 'juez hambriento'.	Cognitivos/ digitalización	Juez hambiento	Sí	El efecto 'juez hambriento' no tiene base sólida para justificar el uso de IA.
Choi, Harris & Shen-Bayh (2022)	Examinar el sesgo étnico en decisiones judiciales en apelaciones criminales en Kenia.	10,000 apelaciones criminales en tribunales de Kenia.	Demográfico	Etnia.	/	Los jueces favorecen a apelantes coétnicos en un 3-5% más.
Concannon & Na (2022)	Analizar la disparidad racial y étnica en las solicitudes de fianza de los fiscales y sus efectos posteriores.	43,971 denuncias por delitos graves en Nueva York.	Demográfico	Raza.	/	Los acusados negros reciben fianzas más altas y mayor tasa de acusaciones.
Romain Dagenhardt (2021)	Examinar la disparidad racial en audiencias de revisión de libertad condicional mediante un enfoque mixto.	350 casos de audiencias de revisión de libertad condicional en una corte de violencia doméstica en EE.UU.	Demográfico	Raza.	/	El género, raza y estado familiar influyen en las decisiones de sanción. Los jueces utilizan discursos racializados al evaluar la responsabilidad y el uso de drogas.
Kastellec (2020)	Examinar los efectos de paneles raciales en casos de pena de muerte en apelaciones.	Datos sobre paneles en tribunales de apelación de EE.UU.	Demográfico	Raza	/	Los paneles con jueces negros son más laxos en casos de pena de muerte.
Harrin & Sen (2019)	Revisar la literatura sobre la influencia de raza, género e ideología en la toma de decisiones judiciales.	Revisión de estudios empíricos sobre decisiones judiciales.	Demográfico	Raza Género Ideología	/	La ideología es el mayor predictor de decisiones judiciales.

Autores	Objetivo	Muestra	Tipo de sesgo		IA	Resultados
			General	Específico		
Ecker, Ennser-Jedenastik & Haselmayer (2019)	Analizar la existencia de sesgo de género en adjudicaciones de asilo en Austria.	40,000 decisiones de asilo en Austria.	Demográfico	Género.	/	Los jueces muestran mayor indulgencia con mujeres en asilos si manejan más casos masculinos.
Oren-Kolbinger (2019)	Medir el efecto del trasfondo social de los jueces en decisiones sobre impuestos.	Casos de impuestos en Israel.	Social	Efecto del trasfondo social en decisiones fiscales.	/	El trasfondo social afecta las decisiones fiscales de los jueces.
Reyes & Reyes (2019)	Analizar empíricamente el impacto de la religión en la toma de decisiones judiciales.	Base de datos de decisiones judiciales y características de jueces.	Demográfico	Religión.	/	La religión influye en ciertas decisiones, especialmente en casos de derechos religiosos.
McKay (2019)	Explorar el uso de herramientas actuariales, algoritmos e IA en la toma de decisiones judiciales.	Análisis de algoritmos utilizados en evaluación de riesgos en tribunales.	Digitalización	Efecto de algoritmos e IA en decisiones judiciales.	Sí	Las herramientas basadas en IA pueden introducir sesgos en las decisiones judiciales.
Gill, Kagan & Marouf (2019)	Estudiar la influencia del género en decisiones de apelaciones de inmigración.	Datos de apelaciones de inmigración en tribunales federales de EE.UU.	Demográfico	Género.	/	Los jueces masculinos son más severos con hombres inmigrantes y más indulgentes con mujeres.
Miller (2018)	Investigar si la experiencia mitiga los sesgos de género en decisiones judiciales.	Jueces y ciudadanos en estudios experimentales.	Demográfico	Género.	/	La experiencia no reduce los sesgos de género en jueces.
Lowder, Morrison, Kroner & Desmarais (2018)	Investigar el sesgo racial en evaluaciones de riesgo en sentencias de libertad condicional.	11,792 casos de libertad condicional en EE.UU.	Demográfico	Raza.	/	Los evaluadores blancos aplican estándares más estrictos a los acusados negros.
Ho, Shlosberg & Lesneskie (2018)	Analizar cómo los errores humanos afectan la validez de la clasificación de riesgo en instrumentos de evaluación delictiva.	1.000 (Risk Device X) + 1.000 (Oregon JCP, juveniles)	Cognitivos	Errores humanos al codificar información	/	Un 10%-20% de error humano en los datos provoca que un 24%-48% de las personas se clasificaran en un nivel de riesgo incorrecto.
Wofford (2017)	Analizar cómo el género influye en las decisiones de los litigantes en el sistema judicial.	Encuesta en línea sobre decisiones de litigantes.	Demográfico	Género.	/	No hay un efecto claro del género en la decisión de litigar.
Gravett (2017)	Investigar cómo el sesgo racial implícito afecta la trayectoria del juicio penal.	Revisión de evidencia empírica sobre sesgo racial implícito.	Demográfico	Raza.	/	El sesgo racial implícito influye en la toma de decisiones judiciales.
Aletras, Tsarapatsanis, Daniel Preoţiuc-Pietro,	Predecir decisiones judiciales del Tribunal Europeo de	Decisiones del Tribunal Europeo de	Lenguaje	Contenido textual.	/	Las decisiones pueden predecirse con un 79% de

Autores	Objetivo	Muestra	Tipo de sesgo		IA	Resultados
			General	Específico		
Vasileios Lampos (2016)	Derechos Humanos mediante NLP.	Derechos Humanos.				precisión basándose en texto.
Recupero, Christopher, Strong, Price & Harms (2015)	Evaluar la influencia del sesgo de género en desafíos testamentarios por influencia indebida.	Casos de impugnaciones testamentarias en EE.UU.	Demográfico	Género	/	Los jueces consideran con mayor frecuencia la influencia indebida en testamentos de mujeres.
Spohn & Sample (2013)	Examinar la relación entre estereotipos raciales y las decisiones de sentencia en casos de drogas.	Sentencias de delitos de drogas en tribunales federales de EE.UU.	Demográfico	Raza.	/	El sesgo racial es más pronunciado en condenas de drogas para afroamericanos.
Williams & Law (2012)	Examinar la influencia de la ideología en las decisiones judiciales en casos de inmigración.	Casos de apelación en inmigración en EE.UU.	Demográfico	Ideología.	/	La ideología judicial tiene influencia, pero es atenuada por factores institucionales.
Bond &Jeffries (2011)	Examinar el impacto de la indigeneidad en la decisión de encarcelamiento en tribunales de Australia Occidental.	Registros de sentencias en tribunales de Australia Occidental (2003-2005).	Demográfico	Indigeneidad	/	Los hombres indígenas tienen mayores tasas de encarcelamiento.
Eerland & Rassin (2010)	Explorar los efectos del sesgo de confirmación y el 'feature positive effect' en la evaluación de evidencia legal.	Estudio con estudiantes de derecho.	Cognitivos	Sesgo de confirmación 'feature positive effect'.	/	El sesgo de confirmación y el FPE influyen en la evaluación de la evidencia,
Elvin (2010)	Investigar el uso de estereotipos sexuales en decisiones judiciales en Inglaterra y Gales.	Casos de apelaciones judiciales en Inglaterra y Gales.	Demográfico	Género.	/	A pesar del entrenamiento, los jueces siguen usando estereotipos sexuales en fallos.
Feldman (2006)	Explorar la influencia de la religión en la toma de decisiones judiciales.	Revisión de literatura y análisis doctrinal.	Demográfico	Religión	/	Las creencias religiosas pueden influir en decisiones judiciales.
Englich, Mussweiler & Strack (2006)	Determinar la influencia de anclajes irrelevantes en decisiones judiciales.	Estudio experimental con jueces alemanes.	Cognitivos	Sesgo de anclaje	/	Incluso jueces expertos son influenciados por anclajes irrelevantes.
Peresie (2005)	Determinar el impacto del género en la toma de decisiones colegiada en tribunales federales de apelación.	556 casos de apelaciones federales en EE.UU.	Demográfico	Género.	/	Las juezas aumentan la probabilidad de fallos a favor de demandantes en casos de género.
Sisk & Heise (2005)	Debatir el impacto de la ideología en la toma de decisiones judiciales utilizando medidas estadísticas.	Análisis de decisiones judiciales mediante estadística aplicada.	Demográfico	Ideología	/	Las decisiones judiciales están influenciadas por la ideología, pero con matices estadísticos.

Autores	Objetivo	Muestra	Tipo de sesgo		IA	Resultados
			General	Específico		
Manning, Carroll & Carp (2004)	Analizar la influencia de la edad en la toma de decisiones en casos de discriminación por edad.	544 casos de discriminación por edad en tribunales federales.	Demográfico	Edad.	/	Jueces mayores son más favorables a demandantes en casos de edad.
Fariña, Arce & Novo (2003)	Aislar el heurístico de anclaje y evaluar su impacto en la toma de decisiones judiciales.	Sentencias judiciales en España.	Cognitivos	Sesgo de anclaje.	/	El 63.6% de las sentencias estaban influenciadas por el anclaje.
Guthrie, Rachlinski& Wistrich (2002)	Explorar el uso de heurísticos y sesgos cognitivos en la toma de decisiones judiciales.	Experimentos con jueces estadounidenses.	Cognitivos	Anclaje. Encuadre Retrospección Representatividad Egocéntricos	/	Los jueces usan heurísticos en su toma de decisiones, lo que introduce sesgos.
Staffensmeier, Ulmer & Kramer (1998)	Analizar la interacción de raza, género y edad en la severidad de sentencias penales.	Registros de sentencias en Pensilvania (1989-1992).	Demográfico	Raza Género Edad	/	Los hombres jóvenes y negros reciben las sentencias más severas en comparación con otros grupos.

Tabla 3.

Resumen de los resultados cuantitativos de la revisión sistemática de la literatura.

Categoría	**n**	**%[a]**	**Subcategoría**	**n**	**%**
Demográficos	23	72	Raza	9	28.1
			Edad	2	6.25
			Sexo/género	9	28.1
			Religión	1	3.2
			Ideología	3	9.4
			Indigeneidad	1	3.2
			Etnia	2	6.25
Cognitivos	6	19	Sesgo de anclaje	3	9.38
			Sesgo de confirmación	1	3.2
			Sesgo de encuadre	1	3.2
			Sesgo de representatividad	1	3.2
			Sesgo de retrospección	1	3.2
			Sesgos egocéntricos	1	3.2
			'feature positive effect'.	1	3.2
			Errores humanos	1	3.2
			Juez hambriento	1	3.2
Digitalización	3	9	Herramientas algorítmicas	2	6.25
			Inteligencia artificial	2	6.25
Social	1	3	-	1	3.2
Lenguaje	1	3	-	1	3.2

[a] Los porcentajes se han calculado en relación con el número total de estudios revisados (N=32). Los porcentajes no suman exactamente 100% porque algunos elementos fueron clasificados en más de una categoría.

En este sentido, se identificó un bloque central que agrupa las principales categorías de sesgo, destacándose los sesgos demográficos, cognitivos, sociales, de digitalización y de lenguaje. Dentro de cada categoría, se encuentran diversas manifestaciones concretas que evidencian cómo estos sesgos pueden afectar la objetividad y equidad en la toma de decisiones. El sesgo demográfico es una de las categorías más recurrentemente aparecía dentro del análisis, al estar compuesto por términos como "raza", "género", "religión", "ideología" y "edad". Estos términos reflejan cómo los factores identitarios pueden influir en la evaluación de las personas dentro del sistema judicial. En particular, se observa una alta frecuencia de sesgos relacionados con raza y género, lo que evidencia la presencia de prejuicios

estructurales en los procesos de toma de decisiones. Por otro lado, la comunidad vinculada a los sesgos cognitivos agrupa términos como "sesgo de anclaje", "sesgo de confirmación", "errores humanos" y "evaluación", los cuales indican que los prejuicios individuales y las limitaciones en la percepción pueden condicionar la interpretación de la información y la toma de decisiones judiciales. La presencia destacada de estos términos sugiere que las decisiones pueden estar influenciadas por información previamente adquirida, generando un efecto de refuerzo en determinadas tendencias de juicio. En cuanto a la digitalización, se identificó una comunidad de términos que resalta el impacto de la inteligencia artificial y los algoritmos en la toma de decisiones judiciales. Palabras como "efecto de los algoritmos", "IA", y "automatización" reflejan la preocupación sobre la posibilidad de que estas tecnologías reproduzcan sesgos existentes en los datos utilizados para su entrenamiento. La interrelación de estos términos sugiere que la digitalización de los procesos judiciales plantea nuevos desafíos en términos de equidad y transparencia. El sesgo social, por su parte, aparece vinculado al trasfondo socioeconómico de los individuos y su influencia en la toma de decisiones fiscales y judiciales. En esta comunidad, términos como "contexto social", "normativa" y "efectos estructurales" sugieren que la justicia penal puede no ser aplicada de manera homogénea, sino que varía en función de factores externos que afectan la percepción y el tratamiento de los casos. Finalmente, el análisis de los sesgos lingüísticos muestra la importancia del contenido textual en la generación y transmisión de prejuicios dentro del ámbito judicial. La manera en que se redactan documentos, sentencias y normativas puede influir en la interpretación y aplicación de la justicia, reforzando ciertos estereotipos o predisposiciones en el lenguaje jurídico.

4.2. Análisis de redes de términos.

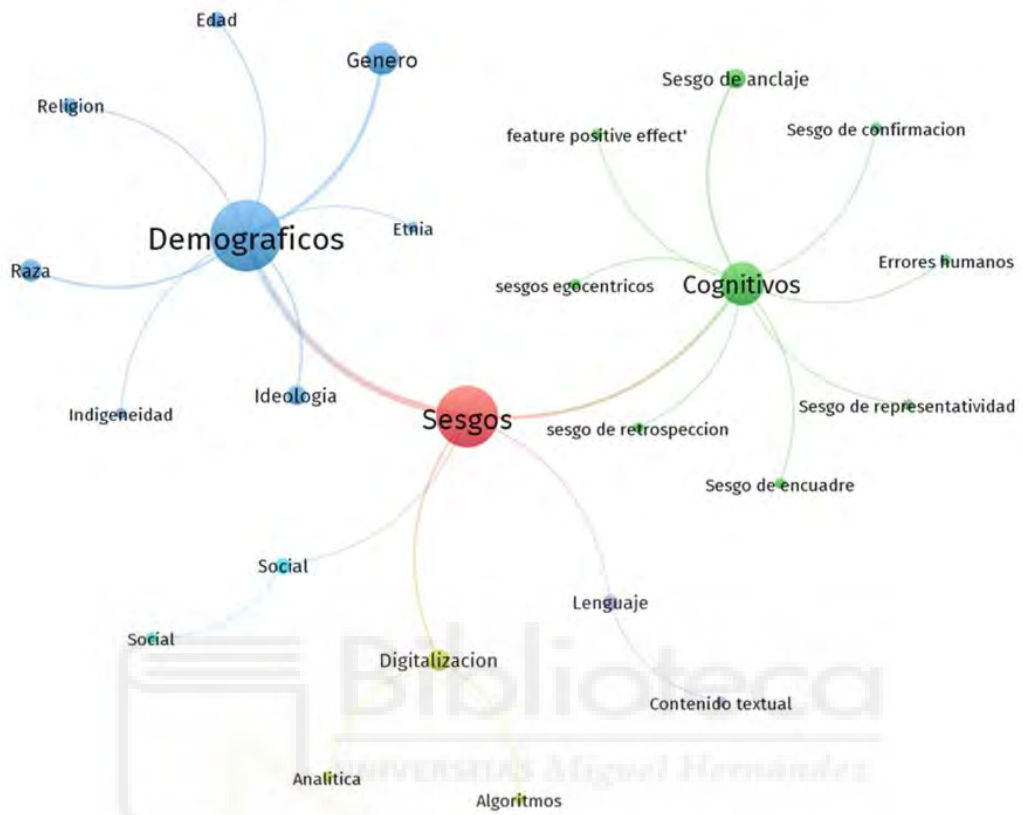
El análisis de redes de términos permite visualizar la estructura conceptual de un campo de estudio y sus interrelaciones. El mapa de términos revela la existencia de cinco clústeres, cada uno representado por un color distinto. En el centro del mapa se encuentra el nodo "Sesgos", que actúa como punto de conexión entre los diferentes grupos. Este nodo se relaciona con cada una de las categorías generales

identificadas en el análisis de contenido, evidenciando la naturaleza multifacética de los sesgos y su presencia en distintos ámbitos.

El primer clúster, representado en azul, agrupa los términos relacionados con factores demográficos, tales como género, edad, raza, religión e indigeneidad. El segundo clúster, en verde, incluye diversos sesgos cognitivos, como el sesgo de anclaje, sesgo de confirmación, sesgo de encuadre y sesgo de retrospectiva. El tercer clúster, en amarillo, agrupa términos vinculados a la digitalización, como algoritmos y analítica de textos. Finalmente, con menor representatividad, el cuarto clúster, en color celeste, está compuesto por términos asociados a lo social; y el quinto clúster, en color morado, agrupa términos relacionados con el lenguaje, como es el contenido del texto que se les explica a los operadores jurídicos. Este tipo de análisis de redes de términos permite poner de manifiesto la complejidad y transversalidad de los sesgos, los cuales pueden clasificarse en diferentes categorías. La interconexión de estos conceptos resalta la necesidad de un enfoque interdisciplinario que permita comprender mejor la influencia de los sesgos en la toma de decisiones y en el desarrollo de herramientas tecnológicas, con el fin de mitigar su impacto en distintos ámbitos de aplicación. Esto se puede apreciar en la Figura 3:

Figura 3.

Distribución y relación de los diferentes sesgos incluidos en los estudios de la revisión.



4.3. Evolución de los factores analizados.

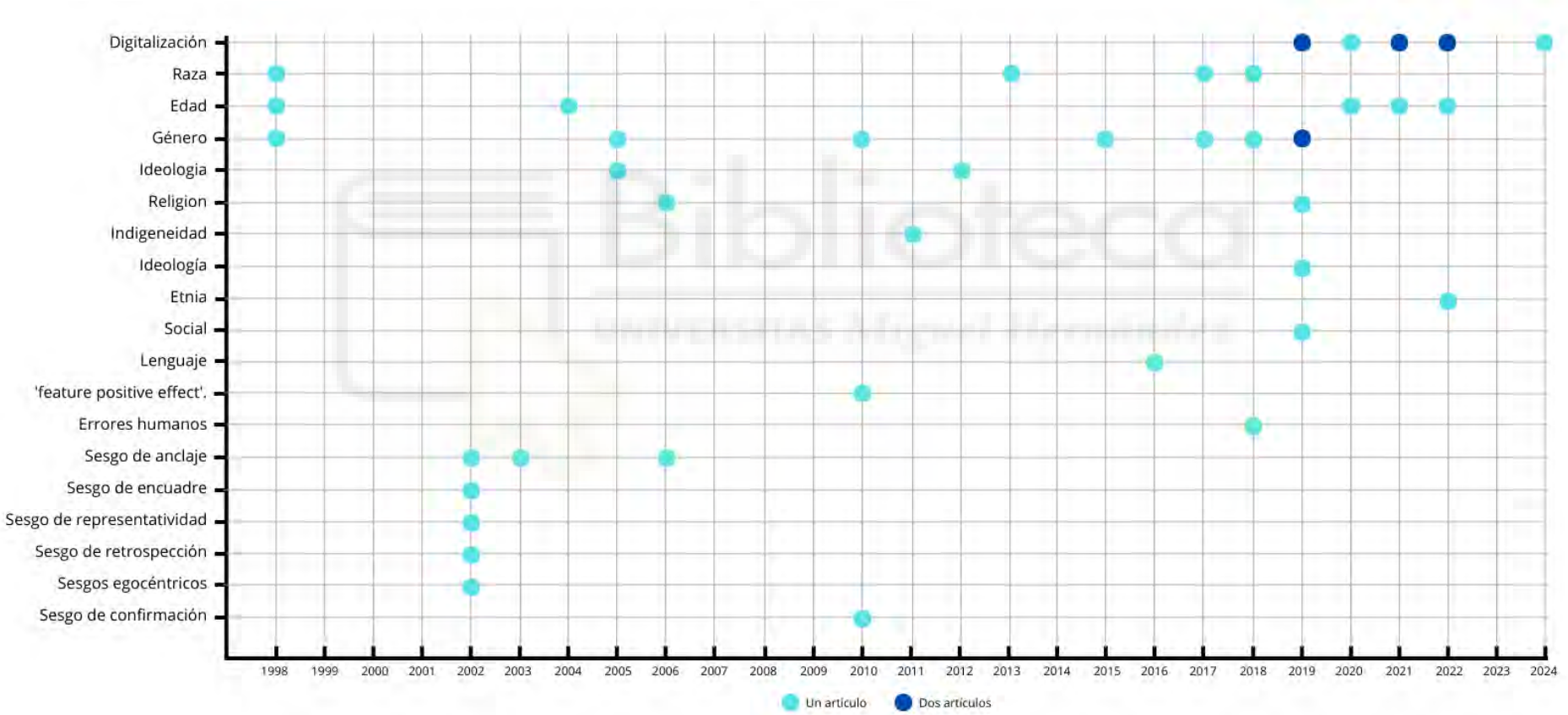
Con la finalidad de valorar la evolución en la discusión académica en materia de sesgos con la aparición de la IA, se examinaron también las variables de estudio en función del año de publicación de los artículos. Para ello, se decidió incluir también los estudios con que abordan la problemática desde una perspectiva teórica, puesto que se obtiene una visión más amplia.

Se observó que, en torno a la década de 1990 y principios del 2000, las primeras investigaciones analizadas se centran mayormente en variables individuales como la raza, la edad y el género, apareciendo estas en publicaciones desde 1998. A partir del año 2002, comienza a evidenciarse un mayor interés en sesgos cognitivos específicos, tales como el sesgo de anclaje y el sesgo de encuadre. A medida que avanzaron los años, la variedad de temas analizados aumentó. En la década de 2010,

el foco se amplía hacia variables relacionadas con el sesgo de confirmación (2010), el sesgo implícito (2015), y progresivamente se integran preocupaciones vinculadas a la transparencia algorítmica y la interpretabilidad en el contexto de la toma de decisiones mediada por inteligencia artificial. En los años más recientes, especialmente desde 2019, los estudios comenzaron a incorporar el análisis de sesgos en los sistemas de IA, explorando su impacto en la equidad y la discriminación algorítmica, así como la implementación de métodos de mitigación de sesgos y la evaluación ética de estas tecnologías (Figura 4).

Figura 4.

Distribución temporal de los sesgos identificados en los artículos.



5. Discusión y conclusiones.

La presente revisión sistemática pone de relieve que la toma de decisiones judiciales no es un proceso plenamente objetivo y que los operadores jurídicos continúan viéndose influidos por una diversidad de sesgos. Los estudios empíricos incluidos en esta revisión muestran que la investigación sobre los sesgos en la justicia sigue centrada mayoritariamente en los operadores humanos, como jueces, fiscales y otros profesionales del sistema judicial. En particular, se ha identificado una persistencia de sesgos demográficos, especialmente aquellos relacionados con la raza, el género y el estatus socioeconómico (Steffensmeier & Demuth, 2000; Lowder et al., 2018). Asimismo, se han constatado sesgos cognitivos, como el sesgo de anclaje, el sesgo de confirmación y el sesgo retrospectivo (Englich, Mussweiler & Strack, 2006; Guthrie, Rachlinski & Wistrich, 2001).

El análisis de los 32 estudios empíricos incluidos en la revisión sistemática proporciona evidencia consistente sobre la existencia de sesgos en la toma de decisiones judiciales. Los datos recogidos muestran que los jueces no son inmunes a los sesgos cognitivos ni a los prejuicios demográficos o sociales, que influyen de manera significativa en sus resoluciones.

En primer lugar, los sesgos demográficos emergen como los más prevalentes en los estudios analizados, representando el 72% de las investigaciones (n = 23). Dentro de esta categoría, destacan los sesgos de género (n = 9) y raciales o étnicos (n = 6). Por ejemplo, estudios como el de Steffensmeier, Ulmer y Kramer (1998) y Lowder et al. (2018) documentan cómo los acusados pertenecientes a minorías raciales, en especial, personas negras o indígenas, reciben penas más severas o tienen mayores probabilidades de detención preventiva que los acusados blancos. De manera similar, Ecker, Ennser-Jedenastik y Haselmayer (2019) evidencian un trato más indulgente hacia mujeres solicitantes de asilo, mientras que Peresie (2005) muestra que la presencia de juezas en tribunales colegiados incrementa la probabilidad de fallos favorables a demandantes en casos de discriminación de género. También se ha podido detectar la persistencia de estos sesgos incluso en contextos de formación y experiencia judicial. Por ejemplo, Miller (2018) señala que la experiencia judicial no mitiga los sesgos de género en la toma de decisiones, mientras que Gravett

(2017) demuestra la influencia del sesgo racial implícito en el desarrollo de procesos penales, a pesar de las políticas de formación en igualdad y diversidad.

En segundo lugar, los sesgos cognitivos también tienen una presencia destacada, representando el 18,8% de los estudios (n = 6). Se identifican sesgos de anclaje, como los descritos por Englich, Mussweiler y Strack (2006), quienes demostraron que incluso jueces experimentados se ven influidos por anclajes irrelevantes, lo que afecta la severidad de sus sentencias. Otros sesgos, como el de confirmación (Eerland & Rassin, 2010), el de encuadre y el retrospectivo, revelan cómo las limitaciones cognitivas de los jueces afectan su interpretación de la evidencia y la valoración de las pruebas presentadas (Guthrie, Rachlinski & Wistrich, 2001).

Además, se ha documentado cómo factores contextuales y trasfondos socioeconómicos de los jueces influyen en sus decisiones. Oren-Kolbinger (2019) muestra que el origen social del juez tiene un impacto significativo en los casos fiscales, condicionando sus apreciaciones sobre equidad tributaria y responsabilidad individual.

En conjunto, los datos analizados sugieren que los jueces están sesgados, tanto por sus propios condicionantes cognitivos y sociales como por factores estructurales que afectan al sistema judicial en su conjunto. La diversidad en la composición de los tribunales y la formación en sesgos cognitivos parecen tener un papel mitigador limitado, lo que refuerza la necesidad de adoptar mecanismos de supervisión externos y auditorías en el proceso de toma de decisiones judiciales.

Un hallazgo relevante en la literatura revisada es que los sesgos tradicionalmente atribuidos a la psicología judicial humana no desaparecen completamente con la digitalización del proceso judicial, sino que en ocasiones pueden trasladarse a través del diseño, entrenamiento y uso de sistemas algorítmicos en la toma de decisiones. Diversos estudios han analizado cómo factores personales y emocionales inciden en las resoluciones judiciales, destacando el trabajo de Chatziathanasiou (2022), quien revisa críticamente el denominado efecto “juez hambriento”. Este autor cuestiona la validez empírica del estudio clásico de Danziger et al. (2011) y concluye que no existe evidencia sólida que respalde la hipótesis de que la fatiga o el hambre influyan

de manera significativa en el endurecimiento de las decisiones judiciales antes de los descansos. Además, advierte sobre el riesgo de utilizar estos argumentos para legitimar la automatización de decisiones judiciales mediante sistemas de inteligencia artificial.

A pesar de las promesas de objetividad asociadas al uso de herramientas algorítmicas, los estudios empíricos que abordan su implementación en el ámbito judicial revelan la persistencia de sesgos, aunque en menor medida en comparación con el factor humano. El estudio de Hochstedler Webb, Riley y Wells (2024), por ejemplo, identifica disparidades significativas en la aplicación del algoritmo VPRAI, que recomienda la detención preventiva con mayor frecuencia a acusados negros que a blancos en situaciones similares. Además, los jueces tienden a agravar esta desigualdad al optar por la detención de personas negras incluso cuando el algoritmo sugiere su liberación. Según los datos del estudio, la tasa de detención errónea asciende al 51,4% en acusados negros frente al 39,7% en blancos, lo que demuestra que tanto las decisiones humanas como las algorítmicas pueden perpetuar y amplificar las desigualdades raciales en el sistema de justicia penal. Por su parte, McKay (2019) advierte sobre los riesgos inherentes al uso de tecnologías predictivas en el proceso penal, subrayando la opacidad de los algoritmos propietarios, la erosión de las garantías del debido proceso y la reproducción de sesgos estructurales ya presentes en los sistemas judiciales. Un caso paradigmático es State v. Loomis, donde el sistema COMPAS influyó en la condena del acusado sin que este pudiera acceder ni impugnar el funcionamiento del algoritmo que determinó su nivel de riesgo. Además, parece ser que se mantiene un patrón de opacidad y falta de transparencia, lo que dificulta la identificación y corrección de los sesgos incorporados (Pasquale, 2015). Sin embargo, la investigación empírica sobre la aparición de sesgos en herramientas digitales sigue siendo limitada. Aunque existe un amplio debate teórico sobre el potencial de los sistemas de inteligencia artificial para reproducir o amplificar los prejuicios humanos (O'Neil, 2016; Eubanks, 2018), los estudios comparativos entre decisiones humanas y algoritmos son escasos. Aquellos trabajos que abordan esta comparación indican que, en ciertos contextos, las herramientas digitales pueden discriminar menos que los operadores humanos, aunque siguen existiendo riesgos derivados de los sesgos

incrustados en los datos y modelos empleados (Berk, 2019; McKay, 2019).

Los resultados de esta revisión no permiten confirmar de manera concluyente las advertencias formuladas por O'Neil (2016) en *Weapons of Math Destruction*, donde sostiene que los algoritmos de toma de decisiones tienden a institucionalizar la discriminación bajo el aparente manto de objetividad matemática. En una línea similar, Eubanks (2018), en *Automating Inequality*, argumenta que los sistemas automatizados refuerzan la marginación de los colectivos más vulnerables al reproducir y amplificar las inequidades preexistentes presentes en los datos utilizados durante su diseño y entrenamiento. No obstante, algunos autores, como Broadhurst et al. (2019), sostienen que los algoritmos, si son diseñados, calibrados y supervisados adecuadamente, pueden aportar mayor objetividad y consistencia al proceso judicial, reduciendo la variabilidad derivada de los prejuicios individuales de los operadores humanos. Sin embargo, los hallazgos de esta revisión otorgan un mayor peso a las posiciones críticas, como las de Peeters & Schuilenburg (2018), quienes alertan sobre el riesgo de consolidar una justicia opaca y tecnocrática cuando los sistemas de inteligencia artificial no están sujetos a mecanismos efectivos de transparencia, auditoría y rendición de cuentas. En definitiva, los resultados sugieren que el sesgo en el ámbito de la justicia digital no se limita a una mera traslación del plano humano al digital, sino que implica una reconfiguración sistémica, determinada por los datos de entrenamiento, los modelos empleados y los objetivos institucionales que guían el diseño y la implementación de los sistemas algorítmicos (Barocas, Hardt & Narayanan, 2023).

Una de las principales limitaciones de la metodología empleada en el presente trabajo radica en la propia naturaleza de las revisiones sistemáticas. Este tipo de metodología tiene su origen en el ámbito de la salud, donde fue desarrollada con el objetivo de sintetizar evidencia empírica que respalde la toma de decisiones clínicas. Por este motivo, uno de los criterios esenciales para la inclusión de estudios es su carácter empírico.

La aplicación de esta metodología al campo del derecho presenta ciertos desafíos. A pesar del creciente desarrollo de la investigación jurídica empírica, la mayor parte de la producción académica en esta área continúa siendo de corte teórico. En

consecuencia, se ha excluido un amplio número de literatura teórica que, si bien no cumple con los criterios establecidos, podría haber enriquecido el análisis desde una perspectiva conceptual más amplia. Esta limitación es relevante, debería plantearse la adaptación de este tipo de metodologías a las ciencias sociales.

En definitiva, esta revisión sistemática confirma que los sesgos tradicionales presentes en la justicia humana no solo persisten en el ámbito digital, sino que, en algunos casos, adoptan nuevas formas que plantean desafíos adicionales para garantizar un sistema judicial justo y equitativo. Aunque el debate sobre los sesgos se ha trasladado al ámbito digital, la evidencia empírica disponible sobre su manifestación en los sistemas de justicia digital es todavía limitada y emergente. Los algoritmos, lejos de eliminar los prejuicios, corren el riesgo de trasladarlos y amplificarlos si no se diseñan e implementan con una atención rigurosa a los principios de equidad y justicia (O'Neil, 2016; Eubanks, 2018). Estas conclusiones subrayan la necesidad de continuar investigando el impacto real de la digitalización sobre la equidad judicial, así como de implementar marcos normativos robustos, mecanismos de transparencia algorítmica y procedimientos de auditoría independientes que aseguren el respeto de los principios fundamentales del derecho (Kroll et al., 2017). En suma, la promesa de una justicia objetiva e imparcial mediante el uso de algoritmos sigue estando condicionada por las limitaciones humanas, ahora incrustadas en los sistemas tecnológicos que diseñamos (Pasquale, 2015). La justicia digital solo será tal si es capaz de combinar el rigor tecnológico con una ética del cuidado y la equidad en cada decisión automatizada.

CAPITULO 5. DIGITALIZACIÓN Y ALGORITMIZACIÓN DE LA JUSTICIA EN EL SISTEMA DE JUSTICIA PENAL Y SUS IMPLICACIONES EN LA PRÁCTICA PROFESIONAL: UN ANÁLISIS CUALITATIVO.

La digitalización del sistema de justicia penal, y especialmente la incorporación progresiva de herramientas basadas en inteligencia artificial, está transformando de manera profunda la gestión de los casos, los procesos de toma de decisiones y la organización del trabajo en los ámbitos judicial, policial y penitenciario. Este proceso de cambio, aunque impulsado por la búsqueda de eficiencia y modernización institucional, plantea al mismo tiempo importantes desafíos éticos, jurídicos y organizativos que afectan a múltiples dimensiones del sistema de justicia penal, y en especial, directamente al ejercicio profesional.

La literatura reciente sugiere que la confianza de los operadores jurídicos y de los cuerpos de seguridad en las herramientas algorítmicas está mediada por factores como el grado de comprensión de su funcionamiento, la percepción de su utilidad real en tareas específicas, la transparencia de los procesos de decisión automatizada y la preocupación por posibles sesgos o errores (Fine, Le & Miller, 2024; Richardson, Schultz & Crawford, 2022; Xu & Wang, 2021). En este contexto, los estudios cualitativos se presentan como un instrumento idóneo para captar la complejidad de estas percepciones, así como para identificar los ámbitos en los que la digitalización genera oportunidades de mejora y aquellos en los que plantea dilemas éticos o tensiones en la práctica profesional.

Es por ello, que analizar esta transformación solo desde una perspectiva técnica o instrumentales es insuficiente, ya que las tecnologías se incorporan en contextos institucionales específicos y en dinámicas de trabajo marcadas por valores, rutinas y juicios humanos. Para comprender sus implicaciones es necesario considerar la experiencia y la percepción de quienes interactúan a diario con estas herramientas, pues jueces, fiscales, abogados, personal penitenciario y fuerzas de seguridad son quienes las aplican en su práctica y les otorgan sentido. A pesar de la relevancia de su papel, la evidencia empírica sobre cómo estos profesionales perciben e integran las tecnologías digitales en su labor cotidiana sigue siendo escasa, especialmente en

el contexto español.

En esta línea, el presente capítulo se centra en los dos estudios cualitativos orientados a examinar las implicaciones y las visiones que los profesionales del ámbito judicial, policial y penitenciario tienen respecto a la digitalización, la algoritmización y el uso de inteligencia artificial en su labor cotidiana. A través del análisis de sus discursos y experiencias, se busca comprender cómo se relacionan con estas tecnologías, cómo las interpretan y de qué modo se incorporan en sus prácticas profesionales, con el objetivo de aportar un conocimiento situado que contribuya a una implementación más reflexiva, legítima y responsable de la justicia digital.

Estudio 1. El impacto de la digitalización en el contexto de la seguridad ciudadana.

1. Justificación.

El proceso de digitalización en el ámbito de la seguridad ha cobrado una relevancia creciente en el contexto de las ciudades inteligentes (*Smart Cities*), donde la tecnología se perfila como un eje fundamental para la optimización de los recursos y la mejora de la calidad de vida de la ciudadanía (Fernández-Güell, 2015; Colado, Gutiérrez, Vives, & Valencia, 2014). Dentro de este marco, las Fuerzas y Cuerpos de Seguridad han comenzado a integrar herramientas de inteligencia artificial, análisis de datos y sistemas de vigilancia inteligente con el objetivo de mejorar su operatividad y capacidad de respuesta ante amenazas delictivas.

Uno de los enfoques predominantes en el debate sobre la digitalización en la seguridad es el tecno-utópico, que sostiene que la incorporación de IA y *Big Data* en la operativa policial traerá beneficios innegables en términos de prevención del delito, asignación eficiente de recursos y reducción de la criminalidad (Meijer & Wessels, 2019). Este enfoque ha llevado a la proliferación de sistemas de *Predictive Policing*, los cuales utilizan grandes volúmenes de datos para identificar patrones delictivos y anticipar la ocurrencia de incidentes (Ratcliffe, 2019; Weisburd & Eck, 2004). Si bien estos sistemas han demostrado cierta utilidad en la orientación estratégica de las patrullas y la asignación de efectivos, es fundamental analizar desde una perspectiva crítica hasta qué punto estas herramientas responden a las necesidades reales de la Policía Nacional y la Guardia Civil en su labor cotidiana, o si su implementación genera nuevas barreras operativas y éticas.

A pesar del creciente interés en la digitalización de la seguridad, existe una falta de estudios empíricos que analicen las percepciones y necesidades de los propios profesionales que integran los cuerpos policiales (Gómez-Bellvís & Esteve Bañón, 2022). La literatura científica ha tendido a centrarse en la aceptación ciudadana de estas tecnologías (Real Castrillo, 2021), dejando en un segundo plano la opinión y experiencias de los agentes que interactúan directamente con estas herramientas. Esta laguna en el conocimiento representa una limitación significativa, ya que la

implementación de tecnologías digitales en la seguridad debe estar alineada con las demandas reales de los profesionales del sector y no solo con las expectativas tecnológicas o políticas.

Además, la digitalización de la seguridad no debe verse únicamente como una herramienta para la persecución del delito, sino como un mecanismo para mejorar la relación entre las Fuerzas y Cuerpos de Seguridad y la ciudadanía. En sociedades democráticas, la seguridad pública no solo implica la reducción de la criminalidad, sino también el fortalecimiento de la confianza de la población en sus instituciones de seguridad (Boba, 2019). Por lo tanto, comprender cómo los agentes perciben estas tecnologías y qué barreras enfrentan en su implementación es un paso fundamental para garantizar que la digitalización de la seguridad en España se realice de manera ética, efectiva y alineada con los valores democráticos.

En este sentido, la presente investigación se justifica en la necesidad de comprender cómo la digitalización está impactando en la labor diaria de la Policía Nacional y la Guardia Civil, permitiendo proporcionar evidencia empírica que permita a las administraciones públicas desarrollar estrategias de digitalización alineadas con las necesidades reales de los profesionales. A diferencia de otros enfoques que han abordado la digitalización desde una perspectiva teórica o tecnológica, esta investigación adopta un enfoque basado en la experiencia directa de los profesionales de la seguridad, permitiendo identificar áreas clave de mejora y posibles obstáculos en la implementación de nuevas tecnologías. El análisis de las percepciones y experiencias de los profesionales de la seguridad permitirá generar un conocimiento más profundo sobre las oportunidades y limitaciones de la digitalización en este ámbito. En última instancia, la evidencia empírica obtenida en esta investigación servirá como base para la formulación de políticas públicas más informadas, asegurando que las herramientas digitales contribuyan de manera efectiva a la seguridad ciudadana sin comprometer los principios de equidad, autonomía profesional y supervisión democrática.

2. Objetivos

El presente estudio tiene como objetivo principal identificar oportunidades de

mejora en la gestión de la seguridad ciudadana a través de la digitalización, considerando los desafíos actuales y las necesidades de las Fuerzas y Cuerpos de Seguridad. Para poder alcanzarlo, se establecen los siguientes objetivos específicos

OE1. Analizar las principales limitaciones tecnológicas en la seguridad ciudadana y su impacto en la eficiencia operativa de los cuerpos policiales.

OE2. Identificar áreas prioritarias para la digitalización que contribuyan a optimizar la gestión de la seguridad pública.

3. Metodología.

Para poder alcanzar los objetivos planteados, se ha usado un enfoque de investigación mixto, una metodología grupal que combina estrategias cualitativas y cuantitativas denominada técnica de grupos nominales (TGN) (Tashakkori & Creswell, 2007). Este método de investigación social mixto permite analizar problemas con un objetivo específico: poder combinar opiniones individuales para permitir la toma de decisiones (Zanón, 1990). Este enfoque metodológico promueve el consenso al realizar en un marco conducente a la convergencia de opiniones. Su metodología facilita la obtención de conclusiones concretas y generando resultados tangibles adaptados al contexto concreto de la problemática. Además, se valora la participación igualitaria en la consideración de todos los participantes, asegurando que el éxito de las propuestas no esté condicionado por la exposición ni por la posición jerárquica de los participantes. En este contexto, se minimizan los conflictos otorgando a cada miembro del grupo la oportunidad de contribuir de manera significativa con sus ideas. Este enfoque inclusivo y participativo facilita la clasificación y clarificación del panorama hasta obtener un conjunto manejable de alternativas, propiciando así una toma de decisiones eficaz y deliberada (Delbecq & Van de Ven, 1971).

3.1. Muestra.

El muestreo en métodos cualitativos, como los grupos nominales, difiere de los métodos cuantitativos. Se centra en muestras intencionales en lugar de aleatorias, buscando seleccionar casos que proporcionen información detallada y relevante

sobre experiencias y necesidades. En este estudio específico, la elección se orientó hacia expertos y profesionales en seguridad ciudadana pertenecientes a las Fuerzas y Cuerpos de Seguridad de Policía Nacional, Policía Local y Guardia Civil. Se estableció un criterio único para la participación en la investigación: una experiencia mínima de cinco años en el respectivo cuerpo.

Para reclutar a los participantes, se envió una invitación para participar de manera voluntaria a egresados y alumnos del grado en Seguridad Pública y Privada de la Universidad Miguel Hernández de Elche. En dicha invitación se detallaba los criterios de participación y la metodología del estudio. Aquellos interesados se comunicaron con el Centro Crímina y se organizaron los grupos según su pertenencia al Cuerpo de Seguridad correspondiente.

Con esta metodología, se llevaron a cabo tres grupos nominales, cada uno integrado por expertos de los diferentes cuerpos de seguridad. El grupo de la Policía Nacional contó con 8 miembros, con una experiencia media en su campo de 13,3 años, compuesto equitativamente por cuatro mujeres y cuatro hombres. El grupo de la Policía Local estuvo compuesto por 7 participantes con una media de 8 años de experiencia, conformado por 4 hombres y 3 mujeres. Finalmente, el grupo de la Guardia Civil estuvo compuesto por 5 miembros (2 hombres y 3 mujeres) con una experiencia media de 17,2 años, según se detalla en la Tabla 4.

Tabla 4.

Datos demográficos de los grupos nominales por cuerpos.

GN	**N**	**Media (años de experiencia).**	**Distribución por sexo.**
Policía Nacional	8	13,3	H: 50% N=4 M: 50% N=4
Policía Local	7	8	H: 57% N=4 M: 43% N=3
Guardia Civil	5	17,2	H: 40% N=2 M: 60% N=3

3.2. Procedimiento.

Para comprender la técnica de los grupos nominales, es esencial explicar las distintas etapas que conforman su desarrollo, ya que siempre sigue la misma estructura (Zanón, 1990). Estas fases se dividen de la siguiente manera:

1. Fase de presentación: En esta primera fase, se instruye a los participantes sobre el propósito del estudio, las preguntas planteadas y el funcionamiento del Grupo Nominal (GN). Además, se les proporciona el documento de consentimiento informado para su firma.
2. Fase de generación silenciosa de ideas por escrito: Durante esta fase, los participantes responden de manera individual a la pregunta planteada por el investigador, expresando ideas concisas. Este enfoque asegura la equidad en la contribución de ideas y previene dinámicas de liderazgo.
3. Fase de registro de ideas: Aquí, las ideas son compartidas en rondas sucesivas, sin interacción directa entre los participantes. En el presente estudio, cada idea es registrada en una presentación de PowerPoint visible para todos, promoviendo la participación equitativa. Aunque no hay intervenciones directas, se fomenta la interacción y retroalimentación, ya que los participantes pueden tomar apuntes e identificar nuevas ideas.
4. Fase de debate aclaratorio: Durante esta fase, se discute cada idea presentada en la presentación de PowerPoint, creando un ambiente interactivo y de debate que subraya la naturaleza grupal de la técnica. Se permite la fusión, subdivisión o reformulación de las ideas presentadas.
5. Fase de votación para la priorización de las ideas: En esta fase, cada participante selecciona individualmente las ideas que considera más relevantes y las ordena según su importancia. Esto facilita la obtención de un consenso que respeta las consideraciones individuales, ofreciendo una perspectiva grupal. En este estudio, los participantes debían asignar puntos a las ideas, otorgando 5 puntos a la más relevante, 4 a la segunda, 3 a la tercera, 2 a la cuarta y 1 a la quinta entre las seleccionadas, permitiendo una ponderación de las ideas en función de su importancia percibida.

Estas cinco fases se realizaron de manera sucesiva para la pregunta relativa y

finalmente, a través de un formulario de Google realizaron la votación de estas.

3.3. Preguntas de investigación.

Para el presente estudio, la cuestión central que guió este estudio se formuló de la siguiente manera:

- "A partir de su experiencia profesional en la (Policía Nacional, Policía Local o Guardia Civil), ¿qué aspectos rutinarios de su trabajo podrían mejorar con la inclusión de la digitalización?".

3.4. Análisis de datos.

Para el análisis de las repuestas recogidas en el formulario de Google, se diseñó y programó una plantilla específica en Microsoft Excel, optimizada para la recogida, procesamiento y visualización automática de los datos obtenidos durante las sesiones. En primer lugar, la plantilla contempla un espacio para el registro de las valoraciones emitidas por cada participante respecto a las ideas generadas en el GN, donde se puntúa de 1 a 5. Cada columna (P1, P2, P3… Pn) corresponde a un participante, mientras que cada fila recoge una de las ideas consensuadas durante la fase de generación y clarificación.

El sistema implementa las siguientes operaciones automáticas:

- Cálculo de puntuaciones totales: Se emplea una fórmula de suma para agregar las puntuaciones otorgadas a cada idea, generando la columna "Puntos". Este valor permite identificar las propuestas mejor valoradas de forma directa.
- Cálculo de porcentajes relativos: Se calcula el peso porcentual de cada idea sobre el total de puntos acumulados en la sesión. Este cálculo se realiza de forma dinámica, de modo que cualquier actualización en las puntuaciones se refleja automáticamente en los resultados.
- Visualización gráfica de resultados: La plantilla incluye gráficos vinculados a las tablas de resultados que muestran, tanto en formato de barras como en sectores, la distribución de puntos y porcentajes. Estos gráficos facilitan la

comparación visual de la relevancia relativa de cada propuesta, agilizando la interpretación de resultados.

4. Resultados.

A continuación, se muestran los resultados derivados de los distintos grupos nominales con profesionales en el ámbito de la seguridad ciudadana. Se ha llevado a cabo un análisis descriptivo individual de cada grupo nominal.

4.1. Grupo nominal con la Policía Nacional.

En relación con el grupo nominal de la Policía Nacional, durante la fase de presentación de ideas se identificaron un total de 16 ideas relacionadas con posibles ámbitos de mejora de su desempeño profesional mediante procesos de digitalización. Durante la fase de discusión de ideas, la fusión de problemas dio lugar a que el total de problemas se redujera a 15 ideas.

Los ámbitos de mejora que recibieron una puntuación más alta fueron la "Bodycam y geolocalización para la seguridad", acumulando un 20,8% (25 puntos) del total de puntos repartidos entre las 15 ideas, seguido por la "localización de compañeros desde la calle" con un 11,7% (14 puntos) del total de puntos y la "identificación dactilar de las personas[7]", con un 9,2% del total de puntos. La "digitalización de documentos (atestados, denuncias…)", el "acceso remoto a expedientes judiciales" y la "mejora de la cobertura y rapidez del sistema Zebra[8]" recibieron el 7,5% de los votos, sin embargo, fue esta última idea la que se tuvo en consideración por un mayor número de participantes. El resto de las ideas identificadas muestran puntuaciones bajas y en su mayoría individuales, sin embargo, se ha de destacar que

[7] Esta propuesta se centra en la integración de un dispositivo que escanea la huella para obtener los datos de identificación de la persona.

[8] Zebra es una empresa responsable de proporcionar a las fuerzas policiales una gama de soluciones de transformación digital, así como de apoyarlas y prepararlas para los desafíos futuros. En este sentido, ha dotado a las Fuerzas y Cuerpos de Seguridad de soluciones consolidadas de informática móvil hasta escáneres, tablets e impresoras portátiles. La inclusión de estos dispositivos ha permitido que los agentes cuenten con los mismos recursos que un departamento de policía, pero con la movilidad de un coche patrulla.

todas las ideas fueron votadas.

En la Tabla 5, se presenta el listado completo de ideas relativas a los ámbitos de mejora de su desempeño profesional mediante procesos de digitalización durante la fase de discusión de ideas, así como las puntuaciones obtenidas por cada una de ellas y la distribución del total de puntos. Por otro lado, se indica el número de puntos obtenidos, el porcentaje de estos y las puntuaciones máximas y mínimas recibidas. A modo resumen, en la Figura 5 se pueden apreciar el total de puntuaciones de cada una de las ideas.

Figura 5.

Puntuación de cada una de las ideas obtenidas en el grupo nominal de Policía Nacional.

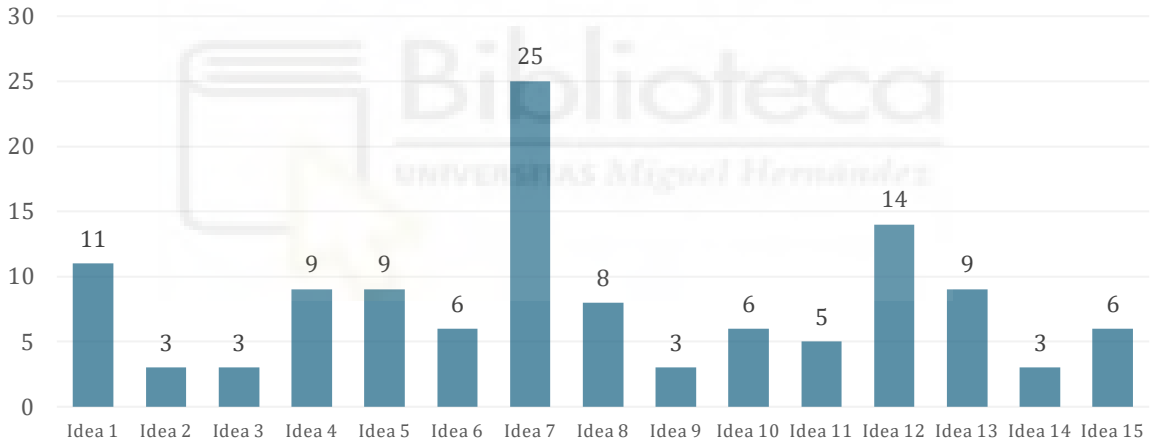


Tabla 5.

Ranking de los ámbitos susceptibles de mejora mediante la digitalización identificados por el grupo de Policía Nacional.

	P1	P2	P3	P4	P5	P6	P7	P8	PUNTOS	%	*M*	Máx.	Mín.
1. Identificación dactilar de las personas	0	0	0	3	4	0	4	0	11	9,2%	1,38	4	0
2. Localización GPS de las llamadas de atención al ciudadano	0	2	0	0	0	0	0	1	3	2,5%	0,38	2	0
3. Acceso a la planimetría de edificios y casas	0	3	0	0	0	0	0	0	3	2,5%	0,38	3	0
4. Acceso remoto a expedientes judiciales	0	0	0	5	0	2	0	2	9	7,5%	1,13	5	0
5. Mejora de la cobertura y rapidez del sistema Zebra	0	0	1	0	2	3	0	3	9	7,5%	1,13	3	0
6. Códigos QR de acceso a las denuncias	1	0	0	1	0	0	0	4	6	5,0%	0,75	4	0
7. Bodycam y geolocalización para la seguridad	5	5	5	2	1	4	3	0	25	20,8%	3,13	5	0
8. Comunicación con organismos externos de manera digital	0	0	3	0	0	0	5	0	8	6,7%	1,00	5	0
9. Equipo de transmisión con elementos gráficos (video, mensajes de voz, imágenes...)	0	0	0	0	3	0	0	0	3	2,5%	0,38	3	0
10. Acceder a imágenes de los operativos desde la sala	4	0	0	0	0	0	2	0	6	5,0%	0,75	4	0
11. Envío automatizado al GPS de los coches patrulla a las localizaciones	0	0	0	4	0	1	0	0	5	4,2%	0,63	4	0
12. Localización de compañeros desde la calle	0	4	0	0	5	5	0	0	14	11,7%	1,75	5	0
13. Digitalización de documentos (atestados, denuncias…)	0	0	4	0	0	0	0	5	9	7,5%	1,13	5	0
14. Autonomía del coche patrulla con información de sala	3	0	0	0	0	0	0	0	3	2,5%	0,38	3	0
15. Comunicación de información de los casos mediante dispositivos electrónicos.	2	1	2	0	0	0	1	0	6	5,0%	0,75	2	0
TOTALES										**120**	**100%**		

4.2. Grupo nominal con la Policía Local.

En el grupo de Policía Local se identificaron un total de 19 ideas durante la fase inicial, las cuales fueron debatidas y fusionadas, reduciéndose finalmente a 13 ideas durante la fase de discusión.

Los ámbitos de mejora que recibieron una puntuación más alta fueron la creación de "bases de datos con informes policiales compartidos entre cuerpos y servicios", acumulando el 19,0% (equivalente a 20 puntos) del total de puntos distribuidos entre las 13 ideas. Le siguió la propuesta de "instalar un lector de documentación que permita detectar falsificaciones con transferencia al acta", con un 16,2% (17 puntos) del total de puntos, y la sugerencia de "instalar tabletas en los vehículos de patrulla con acceso a las bases de datos", representando el 11,4% del total de puntos. Es relevante destacar que, aunque la diferencia entre la primera y la segunda idea es de solo 3 puntos, la idea con una puntuación más baja fue considerada por un mayor número de participantes.

En la Tabla 6, al igual que en el apartado previo, se muestra el listado completo de las ideas relacionadas con la mejora del desempeño profesional a través de procesos de digitalización durante la fase de discusión de ideas, junto con los diversos indicadores mencionados.

Figura 6.

Puntuación de cada una de las ideas obtenidas en el grupo nominal de Policía Local.

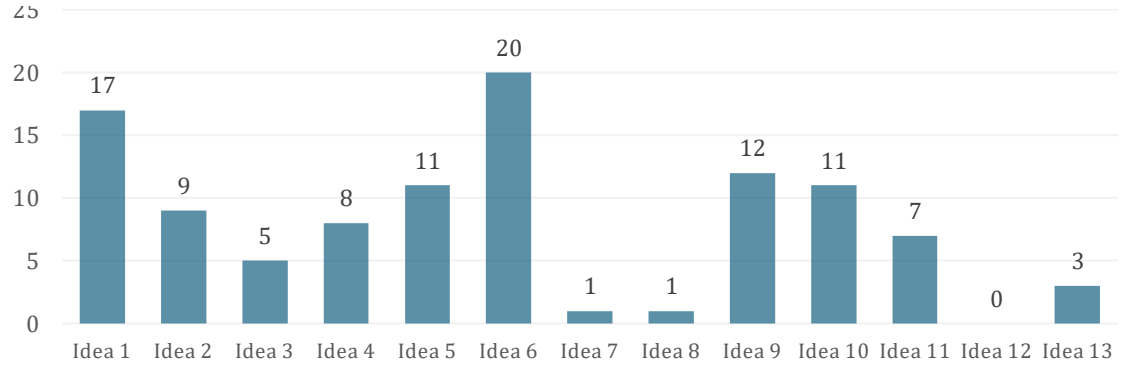


Tabla 6.

Ranking de los ámbitos susceptibles de mejora mediante la digitalización identificados por el grupo de Policía Local.

	P1	P2	P3	P4	P5	P6	P7	PUNTOS	%	*M*	Máx.	Mín.
1. Lector de documentación que permita detectar falsificaciones con traspaso al acta	3	1	2	5	2	4	0	17	16,2%	2,43	5	0
2. GPS con las calles automatizadas	0	3	4	0	1	1	0	9	8,6%	1,29	4	0
3. Organizador de documentos y archivos de manera compartimentada en una nube común	0	2	0	1	0	2	0	5	4,8%	0,71	2	0
4. Herramienta estandarizada de recogida de información	4	4	0	0	0	0	0	8	7,6%	1,14	4	0
5. Herramientas de contacto con otros departamentos y servicios	0	0	0	0	5	3	3	11	10,5%	1,57	5	0
6. Bases de datos con los informes policiales compartida entre cuerpos y servicios	5	5	1	0	4	5	0	20	19,0%	2,86	5	0
7. Posicionamiento GPS de patrullas y de las emisoras en tiempo real	1	0	0	0	0	0	0	1	1,0%	0,14	1	0
8. Base de datos con filtros para traspasar a informes	0	0	0	0	0	0	1	1	1,0%	0,14	1	0
9. Instalar tabletas en los coches patrulla con conexión a las bases de datos	0	0	5	3	0	0	4	12	11,4%	1,71	5	0
10. Acceso y compartir la información de manera automatizada	2	0	0	4	0	0	5	11	10,5%	1,57	5	0
11. Facilitación en la instalación y acceso a cámaras en seguridad pública	0	0	0	2	3	0	2	7	6,7%	1,00	3	0
12. Resúmenes con la información de las bases de datos que permita generar estadísticas	0	0	0	0	0	0	0	0	0,0%	0,00	0	0
13. Geoposicionamiento de la llamada para traspasar la información a las patrullas	0	0	3	0	0	0	0	3	2,9%	0,43	3	0
TOTALES								105	100%			

4.3. Grupo nominal con la Guardia Civil.

Durante la fase de presentación de ideas en el grupo de la Guardia Civil, se identificaron un total de 11 propuestas. Estas se redujeron a 9 posibles áreas de mejora en su desempeño profesional mediante la digitalización, durante la fase de discusión de ideas.

Las áreas de mejora que obtuvieron una puntuación más alta fueron las siguientes: en primer lugar, la creación de una "base de datos común que permita el acceso por parte de todos los cuerpos", la cual acumuló un 26,7% (20 puntos) del total de puntos asignados entre las 9 ideas. En segundo lugar, se destacó la necesidad de "actualizaciones continuas de los dispositivos", con un 14,7% (11 puntos) de los votos. Le siguieron en porcentaje de votos, "Patrullas con recepción de información enviada desde sala" y la "unificación de la recogida de denuncias en una misma aplicación", ambas con un total de 13,3% (10 puntos) cada una. Es importante señalar que estas dos últimas propuestas fueron votadas por un 80% y un 60% de los participantes, respectivamente.

En la figura 6 se muestran todas las puntuaciones obtenidas de cada una de las ideas. Para un análisis más detallado del listado completo de ideas y sus indicadores, se puede consultar la Tabla 7 mencionada anteriormente.

Figura 7.

Puntuación de cada una de las ideas obtenidas en el grupo nominal de guardia civil.

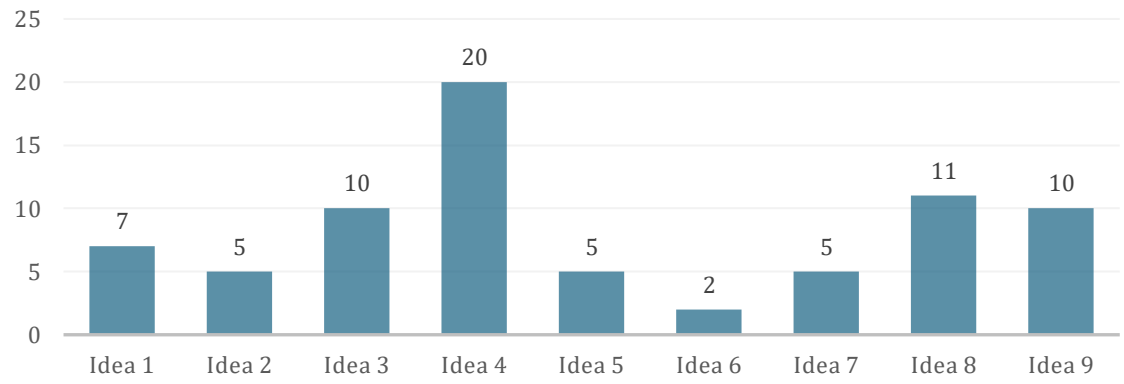


Tabla 7. Ranking de los ámbitos susceptibles de mejora mediante la digitalización identificados por el grupo de Guardia Civil.

	P1	P2	P3	P4	P5	PUNTOS	%	M	Máx.	Mín.
1. Envío de documentos multimedia a las tabletas de los coches patrulla.	0	1	0	2	4	7	9,3%	1,40	4	0
2. APP en los coches patrulla para incluir averías o problemas detectados	0	0	3	0	2	5	6,7%	1,00	3	0
3. Patrullas con recepción de información enviada desde sala	3	3	1	0	3	10	13,3%	2,00	3	0
4. Base de datos común que permita el acceso por parte de todos los cuerpos	4	5	2	4	5	20	26,7%	4,00	5	2
5. Mejora de cobertura de los dispositivos	0	2	0	3	0	5	6,7%	1,00	3	0
6. Almacenamiento de documentación en una nube común	1	0	0	0	1	2	2,7%	0,40	1	0
7. Dispositivos con mayores accesos (sustituyendo el móvil personal)	5	0	0	0	0	5	6,7%	1,00	5	0
8. Actualizaciones continuadas de los dispositivos	2	0	4	5	0	11	14,7%	2,20	5	0
9. Unificación de la recogida de las denuncias en una misma aplicación	0	4	5	1	0	10	13,3%	2,00	5	0
TOTALES						75	100%			

4.4. Comparación de los tres grupos nominales.

Para poder comparar de manera agregada los tres grupos nominales realizados con los diferentes cuerpos de seguridad del estado se realizó, en primer lugar, un listado de todos los ámbitos de mejora de su desempeño profesional mediante procesos de digitalización. Seguidamente, se realizó una clasificación de las ideas creando una categoría general que pudiera fusionar aquellas coincidentes. En este paso se tuvo en cuenta también la discusión generada durante el desarrollo de los grupos nominales. En este proceso, se eliminaron aquellas ideas que no eran coincidentes, de manera que, partiendo de 37 ideas generales, finalmente se redujo a 7 ideas.

En la Tabla 8 pueden observarse las categorías generales que englobarían los diferentes ámbitos de mejora del desempeño profesional mediante la digitalización que requieren de atención.

Tabla 8.

Ideas resultantes de la comparación entre los tres grupos nominales.

Categoría general	**Policía Nacional**	**Policía Local**	**Guardia Civil**
Instalación de dispositivos con envío y recepción de multimedia	Equipo de transmisión con elementos gráficos (video, mensajes de voz, imágenes...).	Instalar tablets en las patrullas con conexión a las bases de datos.	Envío de documentos multimedia a las tablets de las patrullas.
Localización	Localización GPS de las llamadas de atención al ciudadano.	GPS con las calles automatizadas.	-
Bases de datos	Bases de datos con los informes policiales compartida entre cuerpos y servicios	-	Base de datos común que permita el acceso por parte de todos los cuerpos
Posicionamiento GPS	Bodycam y geolocalización para la seguridad. Envío automatizado al GPS de los coches patrulla a las localizaciones	Posicionamiento GPS de patrullas y de las emisoras en tiempo real. Geoposicionamiento de la llamada para traspasar la información a las patrullas	-
Almacenado en línea de documentos	-	Organizador de documentos y archivos de manera compartimentada en una nube común	Almacenamiento de documentación en una nube común
Cobertura	Mejora de la cobertura y rapidez del sistema Zebra	-	Mejora de cobertura de los dispositivos
Estandarización de la recogida de información	-	Herramienta estandarizada de recogida de información	Unificación de la recogida de las denuncias en una misma aplicación

5. Discusión y conclusiones.

El avance de las tecnologías digitales, en particular aquellas vinculadas a la inteligencia artificial, el Internet de las Cosas (IoT) y el análisis de datos, ha generado grandes expectativas sobre su impacto en la seguridad pública y la gestión de ciudades inteligentes (Yigitcanlar, Mehmood & Corchado, 2021). La integración de estas tecnologías en las Fuerzas y Cuerpos de Seguridad se presenta como una oportunidad para mejorar la eficiencia operativa, la capacidad de respuesta y la toma de decisiones informadas. Sin embargo, como muestran los resultados de este estudio, la digitalización no es una solución en sí misma a los problemas sociales o criminales, sino una herramienta que, si se adapta correctamente a las necesidades reales de los profesionales de la seguridad, puede optimizar sus funciones sin reemplazar el criterio humano (Davenport & Kalakota, 2019).

Uno de los principales hallazgos de esta investigación es la identificación de prioridades específicas para la modernización de los cuerpos policiales en España. En particular, los participantes destacaron la necesidad de mejorar los sistemas de geolocalización, incorporando herramientas de posicionamiento GPS que permitan compartir información en tiempo real entre agentes y unidades operativas. Actualmente, los sistemas existentes presentan limitaciones debido a mapas desactualizados o direcciones incompletas, lo que dificulta su eficacia en situaciones de emergencia (Lum, Koper, & Wu, 2022). Este problema se agrava en áreas rurales, donde la cobertura de dispositivos es deficiente y limita la capacidad de respuesta de los agentes.

Otro aspecto clave identificado es la urgencia de crear una base de datos centralizada accesible en línea, una demanda reiterada por la Policía Local y la Guardia Civil. La fragmentación de los sistemas de información dificulta la colaboración interinstitucional y retrasa la transmisión de datos críticos en operaciones policiales. La literatura existente señala que la interoperabilidad de bases de datos es fundamental para mejorar la eficiencia y la coordinación entre agencias de seguridad (Ferguson, 2017). En este sentido, la creación de una infraestructura digital compartida podría optimizar la gestión de la información y reducir tiempos de respuesta en la resolución de casos.

Asimismo, los resultados subrayan la importancia de desarrollar herramientas de almacenamiento remoto de documentos. La posibilidad de acceder a información clave en tiempo real facilitaría la labor policial y reduciría la dependencia de registros físicos, mejorando la movilidad y flexibilidad operativa de los agentes. Además, la digitalización de la recogida de información se percibe como un área prioritaria, con una fuerte demanda por parte de los profesionales para la implementación de una aplicación unificada que permita registrar denuncias y compartir información entre los distintos cuerpos de seguridad (Moses & Chan, 2018). La falta de estandarización en estos procedimientos ha sido identificada como un obstáculo para la eficiencia del trabajo policial y la toma de decisiones basada en datos.

Desde una perspectiva metodológica, el uso de grupos nominales en esta investigación ha permitido generar un debate valioso entre los participantes, facilitando la identificación de problemas concretos y soluciones viables en la implementación de tecnologías digitales en el ámbito policial. Si bien este enfoque presenta limitaciones en términos de generalización de resultados, su capacidad para captar las necesidades específicas de los profesionales de la seguridad proporciona una base sólida para el diseño de políticas públicas y estrategias de modernización adaptadas a la realidad operativa de cada cuerpo policial (Chan, 2021).

En conclusión, los hallazgos de este estudio refuerzan la importancia de una implementación tecnológica basada en la realidad operativa de las Fuerzas y Cuerpos de Seguridad, en lugar de una adopción indiscriminada de innovaciones tecnológicas. La digitalización debe responder a las necesidades concretas de los profesionales y facilitar su trabajo sin generar nuevas barreras operativas. Para ello, es esencial que el desarrollo e implementación de estas soluciones tecnológicas se realice en estrecha colaboración con los usuarios finales, asegurando así su aceptación, eficacia y utilidad en el contexto de la seguridad pública. En última instancia, el éxito de la transformación digital en el ámbito policial dependerá de su capacidad para integrarse de manera eficiente con las prácticas y dinámicas de trabajo existentes, sin perder de vista los límites y responsabilidades de las fuerzas

de seguridad en la gestión de los problemas sociales y criminales.

Estudio 2. El impacto de la digitalización y la inteligencia artificial en operadores del sistema de justicia penal.

1. Justificación.

Como se ha podido comprobar a lo largo de la presente tesis doctoral, la progresiva digitalización del sistema de justicia penal, especialmente a través del uso de tecnologías basadas en inteligencia artificial, está transformando no sólo la gestión de los procedimientos judiciales y penitenciarios, sino también la manera en que los profesionales toman decisiones y ejercen sus funciones. La IA permite automatizar tareas, analizar grandes volúmenes de datos jurídicos y predecir riesgos, facilitando, por ejemplo, la asistencia a jueces en sus resoluciones o la clasificación de personas internas en función de su peligrosidad. Sin embargo, el éxito de estas herramientas no depende exclusivamente de sus capacidades técnicas, sino de su aceptación, comprensión y uso efectivo por parte de quienes las integran en su práctica profesional.

Diversas directrices internacionales, como las promovidas por la UNESCO, insisten en que la implementación ética de la IA debe respetar los derechos fundamentales, promover la transparencia, evitar sesgos y garantizar el control humano sobre los sistemas automatizados. No obstante, su cumplimiento efectivo está mediado por las percepciones que los operadores del sistema penal tienen sobre estas tecnologías. La literatura indica que una falta de conocimiento técnico, la percepción de amenazas a la autonomía profesional o las dudas sobre la transparencia y explicabilidad de los algoritmos pueden generar resistencia, desconfianza o rechazo hacia estas herramientas, incluso si normativamente están avaladas (UNESCO, 2021; Galli & Sartor, 2023).

En este sentido, conocer las percepciones y actitudes de jueces, fiscales, abogados, personal penitenciario y fuerzas de seguridad resulta crucial para entender en qué tareas consideran útil la IA, qué riesgos identifican y qué condiciones creen necesarias para una adopción ética y efectiva (Weinstein, 2022). Este enfoque resulta aún más relevante en un contexto como el español, donde la investigación empírica sobre el impacto de la IA en la práctica profesional del sistema de justicia

penal sigue siendo escasa.

Así, este trabajo se propone contribuir a cerrar la brecha existente entre el diseño normativo de la digitalización judicial y su implementación práctica, integrando la perspectiva de los profesionales como condición indispensable para garantizar una justicia digital que respete los principios del Estado de derecho, preserve la autonomía decisional y fomente una gobernanza algorítmica responsable.

2. Objetivos.

En este sentido, se plantea el presente estudio que tiene como objetivo general analizar la percepción de los operadores del sistema penal y penitenciario respecto a la implementación de herramientas de inteligencia artificial.

De este objetivo general se derivan los siguientes objetivos específicos:

OE1. Explorar las percepciones de profesionales del sistema penal sobre el impacto de las herramientas de inteligencia artificial en su labor cotidiana.

OE2. Identificar los tipos de tareas o funciones para las cuales los operadores consideran que las herramientas algorítmicas presentan un mayor potencial de utilidad en su ámbito profesional.

OE3. Analizar los principales retos, resistencias y desafíos éticos, técnicos y organizativos que los operadores identifican como prioritarios para una implementación adecuada y efectiva de estas tecnologías.

3. Metodología.

Dado que para la presente investigación se empleó la misma metodología del estudio anterior[9], simplemente se hará un breve resumen de la técnica empleada.

[9] Véase la metodología del estudio 1 incluido en el capítulo 5 (pag 194).

Para la presente investigación se empleó nuevamente la Técnica de Grupos Nominales (TGN), un enfoque mixto que combina elementos cualitativos y cuantitativos con el objetivo de recoger, debatir y priorizar ideas a través de la participación estructurada de profesionales. Esta técnica favorece el consenso, garantiza la equidad en la aportación de propuestas y permite obtener resultados concretos y contextualizados a partir de un proceso colaborativo. Su desarrollo se estructura en cinco fases sucesivas: generación individual de ideas, registro colectivo, debate aclaratorio, votación y priorización, lo que facilita una deliberación eficaz y una síntesis compartida de perspectivas.

3.1. Muestra.

La selección de los participantes se realizó mediante la técnica de bola de nieve, cuyos puntos de arranque fueron diferentes contactos profesionales, entre ellos, académicos de diferentes universidades españolas y profesionales de la administración judicial. A todos ellos se les pidió colaboración para trasladar la invitación a participar en los grupos nominales a colegas de los ámbitos indicados. El texto de la invitación incluía una introducción sobre el proyecto, los objetivos de la invitación *("realizar grupos nominales con profesionales de ambos ámbitos para explorar el impacto de los algoritmos predictivos en su trabajo y cómo valoran los retos que son necesarios abordar para afrontar adecuadamente su progreso y su utilidad"*), una explicación sobre la dinámica a realizar "*Los grupos nominales son una técnica de investigación en la que primero se reúne información pidiendo a los participantes que respondan individualmente a las preguntas planteadas por un moderador, y después, de forma anónima, se les pide que prioricen las ideas o sugerencias manifestadas por los miembros del grupo. Este proceso ayuda a identificar experiencias y percepciones individuales, pero especialmente situaciones y preferencias que representan al colectivo*" e información sobre el formato (online) de los grupos y la duración (máximo 2 horas). Así mismo, se adjuntó una carta de invitación con información más detallada y con enlace a un Doodle, a través del cual se recogía la disponibilidad horaria de las personas interesadas.

Del ámbito judicial se solicitó la participación de jueces y fiscales, mientras que, en el ámbito penitenciario, se convocó a profesionales tanto de instituciones

penitenciarias como de los servicios de gestión de penas y medidas alternativas. Aunque inicialmente se obtuvo la aceptación de una decena de profesionales por cada ámbito, finalmente solo fue posible coordinar la disponibilidad de una parte reducida de ellos para llevar a cabo los grupos nominales de forma online.

El grupo nominal del ámbito judicial estuvo compuesto por cuatro participantes: un juez de instrucción y tres magistradas/os de Audiencias Provinciales, procedentes de las jurisdicciones de Cádiz, Alicante, Barcelona y Girona. En cuanto al grupo del ámbito penitenciario, estuvo conformado por cinco profesionales: una directora de centro penitenciario, un educador, un jurista y dos psicólogos, pertenecientes a cinco centros penitenciarios ubicados en diferentes comunidades autónomas (Aragón, Cantabria, Castilla y León, Canarias y Extremadura).

3.2. Preguntas de investigación.

Con la finalidad de alcanzar los objetivos planteados en el presente estudio, se incluyeron 2 preguntas principales que guiaron la dinámica del grupo nominal:

- Considerando su conocimiento sobre las diferentes aplicaciones para las que se están desarrollando en la actualidad las herramientas algorítmicas en el ámbito judicial/penitenciario. ¿Para apoyar qué tipo de funciones propias de su ámbito profesional considera que dichas herramientas pueden tener mayor potencial?
- Para avanzar en el proceso de consolidación de herramientas algorítmicas útiles para apoyar las funciones establecidas en la pregunta anterior ¿Qué tipo de desafíos o retos considera prioritarios abordar?.

4. Resultados.

A continuación, se muestran los resultados derivados del grupo nominal desarrollados con profesionales del sistema de justicia en función de cada una de las preguntas realizadas.

4.1. Usos potencialmente más útiles de la IA.

En cuanto a la primera pregunta, sobre qué usos de la IA perciben potencialmente

más útiles**,** durante la fase de presentación de ideas se identificaron un total de 13 propuestas relacionadas con las posibles aplicaciones de herramientas algorítmicas y procesos de digitalización en la mejora del desempeño profesional. Tras el debate grupal y la priorización individual, se procedió a la fase de votación, en la que se repartieron un total de 60 puntos entre las distintas propuestas.

La idea que obtuvo la puntuación más alta fue la relativa al uso de instrumentos de predicción del riesgo para la adopción de medidas cautelares, que acumuló un 23,3% del total de puntos (10 puntos), seguida de tres ideas que empataron con 8 puntos cada una (18,6%): cumplimiento de la pena de prisión en fase de ejecución, asistente de redacción de resoluciones judiciales con información del juez, y transcripción de vistas orales con selección de fragmentos. Estas ideas fueron las que recibieron mayor respaldo por parte de los participantes, tanto en puntuación como en número de votos individuales.

Otras ideas destacadas fueron el expediente digital, con un 16,3% (7 puntos), y el ajuste de la pena a las características del acusado desde la visión criminológica, con un 11,6% (5 puntos), idéntico porcentaje al que obtuvo la propuesta sobre la investigación policial de los casos. En cambio, propuestas como la síntesis de datos de jurisprudencia o la entrada automatizada al registro central de penados no recibieron puntuación alguna, lo que sugiere un menor grado de prioridad percibida por parte del grupo. En la Tabla 9, se presenta el listado completo de ideas valoradas durante la fase de priorización, así como la puntuación obtenida por cada una, el porcentaje correspondiente, y la distribución de puntuaciones máximas y mínimas. Estos resultados permiten identificar con claridad aquellas áreas que los operadores judiciales consideran prioritarias en el proceso de digitalización e integración de herramientas algorítmicas.

Figura 8.

Puntuación de cada una de las ideas obtenidas en el grupo nominal en el ámbito judicial - Qué usos de la IA perciben potencialmente más útiles.

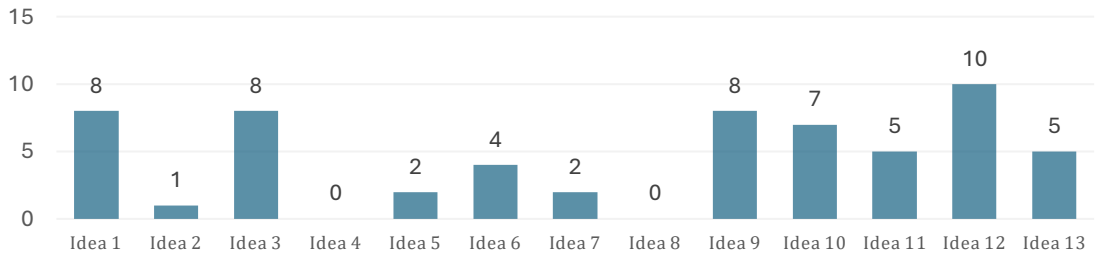


Tabla 9.

Usos potenciales valorados por el grupo del ámbito judicial.

	P1	P2	P3	P4	PUNTOS	%	M	Máx	Mín
1. Cumplimiento de la pena de prisión (fase de ejecución)	0	0	3	5	8	18,6%	2	5	0
2. Facilitar la búsqueda de jurisprudencia	1	0	0	0	1	2,3%	0,25	1	0
3. Asistente redacción de resoluciones judiciales con información del juez	0	3	5	0	8	18,6%	2	5	0
4. Síntesis de los datos de jurisprudencia	0	0	0	0	0	0,0%	0	0	0
5. Redacción mediante dictado de voz o transcripción de documentos	0	2	0	0	2	4,7%	0,5	2	0
6. Decidir la entrada en prisión (suspensión)	0	0	0	4	4	9,3%	1	4	0
7. Generación de modelos de resolución	2	0	0	0	2	4,7%	0,5	2	0
8. Entrada automatizada para la Gestión del registro central de penados	0	0	0	0	0	0,0%	0	0	0
9. Transcripción de vistas orales con selección de fragmentos	4	1	2	1	8	18,6%	2	4	1
10. Expediente digital – facilitar que material se adjunte al expediente según un índice	0	4	1	2	7	16,3%	1,75	4	0
11. Ajuste de la pena a características propias del acusado desde la visión criminológica	5	0	0	0	5	11,6%	1,25	5	0
12. Instrumentos de predicción del riesgo para la adopción de medidas cautelares	3	0	4	3	10	23,3%	2,5	4	0
13. Investigación policial de los casos	0	5	0	0	5	11,6%	1,25	5	0
TOTALES					60	100%			

En cuanto al ámbito penitenciario, se recogieron un total de 16 propuestas centradas en distintas funciones susceptibles de ser apoyadas por herramientas

algorítmicas y procesos de digitalización. Tras la exposición de ideas, el debate grupal y la priorización individual, se procedió a la fase de votación, en la que se distribuyeron un total de 75 puntos entre las diferentes propuestas.

La idea que obtuvo la mayor puntuación fue la búsqueda rápida de jurisprudencia y doctrina para la realización de informes, con 10 puntos (13,3%), reflejando una clara necesidad de apoyo en la elaboración documental. Le siguieron dos propuestas con 9 puntos cada una (12%): la determinación del riesgo e intervención para la prevención de recaídas y la valoración del riesgo de reincidencia, ambas ligadas a la gestión del riesgo y a la toma de decisiones sobre el tratamiento penitenciario. Con 8 puntos (10,7%), se situaron otras dos ideas destacadas: la elección y diseño del plan de tratamiento y el pronóstico de riesgo de violencia en internos de alto riesgo, esta última especialmente relevante para la prevención de incidentes graves en entornos penitenciarios. También obtuvo una puntuación considerable la identificación de factores de riesgo para prevenir el suicidio, con 7 puntos (9,3%), lo que subraya la importancia de la salud mental como prioridad institucional.

Entre las propuestas con puntuaciones medias, se encuentra la valoración del riesgo de quebrantamiento (6 puntos, 8%) y la organización del trabajo con notificaciones internas (5 puntos, 6,7%), junto con la realización de pronósticos como ayuda (5 puntos, 6,7%). Por otro lado, varias ideas no recibieron puntuación, como la resolución telemática de consultas de internos, la identificación de patrones de convivencia o la predicción del suicidio sin indicadores concretos, lo que indica una menor prioridad o utilidad percibida por los participantes.

Figura 9.

Puntuación de cada una de las ideas obtenidas en el grupo nominal en el ámbito penitenciario - Qué usos de la IA perciben potencialmente más útiles.

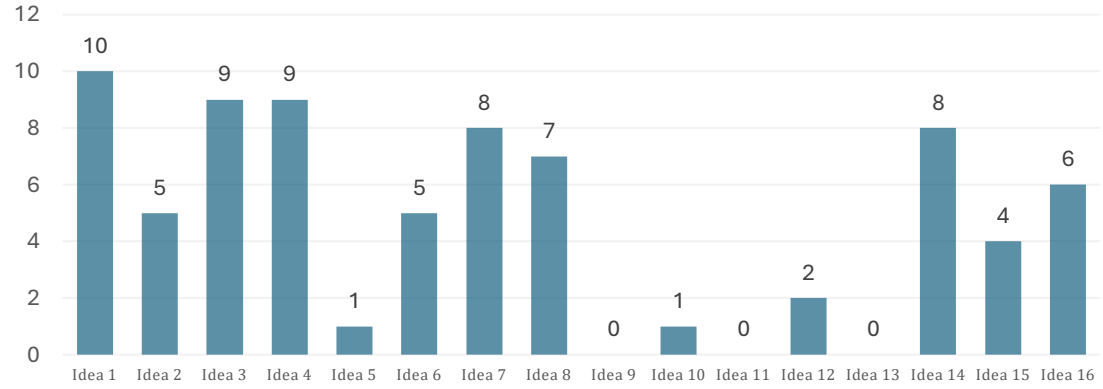


Tabla 10.

Usos potenciales valorados por el grupo del ámbito penitenciario.

	P1	P2	P3	P4	P5	PUNTOS	%	M	Máx	Mín
1. Búsqueda rápida de jurisprudencia y doctrina para la realización de informes	5	0	0	0	5	10	13,3%	2	5	0
2. Realización de pronósticos (como ayuda)	0	0	0	5	0	5	6,7%	1	5	0
3. Determinación del riesgo e Intervención para la prevención de recaídas (identificación de factores de riesgo)	4	0	5	0	0	9	12,0%	1,8	5	0
4. Valoración del riesgo de reincidencia	1	4	0	4	0	9	12,0%	1,8	4	0
5. Concretar los componentes del itinerario tratamentales individualmente según las necesidades (no aplicación trat completo)	0	0	1	0	0	1	1,3%	0,2	1	0
6. Organización de trabajo con notificaciones (agendas, revisiones anuales, plazos de los internos, etc)	0	1	0	0	4	5	6,7%	1	4	0
7. Elegir y diseñar plan de tratamiento	2	0	0	3	3	8	10,7%	1,6	3	0
8. Identificación factores de riesgo y determinación de un valor que permita prevenir el suicidio	0	3	4	0	0	7	9,3%	1,4	4	0
9. Determinar adecuación para aplicación del art 60 o 104 o alternativas a la prisión	0	0	0	0	0	0	0,0%	0	0	0
10. Identificación de entradas y salidas de internos	0	0	0	0	1	1	1,3%	0,2	1	0
11. Resolver consultas de internos, solicitud de	0	0	0	0	0	0	0,0%	0	0	0

	P1	P2	P3	P4	P5	PUNTOS	%	M	Máx	Mín
consultas de manera telemática.										
12. Pronóstico de seguridad interior	0	0	0	2	0	2	2,7%	0,4	2	0
13. Identificar patrones de comportamiento convivencial asociados a variables. ej tiempo, relaciones interpersonales, etc	0	0	0	0	0	0	0,0%	0	0	0
14. Pronóstico de riesgo de violencia dentro de prisión para internos de alto riesgo	3	2	3	0	0	8	10,7%	1,6	3	0
15. Adecuación de un interno para un módulo concreto dentro de prisión	0	0	2	0	2	4	5,3%	0,8	2	0
16. Riesgo de quebrantamiento, (no reingreso, tercer grado, permisos, etc)	0	5	0	1	0	6	8,0%	1,2	5	0
TOTALES						75	100%			

4.2. Retos y desafíos derivados del desarrollo de la IA.

En relación con la segunda pregunta, sobre cuáles son los retos y desafíos que se perciben prioritarios abordar para avanzar en el proceso de desarrollo de la IA en el ámbito penal,

En relación con el grupo nominal del ámbito judicial, durante la fase de presentación de ideas se identificaron un total de 7 propuestas relacionadas con los retos prioritarios para una implementación adecuada de herramientas algorítmicas en el sistema de justicia penal. Tras el proceso de debate y priorización, se distribuyó un total de 60 puntos entre dichas propuestas.

Las ideas que obtuvieron una mayor puntuación fueron la necesidad de formación a los agentes judiciales para que entiendan el funcionamiento de las herramientas y la transparencia del proceso en la toma de decisiones, ambas con 11 puntos (35,5%), lo que evidencia una preocupación compartida por la comprensibilidad y el control humano sobre los sistemas algorítmicos. Le sigue la propuesta de solucionar sesgos en el desarrollo de las herramientas, que alcanzó un 32,3% (10 puntos), reflejando la importancia otorgada a la equidad y fiabilidad en su diseño.

Otras ideas relevantes fueron la fiabilidad de las pruebas gráficas (vídeos, fotos, etc.), con un 29,0% (9 puntos), y la fiabilidad del resultado de las herramientas, que obtuvo un 25,8% (8 puntos). A pesar de recibir menos puntuación, la preocupación

por evitar la sustitución del profesional judicial (22,6%) también fue valorada, señalando la necesidad de preservar el juicio humano en los procesos de decisión. Finalmente, la idea de abordar el riesgo de confianza tecnológica fue la menos respaldada, con un 12,9% (4 puntos), aunque también fue considerada por varios participantes.

En la Tabla 11, se presenta el listado completo de retos priorizados durante el grupo nominal, junto con la puntuación obtenida, su porcentaje respecto al total y las puntuaciones máxima y mínima alcanzadas. Estos resultados permiten identificar las áreas clave que los operadores consideran fundamentales para avanzar en una implementación ética, comprensible y efectiva de la inteligencia artificial en la justicia penal.

Figura 10.

Puntuación de cada una de las ideas obtenidas en el grupo nominal en el ámbito judicial - Retos y desafíos prioritarios percibidos por los operadores del ámbito judicial.

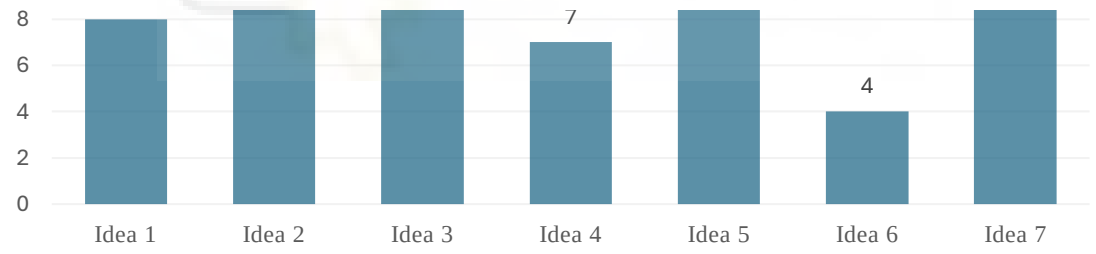


Tabla 11.

Retos y desafíos prioritarios percibidos por los operadores del ámbito judicial.

	P1	P2	P3	P4	PUNTOS	%	M	Máx	Mín
1. Fiabilidad del resultado de las herramientas	0	0	5	3	8	25,8%	2	5	0
2. Transparencia del proceso en la toma de decisiones	4	4	3	0	11	35,5%	2,75	4	0
3. Solucionar sesgos en el desarrollo de las herramientas	3	3	2	2	10	32,3%	2,5	3	2
4. Evitar sustitución del profesional judicial	1	5	0	1	7	22,6%	1,75	5	0
5. Fiabilidad de las pruebas gráficas (videos, fotos, etc.)	0	0	4	5	9	29,0%	2,25	5	0
6. Abordar riesgo de confianza tecnológica.	2	1	1	0	4	12,9%	1	2	0
7. Formación a los agentes judiciales para que entiendan el funcionamiento de las herramientas	5	2	0	4	11	35,5%	2,75	5	0
TOTALES					60	100%			

Respecto a la tercera pregunta, orientada a identificar los principales desafíos o retos que deberían abordarse para avanzar en el proceso de consolidación de herramientas algorítmicas útiles en el ámbito penitenciario, se propusieron un total de 15 ideas a través de los grupos nominales. Tras el debate grupal y el reparto individual de prioridades, se distribuyeron 75 puntos entre las distintas propuestas.

La propuesta que recibió mayor respaldo fue la relativa a la necesidad de transparencia en el funcionamiento de los algoritmos, incluyendo aspectos como qué valoran, cómo se puntúan los ítems, cuáles son los resultados y cómo accede el usuario,, acumulando 16 puntos (21,9%), lo que pone de relieve una fuerte demanda de claridad y trazabilidad en los procesos algorítmicos. En segundo lugar, con 12 puntos (16,4%), se situó la preocupación por los sesgos o discriminaciones hacia colectivos vulnerables, reflejando el interés de los profesionales por evitar efectos adversos derivados del uso de estas herramientas.

Otras dos ideas recibieron un respaldo importante, con 8 puntos cada una (11%): la confrontación científica del uso de algoritmos (especialmente respecto al uso de datos y validación de modelos) y la necesidad de garantizar el respeto a los derechos

fundamentales, mostrando que la dimensión ética y jurídica sigue siendo central en la aceptación tecnológica. Muy cerca, con 7 puntos (9,6%) cada una, se posicionaron tanto la conciencia de la necesidad de cambio o mejora como la confrontación científica del uso de algoritmos (en una formulación complementaria), ambas apuntando a la importancia de una implementación crítica y basada en evidencia.

Otras propuestas con puntuaciones medias incluyeron la conciencia sobre las ventajas económicas y en recursos humanos (5 puntos, 6,8%), la conciencia social del trabajo penitenciario (4 puntos, 5,5%) y la educación digital (3 puntos, 4,1%). En cambio, ideas como la protección de los puestos de trabajo, la unificación de criterios o la mejora general de herramientas obtuvieron poca o ninguna puntuación, lo que podría interpretarse como retos secundarios desde la perspectiva de los participantes.

Figura 11.

Puntuación de cada una de las ideas obtenidas en el grupo nominal en el ámbito penitenciario - Retos y desafíos prioritarios percibidos por los operadores del ámbito penitenciario.

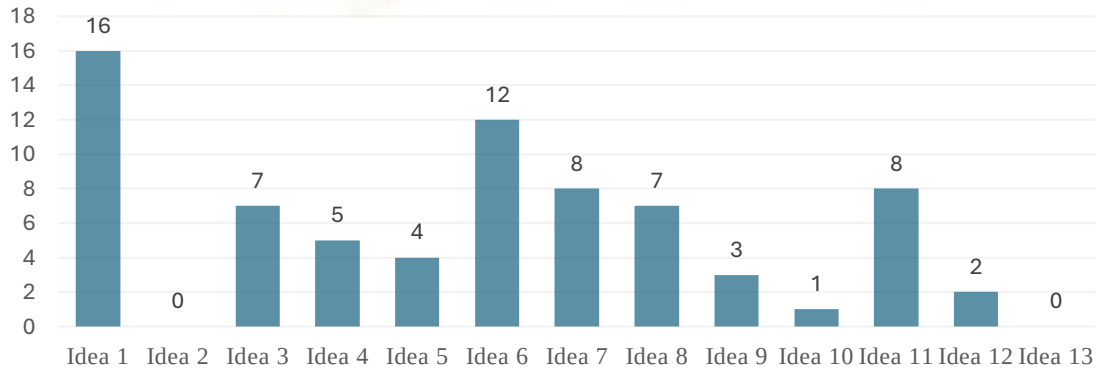


Tabla 12.

Retos y desafíos prioritarios percibidos por los operadores del ámbito penitenciario.

	P1	P2	P3	P4	P5	PUNTOS	%	M	Máx	Mín
1. Transparencia (conocer qué valora, puntuación de los ítems, resultado y acceso del usuario)	5	3	0	5	3	16	21,9%	3,2	5	0
2. Protección y mantenimiento de los puestos de trabajo (factor humano)	0	0	0	0	0	0	0,0%	0	0	0
3. Conciencia de la necesidad de cambio o mejora	0	2	5	0	0	7	9,6%	1,4	5	0
4. Conciencia de las ventajas económica, RRHH.	0	0	0	0	5	5	6,8%	1	5	0
5. Conciencia social del trabajo de los profesionales del ámbito penitenciario	0	0	0	4	0	4	5,5%	0,8	4	0
6. Sesgos o discriminaciones de colectivos vulnerables	4	1	0	3	4	12	16,4%	2,4	4	0
7. Confrontación científica (uso de datos, evaluación algoritmos, etc)	0	4	4	0	0	8	11,0%	1,6	4	0
8. Inversión económica	2	0	3	0	2	7	9,6%	1,4	3	0
9. Educación digital (perdida de miedo al manejo de las herramientas)	0	0	2	0	1	3	4,1%	0,6	2	0
10. Mejora de las herramientas (actualización)	0	0	1	0	0	1	1,4%	0,2	1	0
11.Garantizar el respeto a los derechos fundamentales	3	5	0	0	0	8	11,0%	1,6	5	0
12. Regulación del uso de algorítmicos (dónde y cómo se usan)	1	0	0	1	0	2	2,7%	0,4	1	0
13. Unificación de criterios	0	0	0	0	0	0	0,0%	0	0	0
TOTALES						75	100%			

5. Discusión y conclusiones.

A través del presente estudio se ha podido explorar de forma el impacto actual y el potencial uso de la digitalización y de la inteligencia artificial en la labor de profesionales del ámbito judicial y penitenciario. Los resultados obtenidos coinciden con lo señalado en la literatura comparada al evidenciar que, si bien existe una valoración positiva del potencial de estas herramientas para mejorar la eficiencia operativa, también persisten importantes reservas éticas, técnicas y organizativas sobre su uso (Fine, Le & Miller, 2024; Gerlich, 2023).

En el ámbito judicial, las propuestas que recibieron mayor respaldo se centran fundamentalmente en herramientas de apoyo a la toma de decisiones judiciales y a la gestión documental. En particular, la utilización de instrumentos de predicción del riesgo para la adopción de medidas cautelares fue valorada como la aplicación más relevante (23,3 %), seguida por otras como los asistentes de redacción de resoluciones judiciales, la transcripción automatizada de vistas orales y el expediente digital. Esta tendencia sugiere que los operadores judiciales perciben a la IA como una aliada en procesos que requieren sistematización de información, acceso a datos y automatización de tareas repetitivas, siempre que se preserve la autonomía y responsabilidad de la decisión judicial.

Desde esta perspectiva, se constata una actitud pragmática hacia la digitalización, orientada a mejorar la eficiencia sin comprometer los principios estructurales del proceso penal. No obstante, la escasa puntuación otorgada a herramientas como la síntesis de jurisprudencia o la automatización de entradas al registro penal evidencia que no todas las aplicaciones son consideradas prioritarias, lo que invita a pensar que la percepción de utilidad está vinculada no solo al potencial técnico de la IA, sino también a su encaje con las dinámicas profesionales y al control que los operadores pueden ejercer sobre ella.

En el contexto penitenciario, las propuestas mejor valoradas giran en torno al análisis del riesgo, la prevención de reincidencia y la mejora de la planificación del tratamiento penitenciario. La búsqueda rápida de jurisprudencia y doctrina para la elaboración de informes alcanzó la puntuación más alta (13,3 %), seguida por propuestas como la determinación del riesgo de recaídas, la valoración del riesgo de

reincidencia, el diseño individualizado de planes de tratamiento y la predicción del riesgo de violencia en internos de alto riesgo. Estos resultados reflejan una sensibilidad institucional hacia la gestión del riesgo, la prevención del daño y la mejora de la intervención sobre las personas privadas de libertad.

En ambos ámbitos, los participantes señalaron que la utilidad percibida de la IA no es suficiente para justificar su adopción sin una evaluación crítica de los desafíos que conlleva. En este sentido, los resultados de la última pregunta segundo permiten identificar los principales retos que los profesionales asocian a la implementación de estas herramientas. En el ámbito judicial, la necesidad de formación especializada sobre el funcionamiento de los algoritmos y sus implicaciones éticas y jurídicas fue uno de los elementos más priorizados (35,5 % de los votos), junto con la exigencia de transparencia en el proceso de toma de decisiones algorítmicas (35,5 %) y la corrección de sesgos en los modelos (32,3 %). Estas cifras evidencian que los profesionales valoran la IA en tanto que puedan comprenderla, controlarla y verificar su impacto sobre los principios de legalidad, motivación y equidad.

De forma análoga, en el ámbito penitenciario, el reto más destacado fue la transparencia del funcionamiento de los algoritmos (21,9 %), seguido por la preocupación ante posibles sesgos hacia colectivos vulnerables (16,4 %) y la necesidad de validación científica rigurosa de los modelos utilizados (11 %). Además, los participantes subrayaron la importancia de mantener una supervisión humana constante, así como de evitar la delegación automática de decisiones sensibles, como el diseño de itinerarios de tratamiento o las evaluaciones de riesgo.

Estos hallazgos refuerzan las advertencias formuladas por la literatura científica sobre los riesgos que entraña una implementación acrítica o poco regulada de herramientas algorítmicas en el sistema penal (Galli & Sartor, 2023; UNESCO, 2021). En particular, la opacidad algorítmica, la reproducción de sesgos estructurales y la debilidad de los mecanismos de rendición de cuentas constituyen amenazas reales para los derechos fundamentales si no se acompañan de medidas técnicas, normativas y formativas adecuadas.

La coincidencia entre los dos sectores analizados sugiere que los desafíos de la IA

no dependen exclusivamente del tipo de tarea o del ámbito institucional, sino que responden a condiciones estructurales compartidas: necesidad de comprensión técnica, garantías de transparencia, control humano y alineación ética. En este marco, se refuerza una las hipótesis planteadas en la presente tesis, de que la confianza institucional en la IA no puede darse por sentada, sino que debe construirse progresivamente a través de un enfoque de gobernanza algorítmica centrado en el usuario y en los principios del Estado de derecho.

Estos hallazgos son consistentes con la literatura académica, donde se plantea que las herramientas de IA pueden actuar como mecanismos de apoyo a la toma de decisiones, reduciendo el margen de error humano y aumentando la coherencia jurisprudencial o la seguridad en entornos penitenciarios (Završnik, 2020; Wu, 2019). No obstante, estas oportunidades son percibidas con cautela. Como han advertido numerosos autores, los sistemas algorítmicos pueden actuar como "cajas negras" cuyas decisiones resultan difíciles de auditar o de justificar jurídicamente, lo que entra en conflicto con principios básicos del Estado de derecho como la motivación de las resoluciones judiciales o el derecho a un juicio justo (Richardson, Schultz & Crawford, 2022; Galli & Sartor, 2023).

A partir de estos hallazgos, se pueden extraer diversas conclusiones fundamentales que permiten comprender en profundidad impacto de estas de herramientas de inteligencia artificial en el sistema penal. En primer lugar, la percepción de utilidad de la IA por parte de los operadores jurídicos y penitenciarios está estrechamente vinculada a su carácter instrumental y complementario. Los profesionales valoran positivamente aquellas aplicaciones que contribuyen a optimizar tareas administrativas, sistematizar grandes volúmenes de información o evaluar riesgos con mayor eficiencia, siempre que estas funciones no interfieran con el núcleo de su labor decisional. Es decir, existe una predisposición favorable hacia aquellas herramientas que potencian sus capacidades, pero no hacia las que podrían suplantar el juicio experto.

En segundo lugar, los factores técnicos, éticos y formativos se revelan como determinantes clave para la aceptación de estas tecnologías. La falta de comprensión sobre el funcionamiento de los algoritmos, junto con la opacidad de

los procesos y la posibilidad de sesgos en los resultados, representan barreras significativas para su adopción. Por esta razón, la capacitación continua en competencias digitales, el diseño de algoritmos explicables y la trazabilidad de los resultados emergen como condiciones indispensables para generar confianza institucional y garantizar la rendición de cuentas en contextos de alta exigencia jurídica. Asimismo, los hallazgos del estudio subrayan que la IA debe concebirse, ante todo, como una herramienta de apoyo al juicio humano, y no como un mecanismo que lo sustituya. Las preocupaciones relativas a la deshumanización del proceso judicial, la pérdida de autonomía decisional y el debilitamiento del principio de motivación fueron recurrentes entre los participantes. Esto implica que cualquier desarrollo tecnológico debe orientarse a reforzar el criterio profesional, respetando los márgenes de discrecionalidad y el análisis individualizado de cada caso.

En este marco, la implementación de tecnologías inteligentes en el ámbito penal debería seguir una lógica incremental, reflexiva y participativa. Incorporar la experiencia y conocimiento acumulado de los operadores jurídicos desde las fases iniciales de diseño, prueba y validación de las herramientas resulta clave para fomentar la apropiación tecnológica, facilitar la adaptación institucional y mejorar la eficacia de su aplicación práctica. El diálogo entre desarrolladores, decisores públicos y profesionales del sistema de justicia es una condición estratégica para garantizar que la innovación no se imponga desde una lógica exclusivamente técnica o gerencial.

Los resultados del presente estudio ofrecen, por tanto, una base empírica útil para orientar la elaboración de políticas públicas, planes de formación y marcos regulatorios que promuevan una digitalización del sistema penal alineada con los principios del Estado de derecho y el respeto a los derechos fundamentales. La comprensión de los condicionantes institucionales, técnicos y culturales que influyen en la percepción de la IA resulta esencial para evitar su implementación acrítica o descontextualizada. En definitiva, esta investigación aporta evidencia cualitativa relevante sobre los usos percibidos como más útiles de la inteligencia artificial en la justicia penal, al tiempo que visibiliza los factores que deben ser abordados para asegurar una adopción legítima, eficaz y socialmente aceptada.

CAPITULO 6. ACTITUDES SOCIALES HACIA LA JUSTICIA ALGORÍTMICA: TRES ESTUDIOS SOBRE LA ACEPTACIÓN SOCIAL DEL USO DE HERRAMIENTAS DIGITALES EN EL SISTEMA DE JUSTICIA PENAL

La transformación tecnológica descrita previamente ha puesto de manifiesto que la digitalización, la algoritmización y el uso progresivo de sistemas de inteligencia artificial en el sistema de justicia penal no pueden analizarse sin considerar el papel central del factor humano. Tal como se expuso al contextualizar el Sistema de Justicia Penal 4.0, la digitalización constituye la infraestructura técnica que permite la gestión eficiente de datos y la automatización de procesos (Brennen & Kreiss, 2016; Pereira et al.,2020), mientras que la incorporación de algoritmos y modelos predictivos introduce nuevas formas de análisis, estandarización y apoyo a la decisión (Kleinberg et al., 2018; Monahan & Skeem, 2016). No obstante, como advierten Hildebrandt (2015, 2020) y Susskind (2019), ninguna de estas tecnologías es neutral: su implementación, legitimación y eficacia dependen de las actitudes, creencias y disposiciones de quienes interactúan con ellas.

La literatura internacional en torno a la adopción tecnológica refuerza la relevancia de comprender el papel del factor humano dentro de procesos de transformación institucional como el que atraviesa el sistema de justicia penal. Uno de los aportes más influyentes en este campo es el Technology Acceptance Model (TAM), propuesto por Davis (1989), que demostró que la disposición de una persona a utilizar una tecnología depende fundamentalmente de la utilidad percibida y de la facilidad de uso. Sin embargo, investigaciones posteriores observaron que estos dos factores, aunque esenciales, no bastan para explicar la aceptación en contextos complejos u organizacionalmente sensibles. Esto dio lugar a modelos ampliados, entre ellos el Unified Theory of Acceptance and Use of Technology (UTAUT), formulado por Venkatesh, Morris, Davis y Davis (2003), que integra variables como la influencia social y las condiciones facilitadoras, permitiendo capturar de manera más completa la interacción entre individuo, organización y tecnología.

Posteriormente, nuevos estudios sobre aceptación tecnológica han destacado que los modelos clásicos requieren adaptaciones al mismo tiempo que la tecnología se

actualiza. Sun y Zhang (2006) destacan que diversos factores individuales (de naturaleza emocional y cognitiva) y contextuales pueden modular significativamente la actitud y la aceptación del usuario hacia la tecnología, mientras que Benbasat y Barki (2007) señalan que los modelos tradicionales de aceptación tecnológica, como el TAM, no logran anticipar resistencias derivadas de expectativas no cubiertas, falta de transparencia o cambios en las dinámicas profesionales, que influyen de manera significativa en la adopción de nuevas tecnologías.

Este debate ha ganado especial relevancia con la expansión de la automatización y de la inteligencia artificial. Estudios recientes han evidenciado que variables como la confianza, la percepción de riesgo, la ansiedad tecnológica o las creencias sobre la imparcialidad del sistema ejercen un papel cada vez más central en las decisiones de uso (Parasuraman & Colby, 2015; Wirtz et al., 2019). En sectores como el judicial, donde la tecnología no solo facilita tareas sino que influye en decisiones sensibles, estos factores adquieren un peso determinante.

Asimismo, y entroncando con las diferentes variables que configuran el proceso de aceptación tecnológica, surge una dimensión que adquiere especial relevancia con la incorporación de sistemas de inteligencia artificial y herramientas autónomas: el propio modelo de interacción entre el operador humano y la tecnología. Más allá de los factores psicológicos, organizativos o contextuales que tradicionalmente explican la aceptación tecnológica, una cuestión central actualmente es qué configuración de la relación humano-máquina resulta deseable, aceptable y compatible con las exigencias del sistema de justicia penal. La elección entre modelos con intervención humana significativa (Human-in-the-Loop) o modelos con una mayor autonomía algorítmica (Human-out-of-the-Loop) constituye una vertiente reciente del debate, estrechamente vinculada al despliegue de sistemas de IA avanzada, y que podría actuar como un factor en la aceptación de estas tecnologías.

Sobre esta base, los tres estudios empíricos incluidos en este capítulo responden a una necesidad detectada en el marco teórico: comprender la aceptación tecnológica no como un fenómeno aislado, sino como un proceso multidimensional que integra múltiples variables, y que desde el ámbito del derecho apenas ha sido explorado

empíricamente. Lo que se observa en la literatura revisada es que este tipo de variables resultan decisivas para garantizar una adopción tecnológica adecuada, progresiva y no irruptora, especialmente en sistemas institucionales tan sensibles como el de justicia penal. Sin embargo, a pesar de su relevancia, el derecho se ha centrado en desarrollos normativos sin analizar desde una perspectiva empírica cómo estos factores influyen en la aceptación de estas herramientas.

El primer estudio se plantea para analizar, desde una perspectiva empírica, qué modelo de interacción entre operadores jurídicos y sistemas algorítmicos resulta más aceptable en el sistema de justicia penal, a la luz del marco teórico previamente desarrollado. En este contexto, comparar modelos de decisión supervisados (*Human-in-the-Loop*) y no supervisados (*Human-out-of-the-Loop*) permite examinar los principios que deberían guiar la integración de estas tecnologías. Además, el estudio incorpora la perspectiva diferenciada de la ciudadanía y de los profesionales como un elemento necesario para comprender cómo se articula social e institucionalmente esa relación humano-IA.

El segundo estudio examina de forma sistemática las principales variables de los modelos de aceptación tecnológica evaluando su relevancia y aplicabilidad en el contexto penal. Este análisis comparado es imprescindible porque las organizaciones de justicia presentan características muy distintas a otros sectores donde estos modelos fueron originalmente desarrollados: estructuras jerárquicas rígidas, necesidad de garantías procedimentales, alto escrutinio externo y un fuerte componente profesional en la toma de decisiones. Por ello, este estudio permite identificar qué variables mantienen su poder explicativo (Davis, 1989; Venkatesh et al., 2003) y qué dimensiones emergentes deben incorporarse para ajustar los modelos a la realidad del sistema penal actual.

Finalmente, el tercer estudio, profundiza específicamente en cómo la ciudadanía evalúa el uso de herramientas algorítmicas cuando estas intervienen en decisiones judiciales y penitenciarias, atendiendo no solo al grado de autonomía tecnológica, sino también a la adecuación y proporcionalidad de las sanciones propuestas. Si los dos estudios anteriores se centraban, por un lado, qué modelo de interacción humano–IA resulta más aceptable y, por otro, qué variables explican mejor la

adopción tecnológica en el sistema penal, este tercer estudio amplía el marco al examinar cómo perciben los ciudadanos las decisiones mediadas por sistemas algorítmicos. A través de escenarios experimentales que combinan distintos niveles de intervención humana y diferentes tipos de respuesta sancionadora, el estudio explora qué configuraciones generan mayor aceptación social.

Consideraciones previas a los estudios.

Antes de presentar cada uno de los estudios, es necesario aclarar varios aspectos metodológicos que son comunes a los tres y que, por tanto, se exponen de manera conjunta para evitar reiteraciones innecesarias. Aunque el cuestionario utilizado se estructuró en bloques diferenciados, cada uno diseñado para abordar específicamente las variables y objetivos de cada estudio, la aplicación del instrumento se realizó de manera conjunta.

Como consecuencia, la muestra empleada es la misma para los tres análisis, y también lo es la base de datos de la que se construyen las variables dependientes y las diferentes medidas de aceptación. Por ello, en este apartado se detalla de manera integrada la información relativa a la muestra, el instrumento empleado y el proceso de codificación de las variables, de forma que cada estudio pueda centrarse en sus propios objetivos, hipótesis y análisis sin duplicar explicaciones metodológicas ya expuestas.

1. **Instrumento.**

El instrumento utilizado en los tres estudios fue elaborado específicamente para esta investigación, a partir de una revisión exhaustiva de los principales modelos de aceptación tecnológica y adaptando sus variables a las particularidades del sistema de justicia penal. Su diseño se apoyó en teorías como la Teoría de la Acción Razonada (Fishbein & Ajzen, 1975), la Teoría del Comportamiento Planeado (Ajzen, 1991), el Modelo de Aceptación de la Tecnología (TAM) (Davis, 1989) y sus ampliaciones dentro de la UTAUT (Venkatesh et al., 2003), así como en aportes contemporáneos del Modelo de Difusión de Innovaciones (Rogers, 2003) y del Technology Readiness Index (Parasuraman, 2000). A partir de este marco conceptual, se seleccionaron aquellos componentes más relevantes para evaluar la aceptación de tecnologías en contextos institucionales sensibles como el sistema penal. En este sentido, el instrumento se desarrolló de manera secuencial.

En primer lugar, se elaboró un cuestionario inicial basado en las principales teorías de aceptación tecnológica, incorporando los constructos procedentes de modelos como la Teoría de la Acción Razonada, la Teoría del Comportamiento Planeado, el

TAM, la UTAUT, el Modelo de Difusión de Innovaciones y el Technology Readiness Index. Este primer instrumento incluía dos bloques, ítems de actitudes y creencias vinculados a estos modelos teóricos, diseñados para medir predisposiciones generales hacia la adopción de tecnología; e ítems de aceptación tecnológica vinculados a tareas concretas del sistema de justicia penal, diferenciando entre herramientas autónomas (HOTL) y supervisadas (HITL).[10]

Con este cuestionario inicial se llevó a cabo una prueba piloto con 77 participantes, cuyo análisis permitió evaluar el funcionamiento de los ítems. En particular, se observó que el bloque destinado a medir la aceptación tecnológica contenía un número elevado de ítems, muchos de los cuales tendían a concentrarse en valores intermedios. A partir de los resultados del piloto se decidió reducir el número de ítems del bloque de aceptación, reteniendo únicamente aquellos que mostraban patrones de respuesta más informativos, es decir, los que presentaban mayores niveles de aceptación o rechazo y, por tanto, menor acumulación en valores centrales. En concreto:

1. Primer cuestionario:
    - Incluyó un apartado específico para evaluar la aceptación de herramientas algorítmicas en usos del ámbito del derecho penal. Contenía 8 ítems sobre herramientas autónomas y 8 ítems sobre herramientas supervisadas (16 en total del apartado.
    - Además, se añadieron 16 ítems adicionales, también divididos en herramientas autónomas y supervisadas, para evaluar su aceptación de los usos en otros ámbitos del derecho (32 ítems en el total del

[10] El cuestionario se encuentra publicado en los anexos del artículo Pérez Domínguez, S., & Simón Castellano, P. (2023). Attitudes and perceptions regarding algorithmic judicial judgement: barriers to innovation in the judicial system?. IDP: Revista de Internet, Derecho y Política, (39), 6. https://doi.org/10.7238/idp.v0i39.417206

apartado).

2. Segundo cuestionario (véase Anexo 2 – bloque 4):
    - Consta de 22 ítems, organizados en dos bloques:
        - 11 ítems de usos en el sistema de justicia penal para evaluar la aceptación de herramientas autónomas (*Human-out.of-the-Loop* – HOTL).
        - 11 ítems de usos en el sistema de justicia penal para evaluar la aceptación de herramientas supervisadas (*Human-in-the-Loop* – HITL).

Una vez seleccionados los ítems más adecuados, se construyó el cuestionario definitivo. En él se mantuvo íntegramente el bloque de actitudes derivado de los modelos teóricos y se incorporó el bloque reducido de aceptación tecnológica, ya adaptado a los dos modelos de interacción (autónomo y supervisado). Sobre esta estructura optimizada se añadieron los escenarios experimentales, diseñados específicamente para el estudio principal y orientados a manipular condiciones de autonomía y proporcionalidad en la toma de decisiones. Para la medición de las variables se empleó en ambos cuestionarios una escala tipo Likert de cinco puntos, donde 1 indicaba "totalmente en desacuerdo" y 5 "totalmente de acuerdo".

La versión completa del cuestionario definitivo puede consultarse en el Anexo 2.

2. Procedimiento.

Como se ha comentado anteriormente, el estudio se desarrolló en varias etapas. En primer lugar, se elaboró el cuestionario y se llevó a cabo una prueba piloto que permitió revisar y ajustar los ítems. Tras esta fase, la versión final de la encuesta fue administrada por una empresa especializada en muestreo. Durante la recopilación de datos, cuando se alcanzaron los primeros 100 cuestionarios completos, se realizó una comprobación intermedia para asegurarse de que todos los ítems se estaban registrando correctamente y que no había errores de diseño, redacción o captura. Al confirmarse que no existían incidencias, se continuó con la recogida de datos hasta

completar la muestra total de 1.100 participantes. La participación tuvo una duración aproximada de 25 minutos y el estudio fue aprobado por el Comité de Ética e Investigación de la Universidad Miguel Hernández de Elche (UMH).

Una vez finalizada la fase de campo, la empresa entregó los datos en formato Excel. Estos fueron transformados a una base de datos en SPSS y se procedió a la codificación de todas las variables. Con la base ya preparada, se construyeron las variables dependientes de aceptación social autónoma y aceptación social supervisada, a partir de los 11 ítems vinculados a las diferentes tareas para las que se evaluaba el uso de inteligencia artificial.

Para evaluar la consistencia interna de ambas escalas, se calculó el alfa de Cronbach. El valor obtenido para la escala de herramientas autónomas fue de 0.891 y para la escala de herramientas supervisadas, de 0.918, lo que indica una fiabilidad elevada y una buena coherencia interna entre los ítems. Por tanto, para construir ambas variables dependientes los ítems fueron sumados y calculado la media.

3. Muestra.

En la prueba piloto, la muestra estuvo compuesta por 77 participantes, de los cuales el 51,9% (N = 50) eran mujeres y el 48,1% (N = 37) eran hombres. La edad media fue de 38,8 años (DE = 11,89). En cuanto al nivel educativo, el 84,4% (N = 65) de los participantes poseía un título universitario. Además, se analizó el conocimiento y la experiencia en el sistema de justicia penal y en herramientas automatizadas. Se encontró que el 47,4% de los participantes tenía un grado en derecho, el 17,1% había ejercido la abogacía, otro 17,1% trabajaba como consultor en servicios jurídicos y el 7,9% era magistrado o juez. En el ámbito tecnológico, el 5,3% había trabajado o trabajaba actualmente en empresas de LegalTech, el 7,9% se desempeñaba como técnico en análisis de datos, el 6,6% había participado en el desarrollo de soluciones tecnológicas y el 10,5% tenía conocimientos prácticos sobre inteligencia artificial. Finalmente, el 34,2% de los participantes no tenía conocimientos en ninguna de las áreas mencionadas.

En la segunda fase, se amplió la muestra a 1.100 participantes de la población española, con una distribución de 52,3% (N = 575) mujeres y 47,7% (N = 525)

hombres. La edad media en esta fase fue de 43,2 años (DE = 12,04). Respecto al nivel educativo, el 62,9% (N = 692) contaba con un título universitario, mientras que el 28,3% (N = 311) había completado estudios de bachillerato o formación profesional. Solo el 8,8% de la muestra no había alcanzado el nivel de educación secundaria, de los cuales el 1,8% tenía educación primaria y el 7,0% contaba con estudios secundarios. En relación con el conocimiento sobre el sistema de justicia penal o el desarrollo y análisis de herramientas automatizadas, el 25,7% de los participantes tenía algún conocimiento en el ámbito jurídico, mientras que el 5,5% poseía conocimientos sobre sistemas de inteligencia artificial. Por otro lado, el 67,2% (N = 739) de los participantes formaba parte de la población general sin conocimientos en estas áreas.

Estudio 1. Actitudes de la ciudadanía y los profesionales relativas a la justicia algorítmica: ¿barreras para la innovación en el sistema de justicia penal?.

1. Justificación.

La incorporación de sistemas algorítmicos en el sistema de justicia penal ha abierto un debate profundo sobre el papel que debe desempeñar el ser humano en los procesos de toma de decisiones y sobre las condiciones bajo las cuales estas tecnologías pueden considerarse legítimas, fiables y socialmente aceptables. Como se expone en el marco teórico, la digitalización de la justicia no se limita a introducir herramientas técnicas, sino que implica repensar la arquitectura completa del proceso judicial y, en particular, redefinir la relación entre los operadores jurídicos y los sistemas de inteligencia artificial. En este contexto, los modelos de interacción humano-IA, especialmente los modelos supervisados (Human-in-the-Loop, HITL) y los no supervisados (Human-out-of-the-Loop, HOTL), se han convertido en elementos clave para valorar hasta qué punto la automatización puede integrarse en decisiones que afectan directamente a derechos fundamentales como la libertad, la igualdad o el acceso a un juicio justo.

La literatura científica y las recomendaciones internacionales coinciden en señalar que la intervención humana significativa constituye una condición esencial para preservar la legitimidad democrática de las decisiones judiciales automatizadas (UNESCO, 2021; European Commission, 2024). Sin embargo, también muestran que no todas las tareas requieren el mismo grado de supervisión, y que la aceptación social de estas herramientas no depende únicamente del diseño técnico del sistema, sino también de la percepción que tanto la ciudadanía como los profesionales del derecho tienen sobre su fiabilidad, imparcialidad y propósito.

El presente estudio se plantea precisamente con este objetivo: evaluar qué modelo de interacción entre operadores jurídicos y sistemas algorítmicos genera mayor aceptación en el sistema de justicia penal. Este estudio aborda, por tanto, dos dimensiones fundamentales. La primera es el grado de autonomía tecnológica, comparando las percepciones sociales ante escenarios en los que la IA actúa como herramienta de apoyo (HITL) frente a aquellos en los que opera de forma autónoma

(HOTL). Además, el estudio incorpora la comparación entre población general y profesionales del derecho, entendiendo que sus percepciones no solo pueden diferir, sino que esa diferencia constituye información relevante para anticipar resistencias y oportunidades en la implementación real de estos sistemas. Mientras que la ciudadanía actúa como destinataria final de las decisiones algorítmicas, los operadores jurídicos representan a quienes deberán integrar, supervisar o delegar parte de su trabajo en estas herramientas. En conjunto, este estudio se justifica por la necesidad de generar evidencia que permita orientar el diseño de políticas públicas, protocolos tecnológicos y modelos de supervisión que respeten las expectativas sociales.

2. Objetivos.

El objetivo general del presente estudio, tanto en la prueba piloto como en el estudio final, fue analizar las actitudes y el nivel de aceptación de la población general y de los profesionales del derecho respecto al uso de herramientas algorítmicas, tanto autónomas como supervisadas, en el análisis judicial. No obstante, los objetivos específicos y las hipótesis fueron ajustados a lo largo del proceso para garantizar una mayor adecuación al desarrollo del estudio.

2.1. Objetivos estudio piloto.

Para el primer análisis se incluyeron los siguientes objetivos específicos:

OE1. Analizar la aceptación de la población general y de los profesionales del derecho respecto a la implementación de herramientas algorítmicas para el análisis judicial.

OE2. Comparar los niveles de aceptación de la población general y de los profesionales del derecho.

OE3. Analizar si la aceptación de la inclusión de herramientas algorítmicas para el análisis judicial depende del nivel de automatización.

Así, se plantearon las siguientes hipótesis se proponen para alcanzar los objetivos mencionados anteriormente:

H1. Los profesionales del derecho consideran que la IA no puede reemplazar su papel en la administración de justicia.

H2. La aceptación por parte de la población general del uso de herramientas algorítmicas para el análisis judicial es mayor en comparación con la de los profesionales del derecho.

H3. Las tareas que requieren un procesamiento más simple tendrán una mayor aceptación entre la población.

H4. La inclusión de herramientas algorítmicas tendrá una mayor aceptación si el

proceso es supervisado por un factor humano.

2.2. Objetivos estudio definitivo.

Por otro lado, para el presente estudio se plantearon los siguientes objetivos específicos.

OE1. Evaluar el nivel de aceptación de las herramientas automatizadas en el sistema judicial por parte de la población general en España, con especial atención a la percepción sobre la necesidad de intervención humana en los procesos de toma de decisiones.

OE2. Identificar las tareas judiciales automatizadas que cuentan con mayor aceptación social, diferenciando entre herramientas autónomas y aquellas que operan bajo supervisión humana.

OE3. Analizar las actitudes de la población general y de los profesionales del derecho respecto a la integración de herramientas algorítmicas en las tareas judiciales.

OE4. Examinar cómo la supervisión humana afecta la aceptación social del uso de herramientas algorítmicas en los procesos judiciales.

Y las siguientes hipótesis:

H1. La población general muestra un mayor nivel de aceptación hacia el uso de herramientas algorítmicas en el sistema de justicia penal que los profesionales del derecho.

H2. La presencia de intervención humana en los procesos judiciales incrementa el nivel de aceptación de las herramientas algorítmicas.

H3. Las personas más jóvenes presentan un nivel de aceptación más alto hacia el uso de herramientas algorítmicas en el sistema judicial que las personas de mayor edad.

3. Metodología.

3.1. Muestra.

Tal como se detalla en las consideraciones previas, este estudio utiliza la misma muestra de 1.100 participantes, correspondiente a la segunda fase de la investigación. Por ello, para evitar repeticiones, no se describe aquí de nuevo su composición, que puede consultarse en dicha sección. En este estudio participaron la totalidad de los encuestados, ya que todos respondieron al bloque de aceptación de herramientas algorítmicas autónomas y supervisadas.

3.2. Variables e instrumento.

El estudio emplea específicamente los ítems destinados a evaluar la aceptación de herramientas algorítmicas en el ámbito judicial, diferenciando entre sistemas completamente autónomos (HOTL) y herramientas con supervisión humana (HITL).

La variable dependiente se construyó mediante dos índices: aceptación de herramientas autónomas y aceptación de herramientas supervisadas, cada uno calculado como la media de los 11 ítems correspondientes. Las variables sociodemográficas (sexo, edad, nivel educativo y experiencia jurídica o tecnológica) se usan como variables independientes.

A continuación, se presentan en una tabla todas las variables incluidas en el estudio:

Tabla 13.

Resumen de las variables incluidas en el estudio 1.

Categoría	Variable	Descripción
Sociodemográficas	Sexo	Sexo del participante
	Edad	Edad del participante
	Nivel educativo	Máximo nivel de estudios alcanzado
Dependiente	Aceptación de herramientas algorítmicas autónomas y supervisadas	Grado de aceptación de las herramientas algorítmicas
Independientes	Modo de uso de herramientas algorítmicas	Modelos de uso HOTL o HITL.
	Conocimiento legal	Formación o experiencia previa en el ámbito legal.

3.3. Análisis de datos.

En cuanto al análisis de los datos, En primer lugar, se realizaron análisis descriptivos con el fin de obtener una visión general del patrón de respuestas de la muestra. A continuación, se comprobaron los supuestos de normalidad, independencia y homogeneidad. Dado que estos no se cumplían, se emplearon pruebas no paramétricas en los análisis posteriores.

- Para comparar la aceptación de herramientas autónomas y supervisadas dentro de los mismos participantes, se utilizó la prueba de Wilcoxon. Esta prueba fue elegida porque es particularmente adecuada para comparar dos muestras relacionadas cuando los datos son ordinales o cuando no se cumplen los supuestos de normalidad requeridos por las pruebas paramétricas.
- Para examinar las diferencias entre grupos (profesionales del ámbito jurídico frente a población general), se aplicó la U de Mann–Whitney.
- Para estudiar la relación entre variables sociodemográficas (edad, sexo, formación) o experiencia previa y los niveles de aceptación, se utilizaron correlaciones de Spearman. Este coeficiente estima tanto la intensidad como la dirección de la relación entre dos variables, indicando si evolucionan conjuntamente (relación directa) o en direcciones opuestas (relación inversa). Sus valores oscilan entre –1 y +1, de manera que valores próximos a +1 reflejan asociaciones directas muy fuertes, valores cercanos a –1 evidencian asociaciones inversas muy fuertes y valores próximos a 0 sugieren ausencia de relación. Según los criterios de Dancey y Reidy (2017), correlaciones entre 0.00 y 0.19 se consideran muy débiles, entre 0.20 y 0.39 débiles, entre 0.40 y 0.59 moderadas, entre 0.60 y 0.79 fuertes y entre 0.80 y 1.00 muy fuertes. Esta técnica resulta adecuada para explorar las asociaciones entre las actitudes analizadas como objetividad, imparcialidad, justicia, sesgos o necesidad de regulación y los niveles de aceptación de la inteligencia artificial en escenarios diferenciados, tanto en su aplicación autónoma como bajo supervisión humana.

Finalmente, se valoró la posibilidad de estimar un modelo de regresión ordinal con

el fin de explorar en mayor profundidad la influencia conjunta de las variables sociodemográficas y experienciales en la aceptación de herramientas algorítmicas. No obstante, tras examinar los resultados de las correlaciones, se consideró que no había suficientes evidencias para desarrollar un modelo predictivo.

4. Resultados.

4.1. Resultados del primer estudio piloto.

Dado que los resultados completos del estudio piloto se encuentran publicados[11], a continuación, se presentan de forma resumida los principales hallazgos, que sirvieron de base para orientar el desarrollo del estudio ampliado.

En términos generales, los datos mostraron una aceptación moderada del uso de herramientas algorítmicas en la administración de justicia. Una proporción relevante de participantes consideró que estas tecnologías podrían resultar útiles en determinadas tareas, especialmente en funciones de apoyo técnico o administrativo, aunque también se observaron reservas significativas ante su aplicación en procesos de decisión judicial.

Los niveles de aceptación variaron en función del tipo de tarea atribuida al sistema algorítmico. Las actividades de carácter más técnico o rutinario, como el uso de *chatbots* para consultas frecuentes o el análisis de datos en expedientes judiciales, registraron niveles de acuerdo más altos. Por el contrario, las funciones decisorias, como la determinación del fallo o la concesión del tercer grado penitenciario, obtuvieron una menor aceptación, especialmente cuando se planteaba un uso autónomo de la herramienta sin intervención humana.

El análisis comparativo entre perfiles profesionales evidenció diferencias estadísticamente significativas. Los participantes con formación en inteligencia

[11] Pérez Domínguez, S., & Simón Castellano, P. (2023). Attitudes and perceptions regarding algorithmic judicial judgement: barriers to innovation in the judicial system?. IDP: Revista de Internet, Derecho y Política, (39), 6. https://doi.org/10.7238/idp.v0i39.417206

artificial o análisis de datos mostraron una mayor aceptación del uso de estas herramientas que los profesionales del ámbito jurídico. Dentro de este último grupo, la aceptación fue más baja tanto en las tareas automatizadas como en las supervisadas, lo que sugiere una actitud más conservadora respecto a la introducción de sistemas algorítmicos en la práctica judicial.

Los resultados obtenidos permitieron identificar patrones iniciales de aceptación distintos en función del tipo de tarea y la formación de los participantes, lo que llevó a ajustar el diseño del estudio principal, ampliar el tamaño de la muestra y analizar las diferencias entre la población general y los profesionales del Derecho.

4.2. Resultados del estudio ampliado.

4.2.1. Aceptación social del uso de herramientas algorítmicas.

A continuación, se presentan los resultados obtenidos respecto al uso de herramientas algorítmicas automatizadas e informacionales en los procesos judiciales. En primer lugar, la Tabla 14 resume las respuestas de los participantes en relación con diversas aplicaciones de la inteligencia artificial en el ámbito legal, evaluando su nivel de acuerdo o desacuerdo con la implementación de estas tecnologías de manera totalmente autónoma. Por otro lado, la Tabla 15 muestra el nivel de acuerdo con el uso de la inteligencia artificial en el ámbito legal cuando esta funciona como un accesorio para el profesional.

Tabla 14.

Uso de herramientas algorítmicas automatizadas (Human-Out-of-the-Loop).

	Totalmente en desacuerdo		En desacuerdo		Ni de acuerdo ni en desacuerdo		De acuerdo		Totalmente de acuerdo	
	n	%	n	%	n	%	n	%	n	%
Predicción del nivel de riesgo de reincidencia.	200	18,2	193	17,5	375	34,1	225	20,5	107	9,7
Decisión sobre el fondo o determinación del fallo.	319	29	197	17,9	386	35,1	143	13	55	5
Decisión de concesión del tercer grado	313	28,5	236	21,5	359	32,6	139	12,6	53	4,8
Chatbots para responder preguntas comunes y programar citas.	123	11,2	108	9,8	306	27,8	300	27,3	263	23,9
Averiguación del patrimonio o del domicilio.	87	7,9	81	7,4	302	27,5	319	29	311	28,3
Adopción de la decisión judicial en el ámbito laboral.	267	24,3	199	18,1	394	35,8	168	15,3	72	6,5
Adopción de la decisión judicial en la jurisdicción civil (disputas en temas como propiedad, contratos, familia y herencias).	262	23,8	179	16,3	402	36,5	179	16,3	78	7,1
Monitorear el cumplimiento de las obligaciones contractuales.	98	8,9	102	9,3	371	33,7	310	28,2	219	19,9
Análisis de tratados y acuerdos internacionales (revisión del texto e identificación de compromisos específicos).	123	11,2	122	11,1	411	37,4	310	28,2	269	19,9
Determinación de la cuantía de la indemnización en accidentes.	136	12,4	127	11,5	366	33,3	296	26,9	175	15,9
Determinar la veracidad de una denuncia.	252	22,9	163	14,8	349	31,7	224	20,4	112	10,2
Aceptación general	76	6,9	228	20,7	569	51,7	194	17,6	33	3,0

Tabla 15.

Uso de herramientas algorítmicas con supervisión (Human-in-the-loop).

	Totalmente en desacuerdo		En desacuerdo		Ni de acuerdo ni en desacuerdo		De acuerdo		Totalmente de acuerdo	
	n	%	n	%	n	%	n	%	n	%
Predicción del nivel de riesgo de reincidencia	172	15,6	196	17,8	360	32,7	237	21,5	135	12,3
Decisión sobre el fondo o determinación del fallo	246	22,4	193	17,5	411	37,4	153	13,9	97	8,8
Decisión de concesión del tercer grado	264	24	233	21,2	365	33,2	146	13,3	92	8,4
Chatbots para responder preguntas comunes y programar citas	92	8,4	104	9,5	329	29,9	291	26,5	284	25,8
Averiguación del patrimonio o del domicilio	81	7,4	95	8,6	322	29,3	295	26,8	307	27,9
Adopción de la decisión judicial en el ámbito laboral.	219	19,9	195	17,7	385	35	195	17,7	106	9,6
Adopción de la decisión judicial en la jurisdicción civil (disputas en temas como propiedad, contratos, familia y herencias)	219	19,9	209	19	370	33,6	195	17,7	107	9,7
Monitorear el cumplimiento de las obligaciones contractuales	100	9,1	118	10,7	362	32,9	294	26,7	226	20,5
Análisis de tratados y acuerdos internacionales (revisión del texto e identificación de compromisos específicos)	97	8,8	133	12,1	385	35	295	26,8	190	17,3
Determinación de la cuantía de la indemnización en accidentes.	115	10,5	138	12,5	355	32,3	289	26,3	203	18,5
Determinar la veracidad de una denuncia.	195	17,7	175	15,9	355	32,3	240	21,8	135	12,3
Aceptación general	70	6,4	209	19	533	48,5	222	20,2	66	6,0

4.2.2. Aceptación basada en herramientas algorítmicas autónomas vs. supervisadas.

El análisis de las Tablas 14 y 15 revela diferencias en la aceptación de las herramientas algorítmicas, comparando una que opera de manera autónoma y toma decisiones de forma independiente con otra que actúa como un asistente informacional, donde una persona toma la decisión final. Para determinar si estas diferencias observadas son estadísticamente significativas, se utilizó la prueba de Wilcoxon.

Tabla 16.

Diferencias en la aceptación de herramientas con y sin supervisión humana.

Resumen de prueba de rangos con signo de Wilcoxon para muestras relacionadas	
N total	1100
Estadístico de prueba	225625,500
Error estándar	7570,850
Estadístico de prueba estandarizado	3,968
Sig. asintótica (prueba bilateral)	<,001

Los valores presentados en la Tabla 16 indican una diferencia significativa entre las condiciones evaluadas. Además, el tamaño del efecto, medido a través del coeficiente *r* de Rosenthal, fue aproximadamente 0.120, lo que sugiere un efecto pequeño. Estos resultados respaldan la hipótesis. La dirección de la diferencia indica que la aceptación de la IA es mayor cuando existe supervisión humana.

4.2.3. Aceptación basada en la formación (Población general vs. Profesionales del ámbito legal)

Por otro lado, se comparó los niveles de aceptación del uso de IA autónoma e informacional para determinar si estos dependían de la formación de los participantes; es decir, comparamos los niveles de aceptación entre la población general y los profesionales del ámbito legal.

Tabla 17.

Comparación de la aceptación del uso de herramientas entre profesionales del derecho y población general.

	Aceptación de herramientas algorítmicas autónoma – (HOTL)	**Aceptación de herramientas algorítmicas supervisadas (HITL).**
U de Mann-Whitney	92061,000	95577,000
Z	-2,965	-2,132
Sig. asintótica(bilateral)	,003	,033
r de Rosenthal	−0.093	−0.067

a. Variable de agrupación: Formación

Los resultados en la Tabla 17 muestran un valor de *Z* negativo en ambos casos. Esto sugiere que el grupo con el rango promedio más bajo, en este caso, los profesionales del ámbito legal, presenta una menor aceptación de la inteligencia artificial en comparación con la población general. Además, ambos valores de *p* están por debajo de 0.05, lo que indica que las diferencias en la aceptación entre los dos grupos son estadísticamente significativas.

Esto confirma que la aceptación de las herramientas algorítmicas varía entre la población general y los profesionales del ámbito legal, siendo mayor en la población general. Para cuantificar la magnitud de esta diferencia, se calculó el tamaño del efecto utilizando el coeficiente *r* de Rosenthal. Aunque estos valores fueron pequeños y negativos, resultaron estadísticamente significativos, lo que refuerza la existencia de diferencias en la aceptación de la IA entre los grupos estudiados. Un tamaño del efecto pequeño a moderado indica que, si bien la diferencia es evidente, no es extremadamente pronunciada, aunque sigue siendo lo suficientemente notable como para ser considerada.

4.2.4. Correlación de Spearman.

La Tabla 18 muestra los coeficientes de correlación de Spearman entre las variables predictoras (edad, nivel educativo, género, familiaridad con los algoritmos judiciales y formación) y las variables dependientes de aceptación. Este análisis permite

observar la fuerza y dirección de las relaciones entre las variables e identificar aquellas que son estadísticamente significativas.

A partir del análisis de correlaciones, se observa que la aceptación de las herramientas algorítmicas totalmente autónomas (HOTL) muestra una relación significativa, aunque débil, únicamente con el nivel de familiarización con los algoritmos, lo que indica que a mayor conocimiento tecnológico, ligeramente mayor es la predisposición para aceptar sistemas de decisión automatizados sin supervisión humana. En contraste, la aceptación de las herramientas algorítmicas supervisadas (HITL) presenta una correlación muy elevada con la aceptación de las herramientas autónomas, evidenciando que ambos modelos comparten un mismo patrón subyacente de confianza en la tecnología. Asimismo, el sexo muestra una relación significativa pero de baja magnitud con la aceptación del modelo supervisado, mientras que el resto de las variables sociodemográficas no presentan asociaciones relevantes con ninguna de las dos medidas. En conjunto, los resultados sugieren que la aceptación ciudadana de herramientas algorítmicas, tanto autónomas como supervisadas, depende fundamentalmente de la familiaridad tecnológica y de una predisposición general hacia este tipo de sistemas, más que de factores demográficos.

Tabla 18.

Matriz de correlación de Spearman de las variables relacionadas con la aceptación de herramientas algorítmicas.

Variables	1	2	3	4	5	6	7
1. Edad	-						
2. Nivel educativo	-.109**	-					
3. Sexo	-.144**	.109**	-				
4. Familiarización con los algoritmos.	.080*	-.197**	.040	-			
5. Formación	-.013	-.388**	-.044	.228**	-		
6. Aceptación de herramientas algorimicas autónomas (HOTL)	-.033	-.028	-.052	.087**	.060	-	
7. Aceptación de herramientas algorítmicas supervisadas (HITL).	-.016	.032	-.062*	.034	-.044	.725**	-

Nota. *p < .05, **p < .01. Los coeficientes de correlación fueron calculados utilizando el método de Spearman.

Dado que las dos medidas de aceptación están fuertemente correlacionadas entre sí y su variabilidad depende principalmente de un factor común (actitud general hacia IA), un modelo predictivo aportaría poco valor interpretativo por lo que en este caso se optó por no desarrollar el modelo.

5. Discusión y conclusiones.

La digitalización de la sociedad y, específicamente, de los sistemas judiciales, ha generado un debate continuo sobre el papel y las limitaciones de las herramientas algorítmicas en el sector de la justicia. La inteligencia artificial se presenta como un recurso potencialmente beneficioso en áreas que van desde la gestión administrativa hasta la predicción del riesgo de reincidencia. Sin embargo, su implementación también genera intensas discusiones sobre los valores éticos, técnicos y legales que deben guiar su uso, especialmente cuando se considera la posibilidad de una toma de decisiones autónoma. Este estudio contribuye a esta discusión al centrarse en la aceptación de las herramientas algorítmicas dentro del sistema judicial español, comparando las percepciones entre la población general y los profesionales del ámbito legal.

Diversos estudios destacan que la inteligencia artificial puede mejorar la eficiencia de los sistemas judiciales al procesar grandes volúmenes de datos y realizar tareas repetitivas con alta precisión. En este sentido, como se estableció en la introducción, algunos autores sugieren que la inteligencia artificial puede servir como una valiosa herramienta de apoyo en ciertas fases del proceso judicial (Cotino Hueso, 2024; Simón Castellano, 2021). Sin embargo, otros plantean serios desafíos en cuanto a transparencia y rendición de cuentas, particularmente en la toma de decisiones judiciales. La teoría de la justicia algorítmica respalda el uso complementario de la inteligencia artificial en áreas específicas, siempre bajo supervisión humana, para garantizar la equidad y la justicia (Marchena, 2022).

En este sentido, la supervisión humana, independientemente del tipo de tarea judicial de que se trate (Barysé & Sarel, 2023), se presenta como un elemento clave para la incorporación de la inteligencia artificial, ya que influye directamente en su aceptación social. La literatura previa señala una tendencia a preferir el uso

complementario de la inteligencia artificial en la administración de justicia, donde el juicio humano se percibe como garantía de imparcialidad, transparencia y objetividad. En esta línea, los resultados del presente estudio confirman una clara inclinación hacia la inteligencia artificial como instrumento de apoyo más que como sistema autónomo, ya que tanto la población general como los profesionales del ámbito jurídico muestran una mayor disposición a aceptarla cuando refuerza, y no sustituye, la intervención humana. Además, varios estudios consideran la supervisión humana un elemento esencial en el diseño e implementación de la IA en la toma de decisiones judiciales (Caterini, 2022; Castro-Toledo, 2022). Esta preferencia por utilizar la IA como un complemento, en lugar de sustituir el papel del juez, refleja preocupaciones sobre el mantenimiento de un enfoque humanista en los procesos judiciales, donde la intervención humana es percibida como esencial para garantizar la imparcialidad, la transparencia y la objetividad.

En este contexto, el presente estudio evalúa la aceptación social actual de las herramientas algorítmicas en el ámbito judicial y explora cómo factores éticos y sociales, como la supervisión humana y el tipo de tarea, influyen en las actitudes hacia su uso. Los resultados revelan una clara preferencia por la incorporación de modelos *human-in-the-loop,* donde se incorpora la intervención humana en alguna fase del proceso de toma de decisiones. Los datos muestran que los participantes, tanto de la población general como del ámbito jurídico, expresan una mayor disposición a aceptar las herramientas cuando actúa como un complemento en lugar de un reemplazo de los operadores humanos en el sistema judicial. En este sentido, la implementación de este tipo de herramientas de manera supervisadas podría mejorar la eficiencia sin comprometer los principios de justicia, ya que facilita ciertos procesos mientras se mantiene el juicio humano para decisiones más complejas.

Esta preferencia es significativamente evidente en la aceptación de aplicaciones de IA en tareas administrativas o de bajo riesgo, como la programación de citas o el uso de chatbots para consultas simples. Esto sugiere que la sociedad valora la tecnología algorítmica para tareas menos complejas donde su intervención no pone en riesgo la integridad del proceso judicial. Sin embargo, cuando se trata de decisiones con

consecuencias graves para los usuarios, como la resolución de un caso penal o la concesión de libertad condicional en procesos de reinserción, la aceptación de herramientas algorítmicas automatizadas disminuye. En estos ámbitos, los participantes expresan escepticismo sobre la capacidad de un algoritmo para tomar decisiones que implican un análisis profundo de circunstancias personales, sociales y contextuales, así como la interpretación de principios jurídicos complejos tradicionalmente asociados al juicio humano.

El estudio también revela diferencias significativas entre la población general y los profesionales del ámbito legal en cuanto a la aceptación de la IA en el sistema judicial. En relación con la primera hipótesis del estudio, se puede afirmar que los profesionales del ámbito legal muestran mayor resistencia a la incorporación de la IA, especialmente en tareas que implican autonomía en la toma de decisiones. Este grupo valora el juicio humano y el conocimiento especializado como elementos fundamentales del sistema judicial, viendo la IA más como un riesgo potencial que como una herramienta útil. La resistencia de los profesionales jurídicos también podría explicarse por el temor a que su rol sea reemplazado o devaluado, o incluso por una sobrevaloración de su propia función, dado que la IA representa un cambio significativo en la estructura tradicional de la justicia.

Por otro lado, la población general exhibe una mayor apertura al uso de la IA en el sistema judicial. Esta disposición puede estar impulsada por la percepción de la IA como una herramienta que moderniza y mejora la eficiencia del sistema judicial. La población general, menos familiarizada con las implicaciones técnicas y normativas de cada decisión, podría considerar el desarrollo de nuevas herramientas de IA como una solución a problemas comunes, como la lentitud de los procesos judiciales y la gestión de grandes volúmenes de información.

Respecto a la segunda hipótesis, los resultados revelan una mayor aceptación cuando las herramientas de IA están diseñadas como ayudas al proceso en lugar de como decisores únicos. La idea de que un operador humano pueda intervenir o revisar el análisis realizado por una herramienta algorítmica se percibe como un factor que mejora la aceptación y legitima el uso de la IA en el ámbito judicial. La supervisión humana actúa como un puente de confianza entre la sociedad y la

tecnología, permitiendo una adopción más gradual y controlada de herramientas algorítmicas en el ámbito judicial.

Finalmente, y en relación con la última hipótesis, el estudio analiza cómo ciertas variables sociodemográficas influyen en la aceptación de la IA en el sistema judicial. Los resultados indican que estas variables no parecen tener un impacto determinante, ya que no se observaron diferencias significativas en la aceptación entre encuestados más jóvenes y mayores.

En conclusión, los hallazgos de este estudio revelan una aceptación condicional de las herramientas algorítmicas en el sistema judicial español, con una clara preferencia por el modelo de Human-In-the-Loop en lugar de como herramientas autónomas (HOTL). Tanto la población general como los profesionales del ámbito legal expresan escepticismo ante la idea de que la IA tome decisiones judiciales de forma independiente, lo que subraya la importancia del juicio humano en el sistema de justicia. La supervisión humana surge como un requisito fundamental para la aceptación social de estas tecnologías, garantizando el control sobre los procesos y proporcionando una percepción de justicia más transparente y ética. Este estudio proporciona datos empíricos que respaldan la necesidad de un enfoque cauteloso y ético en la digitalización de la justicia, equilibrando la eficiencia algorítmica con los principios de transparencia, rendición de cuentas y sensibilidad humana que caracterizan un sistema judicial justo y legítimo.

Estudio 2. Actitudes ciudadanas hacia las herramientas algorítmicas en el sistema de justicia penal: un análisis de sus factores explicativos.

1. Justificación.

En el presente estudio se propone analizar cómo la sociedad percibe el uso de herramientas algorítmicas en los procesos judiciales y qué factores influyen en la formación de actitudes favorables o desfavorables respecto a su adopción. Para ello, el diseño de la investigación se apoya en diversos modelos teóricos sobre adopción tecnológica y comportamiento social, entre ellos la Teoría de la Acción Razonada (TRA) (Fishbein y Ajzen, 1975), la Teoría del Comportamiento Planeado (TPB) (Ajzen, 1991), el Modelo de Aceptación de la Tecnología (TAM) (Davis, 1989), la Teoría Unificada de Aceptación y Uso de la Tecnología (UTAUT) (Venkatesh et al., 2003), el Modelo de Difusión de Innovaciones (Rogers, 2003) y el Technology Readiness Index (TRI) (Parasuraman, 2000). Estos marcos permiten identificar variables, como familiaridad, confianza, percepción de utilidad, imparcialidad u objetividad, que podrían explicar la aceptación social de la IA en el ámbito judicial. Así mismo, la literatura tambien ha señalado que, si bien las herramientas algorítmicas pueden contribuir a una mayor eficiencia y coherencia en las resoluciones judiciales (Binns, 2018; Zalnieriute et al., 2019), también generan inquietudes relacionadas con la falta de transparencia, la opacidad en la toma de decisiones y el riesgo de reproducir o amplificar sesgos existentes (Eubanks, 2018; Noble, 2018). En consecuencia, la aceptación social de estas tecnologías se encuentra estrechamente asociada a la percepción de su imparcialidad, su objetividad y su capacidad para reducir errores humanos (Zavrsnik, 2020).

En este escenario, el presente estudio busca identificar las variables que determinan la aceptación de herramientas algorítmicas en el sistema de justicia penal. El análisis empírico permitirá evaluar en qué medida los factores señalados por los modelos teóricos explican las actitudes ciudadanas hacia la utilización de IA, tanto en configuraciones totalmente automatizadas (HOTL) como en aquellas que incluyen supervisión humana (HITL). En conjunto, la investigación pretende tiene como objetivo analizar qué variables permitirán que el nuevo desarrollo tecnológico sea adoptado de manera socialmente aceptable en el sistema de justicia penal.

2. Objetivos.

El objetivo general de este estudio es analizar en qué medida las variables de los modelos de aceptación social y tecnológica influyen en las actitudes de la ciudadanía hacia la implementación de herramientas algorítmicas en el sistema de justicia penal. Para profundizar en este análisis, se plantean los siguientes objetivos específicos:

OE1. Explorar las actitudes de la sociedad española respecto al uso de herramientas algorítmicas autónomas y supervisadas en el ámbito judicial.

OE2. Identificar los factores que determinan la aceptación del uso de herramientas algorítmicas autónomas y supervisadas.

A partir de los objetivos planteados, se derivan las siguientes hipótesis de investigación:

H1. La percepción de imparcialidad en las decisiones emitidas con el apoyo de herramientas algorítmicas supervisadas o autónomas se asocia positivamente con actitudes favorables hacia su utilización en el sistema judicial.

H2. La creencia de que las herramientas algorítmicas supervisadas o autónomas contribuyen a reducir los errores propios de la decisión judicial humana se asocia positivamente con actitudes favorables hacia su incorporación en los procesos judiciales.

H3. La percepción de riesgo derivada de posibles sesgos algorítmicos se asocia negativamente con las actitudes hacia el uso de herramientas algorítmicas en el ámbito de la justicia penal.

H4. Un mayor nivel de conocimientos tecnológicos y de familiarización con herramientas digitales se asocia positivamente con actitudes favorables hacia el uso de herramientas algorítmicas en el ámbito de la justicia penal.

H5. Factores sociodemográficos como el sexo, la edad y la familiaridad previa con los algoritmos influyen significativamente en las actitudes hacia las herramientas

algorítmicas en el sistema de justicia penal.

3. Metodología.

3.1. Muestra.

La información detallada sobre la composición de la muestra utilizada en este estudio se presenta en las *Consideraciones previas*, ya que se trata de la misma muestra empleada en los tres estudios. En el caso concreto del Estudio 1, participaron los 1.100 sujetos que completaron el cuestionario final, todos ellos expuestos al bloque de preguntas relativas a la aceptación de herramientas algorítmicas.

3.2. Variables e instrumento.

El cuestionario se centra en las diversas variables y modelos teóricos sobre la aceptación social de nuevas tecnologías. Se incorporaron ítems adaptados de los modelos anteriores para casos del sistema de justicia penal, permitiendo evaluar las actitudes específicas de los participantes, en concreto:

- "Los jueces y los abogados deben recibir capacitación sobre el posible uso de herramientas algorítmicas (HA) en el sistema de justicia" (condiciones facilitadoras, UTAUT; normas sociales, TCP).
- "El uso de las herramientas algorítmicas en el sistema de justicia puede perpetuar sesgos y discriminación existentes" (riesgo percibido, TRI; actitud negativa, TRA/TCP).
- "El uso de las herramientas algorítmicas en el sistema de justicia podría ser susceptible a manipulación o sabotaje" (inseguridad, TRI; percepción de riesgo, TCP).
- "Las herramientas algorítmicas deberían ser reguladas para evitar posibles discriminaciones o sesgos en las decisiones judiciales" (normas subjetivas, TCP).
- "Las herramientas algorítmicas utilizadas en el sistema de justicia son objetivas en sus evaluaciones" (utilidad percibida, TAM; actitud positiva, TRA).

- "Las herramientas algorítmicas son más justas en comparación con la intuición humana" (ventaja relativa, MDI ; rendimiento esperado, UTAUT).
- "Las herramientas algorítmicas son más imparciales que los jueces humanos en la toma de decisiones judiciales" (percepción de imparcialidad, TPP/TAM).

Además de las variables mencionadas, se utilizaron como variables dependientes la aceptación de las herramientas algorítmicas en el sistema judicial[12], con la finalidad de relacionar las concepciones anteriores con la aceptación específica de ciertas tareas en el ámbito judicial.

A continuación, se muestra una tabla resumen con las variables:

Tabla 19.

Resumen de las variables incluidas en el estudio 2.

Categoría	**Variable**	**Descripción**
Sociodemográficas	Sexo	Sexo del participante
	Edad	Edad del participante
	Nivel educativo	Máximo nivel de estudios alcanzado
	Conocimiento legal	Formación o experiencia previa en el ámbito legal.
Independientes	Actitudes sobre las herramientas.	Valoraciones cognitivas que las personas tienen respecto al uso de herramientas algorítmicas en el sistema de justicia penal.
Dependiente	Aceptación de herramientas algorítmicas autónomas y supervisadas	Grado de aceptación de las herramientas algorítmicas.

[12] Véase el apartado de procedimiento en las consideraciones previas del Capítulo 6.

3.3. Análisis de datos.

En cuanto al análisis de los datos, en una primera fase se realiza un análisis descriptivo de frecuencias con el propósito de examinar la distribución de las respuestas relativas a las actitudes sobre las herramientas algorítmicas y a los niveles generales de aceptación de estas tecnologías en el ámbito judicial. Este procedimiento inicial permite identificar las principales tendencias actitudinales de la muestra y ofrecer una aproximación preliminar al posicionamiento ciudadano frente a la digitalización de la justicia penal.

A continuación, se comprobó si las variables cumplían los supuestos básicos necesarios para la aplicación de pruebas paramétricas, concretamente los de normalidad, homogeneidad e independencia. Para ello, se utilizó la prueba de Shapiro–Wilk y la inspección visual de los gráficos Q–Q, así como contrastes adicionales para valorar la estructura de las distribuciones. A raíz de estas comprobaciones se observó que las variables no cumplían adecuadamente los supuestos requeridos para la aplicación de técnicas paramétricas. Por este motivo, se optó por emplear análisis no paramétricos equivalentes, más apropiados para este tipo de datos y para las características de la muestra.

Posteriormente se aplicó la correlación de Spearman, atendiendo al carácter ordinal de las variables y a la ausencia del supuesto de normalidad. Posteriormente se estimaron modelos de regresión lineal múltiple, dado que esta técnica es robusta frente a desviaciones moderadas de normalidad y permite analizar el efecto conjunto de las variables predictoras.

En una tercera etapa se estiman modelos de regresión lineal múltiple, cuyo objetivo es evaluar el peso relativo de diferentes variables sociodemográficas y actitudinales en la explicación de la aceptación de herramientas algorítmicas. A diferencia de la correlación, que se limita a examinar asociaciones bivariadas, la regresión lineal permite determinar el efecto específico de cada variable al considerar de manera simultánea la influencia de las demás. Este enfoque posibilita analizar de forma integrada factores como la edad, el sexo, el nivel educativo, la formación y distintas creencias sobre el funcionamiento de los algoritmos, diferenciando en todo

momento entre la aceptación de sistemas completamente autónomos y aquellos sometidos a supervisión humana. La interpretación de los coeficientes se fundamenta en la significación estadística ($p < .05$), que indica si la relación observada es improbable que se deba al azar.

Finalmente, se construyeron árboles de decisión como técnica de análisis no paramétrica orientada a identificar combinaciones de variables que expliquen de forma jerárquica las decisiones de aceptación o rechazo (Berlanga, Rubio Hurtado & Vilà Baños, 2013). Esta técnica facilitó una interpretación visual e intuitiva de los perfiles de respuesta, y se utilizó tanto con variables demográficas como con variables teóricas y de experiencia, diferenciando además entre escenarios de herramientas totalmente automatizadas y aquellas utilizadas con supervisión humana (*human-in-the-loop*). El análisis se llevó a cabo a través del software estadístico SPSS v.29, que ofrece una implementación robusta de algoritmos de árbol (CHAID), que divide progresivamente la muestra en función de la variable predictora que maximiza la reducción de heterogeneidad (medida mediante estadísticos χ^2 y ajustes de significación) en la variable dependiente, en este caso, la aceptación de las herramientas algorítmicas. Las categorías de las respuestas (escala Likert 1-5) se agrupan automáticamente cuando no existen diferencias estadísticamente relevantes entre ellas en términos de aceptación, lo que permite simplificar la estructura y mejorar la estabilidad del modelo. Las variables predictoras se seleccionan secuencialmente: solo aparecen aquellas que aportan información explicativa significativa en cada nodo; por tanto, el hecho de que algunas variables no figuren en el árbol no implica irrelevancia conceptual, sino que no contribuyeron a mejorar la clasificación una vez incorporadas las variables anteriores. Este comportamiento es característico de los árboles de decisión, ya que priorizan la ganancia incremental de información y no la inclusión obligatoria de todas las variables. Lo que se busca como resultado final es una estructura jerárquica interpretable que permite identificar las creencias que mejor diferencian perfiles, en este caso, de mayor o menor aceptación hacia el uso de algoritmos en el ámbito judicial. Estos análisis se estructuraron en torno a cuatro árboles de decisión independientes, cada uno orientado a responder a las preguntas específicas de investigación:

- Árbol 1 – Variables demográficas y aceptación de herramientas algorítmicas con supervisión humana (HITL).

El primer modelo expone la relación entre variables sociodemográficas, como edad, sexo y la familiaridad con los algoritmos y la aceptación de herramientas algorítmicas que operan como apoyo al juicio humano. Este análisis permite identificar patrones iniciales de predisposición en función de características demográficas.

- Árbol 2 – Actitudes y aceptación de herramientas algorítmicas con supervisión humana (HITL).

El segundo modelo se construyó a partir de variables derivadas de marcos teóricos sobre adopción tecnológica (TAM, UTAUT, TPB y TRI). Su propósito es explorar cómo dimensiones como la confianza, la percepción de imparcialidad, la transparencia o la regulación influyen en la aceptación de las herramientas algorítmicas cuando se emplean bajo supervisión humana.

- Árbol 3 – Variables demográficas y aceptación de herramientas autónomas (HOTL).

El tercer modelo replica el análisis realizado en el primer árbol, pero aplicado a herramientas algorítmicas que operan de manera autónoma, es decir, sin supervisión humana en la toma de decisiones judiciales. De este modo, se examina si la ausencia de intervención humana altera las asociaciones previamente detectadas entre factores sociodemográficos y niveles de aceptación.

- Árbol 4 – Actitudes y herramientas algorítmicas autónomas (HOTL).

Finalmente, el cuarto modelo incorpora variables relacionadas con creencias, actitudes y factores académicos, como el nivel de estudios, la formación en nuevas tecnologías o el grado de familiarización con herramientas algorítmicas. Este análisis permite identificar en qué medida dichos factores influyen en la aceptación de tecnologías que actúan con plena autonomía en el proceso judicial.

Este enfoque metodológico permite detectar si existen diferencias significativas en

la aceptación de las herramientas algorítmicas en función de diversas variables individuales, tales como las creencias específicas sobre su funcionamiento, las características sociodemográficas de la ciudadanía o el nivel de familiarización tecnológica. La utilización de árboles de decisión, además, proporciona una herramienta visual y accesible para la interpretación de los resultados, facilitando la identificación de perfiles y patrones dentro de la población analizada.

El análisis de los datos se ha realizado haciendo uso del paquete estadístico SPSS v.29 y RStudio. Así mismo, el código utilizado en R se incluye en los anexos.

4. Resultados.

4.1. Análisis descriptivo de las variables.

Antes de abordar los análisis inferenciales, se realizó un estudio descriptivo de frecuencias con el objetivo de identificar las tendencias generales en las respuestas de la muestra en relación con las creencias sobre la inteligencia artificial y la aceptación de su uso en el sistema de justicia penal. Este análisis permite conocer el grado de acuerdo o desacuerdo de los participantes con respecto a aspectos clave como la transparencia, la imparcialidad, el sesgo algorítmico, la necesidad de regulación o la confianza en la tecnología, así como su actitud general hacia el uso de herramientas algorítmicas, tanto autónomas como supervisadas. Los resultados obtenidos ofrecen una visión inicial del posicionamiento social ante la implementación de estas tecnologías, y sirven como base para el posterior análisis relacional y explicativo.

Tabla 20.

Frecuencia de las variables independientes.

	Totalmente en desacuerdo		En desacuerdo		Ni de acuerdo ni en desacuerdo		De acuerdo		Totalmente de acuerdo	
	n	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Los jueces y los abogados deben recibir capacitación sobre el posible uso de herramientas algorítmicas en el sistema de justicia.	68	6,2	87	7,9	256	23,3	232	21,1	457	41,5
El uso de las herramientas algorítmicas en el sistema de justicia puede perpetuar sesgos y discriminación existentes.	77	7,0	89	8,1	389	35,4	264	24,0	281	25,5
El uso de las herramientas algorítmicas en el sistema de justicia podría ser susceptible a manipulación o sabotaje.	38	3,5	74	6,7	294	26,7	301	27,4	393	35,7
Las herramientas algorítmicas deberían ser reguladas para evitar posibles discriminaciones o sesgos en las decisiones judiciales.	24	2,2	52	4,7	222	20,2	263	23,9	539	49,0
Las herramientas algorítmicas utilizadas en el sistema de justicia son objetivas en sus evaluaciones.	118	10,7	153	13,9	542	49,3	176	16,0	111	10,1
Las herramientas algorítmicas son más justas en comparación con la intuición humana.	226	20,5	253	23,0	422	38,4	134	12,2	65	5,9
Las herramientas algorítmicas son más imparciales que los jueces humanos en la toma de decisiones judiciales.	98	8,9	141	12,8	446	40,5	266	24,2	149	13,5
Aceptación general – HITL.	70	6,4	209	19	533	48,5	222	20,2	66	6,0
Aceptación general – HOTL.	76	6,9	228	20,7	569	51,7	194	17,6	33	3,0

4.2. Correlación de Sperman.

Una vez descritas las tendencias generales de las variables, se procedió a realizar un análisis de correlación de Spearman con el fin de explorar la existencia de relaciones estadísticas significativas entre las creencias identificadas y la aceptación de las herramientas algorítmicas en el sistema de justicia penal. Dado que las variables en estudio son de naturaleza ordinal y no se parte del supuesto de normalidad, la correlación de Spearman se presenta como la técnica más adecuada para estimar la intensidad y dirección de dichas asociaciones. Este análisis permite identificar cuáles de las creencias, como la percepción de transparencia, la imparcialidad o el sesgo algorítmico, guardan una relación significativa con las actitudes ciudadanas hacia la implementación de inteligencia artificial, tanto en escenarios de autonomía total como en aquellos que contemplan supervisión humana. Los resultados se muestran en la siguiente tabla (Tabla 21):

Tabla 21.

Matriz de correlación de Spearman entre las variables dependientes y los modelos de interacción humano-máquina.

	Aceptación de IA modelo HOTL	**Aceptación de IA modelo HITL**
Los jueces y los abogados deben recibir capacitación sobre el posible uso de herramientas algorítmicas en el sistema de justicia	,187**	,249**
El uso de las herramientas algorítmicas en el sistema de justicia puede perpetuar sesgos y discriminación existentes.	-,083**	-,039
El uso de las herramientas algorítmicas en el sistema de justicia podría ser susceptible a manipulación o sabotaje	-,134**	-,096**
Las herramientas algorítmicas deberían ser reguladas para evitar posibles discriminaciones o sesgos en las decisiones judiciales.	0,14	,098**
Las herramientas algorítmicas utilizadas en el sistema de justicia son objetivas en sus evaluaciones.	,382**	,391**
Las herramientas algorítmicas son más justas en comparación con la intuición humana.	,517**	,490*
Las herramientas algorítmicas son más imparciales que los jueces en la toma de decisiones judiciales.	,416**	,412**

En el análisis de las creencias sobre el uso de herramientas algorítmicas en el sistema judicial, se identificaron relaciones significativas entre la aceptación de la inteligencia artificial y diversas creencias relacionadas con su implementación.

Los resultados mostraron una correlación positiva de magnitud débil entre la aceptación de la inteligencia artificial en su modalidad autónoma y la creencia en la necesidad de formación especializada sobre estas tecnologías ($r = .187$, $p < .01$). Una correlación similar se observó entre la aceptación de la IA supervisada (*human-in-the-loop*) y dicha creencia ($r = .249$, $p < .01$). También se observaron correlaciones negativas entre la aceptación de la IA y la creencia de que estas herramientas pueden reproducir sesgos y discriminaciones preexistentes. En el caso de la IA autónoma, la correlación fue significativa, aunque de magnitud baja ($r = –.083$, $p < .01$), mientras que para la IA supervisada no se alcanzó significación estadística ($r = –.039$, $p > .05$). Asimismo, se identificó una correlación negativa entre la aceptación de la IA autónoma y la creencia de que estas herramientas pueden ser objeto de manipulación o sabotaje ($r = –.134$, $p < .01$). Se registraron correlaciones positivas y significativas entre la aceptación de la IA y la creencia de que las herramientas son objetivas en sus evaluaciones. En el caso de la IA autónoma, la correlación fue de $r = .382$ ($p < .01$), y en el de la IA supervisada, $r = .391$ ($p < .01$). Finalmente, se observaron correlaciones positivas y significativas entre la aceptación de la IA y las creencia de que las herramientas son más justas en comparación con la intuición humana ($r = .517$, $p < .01$ para IA autónoma; $r = .490$, $p < .05$ para IA supervisada), así como con la creencia de que son más imparciales que los jueces humanos ($r = .416$, $p < .01$ para IA autónoma; $r = .412$, $p < .05$ para IA supervisada).

4.3. Modelos de regresión lineales.

Con el objetivo de profundizar en el análisis de los factores que influyen en la aceptación de las herramientas algorítmicas en el ámbito de la justicia penal, se procedió a la estimación de modelos de regresión lineal. Este tipo de análisis permite evaluar la probabilidad de aceptación en función de un conjunto de variables independientes, considerando tanto las creencias previamente identificadas como las tareas específicas en las que la inteligencia artificial podría ser implementada dentro del proceso judicial. Se diseñaron distintos modelos, uno

por cada tarea concreta planteada (por ejemplo, evaluación del riesgo de reincidencia, clasificación de gravedad, recomendación de medidas, entre otras), lo que permitió identificar patrones diferenciados según el tipo de aplicación propuesta.

4.3.1. Modelos de regresión lineal múltiple de la aceptación de herramientas supervisadas (Human-in-the-loop).

Previo a la exposición de los resultados, es necesario recordar que varias de las variables incluidas en el modelo se midieron mediante escalas tipo Likert de cinco niveles, diseñadas para recoger grados crecientes de acuerdo con cada una de las creencias específicas. Estas variables fueron incorporadas al modelo como variables categóricas tomando como referencia el nivel 1 (totalmente en desacuerdo) hasta el nivel 5 (totalmente de acuerdo). Este enfoque permite estimar el efecto de cada nivel sobre la variable dependiente, en este caso, la aceptación de herramientas algorítmicas supervisadas, y facilita la comparación entre diferentes grados de acuerdo.

Los resultados del modelo indican que la aceptación de herramientas algorítmicas supervisadas (Human-in-the-Loop) se explica sobre todo por las actitudes de las personas que consideran que estas herramientas son objetivas, justas e imparciales. Además, el modelo revela que las personas que creen que los profesionales deben recibir formación específica para utilizar estas herramientas también presentan niveles significativamente mayores de aceptación, de modo que la demanda de capacitación aparece como un segundo eje que impulsa el apoyo hacia estos sistemas supervisados. En conjunto, estos resultados subrayan que la aceptación no depende tanto de características personales, sino de la convicción de que los algoritmos aportan mayor equidad y de que su uso requiere una preparación adecuada.

Por el contrario, las variables sociodemográficas muestran un patrón menos claro. Aunque el sexo presenta un efecto pequeño pero significativo ($B = -0.067$, $p = .038$), la edad apenas exhibe asociaciones robustas: solo el grupo de edad 30–44 años muestra un coeficiente marginalmente significativo ($B = -0.057$, $p = .040$), mientras

que los restantes tramos no alcanzan los umbrales de significación ($p > .05$). De modo similar, las creencias relativas a la perpetuación de sesgos, la posibilidad de manipulación de las herramientas o la necesidad de regulaciones legales tampoco muestran efectos significativos en ninguno de sus niveles ($p > .05$ en todos los casos), lo que sugiere que estas dimensiones no explican diferencias relevantes en la aceptación.

En cuanto a los predictores significativos, la percepción de que los profesionales deben recibir formación sobre estas herramientas emerge como un factor importante: sus niveles 2 ($B = 0.188$, $p = .048$), 3 ($B = 0.246$, $p = .003$), 4 ($B = 0.326$, $p < .001$) y 5 ($B = 0.349$, $p < .001$) presentan asociaciones positivas y estadísticamente significativas. Este patrón indica que, cuanto mayor es la creencia de que la capacitación profesional es necesaria, mayor es la disposición a aceptar herramientas algorítmicas supervisadas. Una de las relaciones más sólidas del modelo proviene de la creencia de que las herramientas son objetivas, con coeficientes significativos en los niveles 2 ($B = 0.123$, $p < .001$), 3 ($B = 0.197$, $p < .001$), 4 ($B = 0.189$, $p < .001$) y 5 ($B = 0.470$, $p < .001$). Del mismo modo, la idea de que las herramientas son más justas que los humanos presenta uno de los gradientes más intensos de todo el análisis, con asociaciones fuertes en los niveles 2 ($B = 0.197$, $p < .001$), 3 ($B = 0.402$, $p < .001$), 4 ($B = 0.785$, $p < .001$) y 5 ($B = 1.030$, $p < .001$). Finalmente, la percepción de que las herramientas son *más imparciales que los jueces* también muestra efectos relevantes, aunque más moderados que los anteriores: los niveles 4 ($B = 0.361$, $p = .011$) y 5 ($B = 0.578$, $p < .001$) resultan significativos, mientras que los niveles 2 y 3 no alcanzan la significación estadística.

El modelo mostró un ajuste adecuado ($R^2 = 0.393$; R^2 ajustado = 0.372), indicando que aproximadamente el 37% de la variabilidad de la variable dependiente queda explicada por los predictores incluidos. El error estándar residual fue de 0.712, lo que refleja una desviación moderada entre valores predichos y observados. La prueba global del modelo resultó significativa ($F(36, 1060) = 19.06$, $p < .001$).

Tabla 22.

Resultados del análisis de regresión lineal múltiple para la aceptación de herramientas supervisadas (Human-in-the-Loop).

	B	SE B	β	p[a]	IC 95%
Intercepto	1.630	.181	-.040	.000***	[1.27; 1.982]
Sexo nivel 2 - mujer	-.067	.041	-.068	.038*	[-.149; .014]
Grupo edad (30-44)	-.057	.060	-.103	.040*	[-.174; .060]
Grupo edad (45-54)	-.086	.064	.024	.061.	[-.212; .041]
Grupo edad (55-64)	.020	.070	.142	.778	[-.118; .157]
Estar familiarizado con el uso de algoritmos – nivel 2	.118	.056	.123	.981	[.008; .229]
Estar familiarizado con el uso de algoritmos – nivel 3	.103	.057	.196	.634	[-.009; .214]
Estar familiarizado con el uso de algoritmos – nivel 4	.163	.090	.435	.081.	[-.014; .341]
Estar familiarizado con el uso de algoritmos – nivel 5	.362	.120	.225	.503	[.127; 0.597]
Los profesionales deben recibir formación – nivel 2	.188	.119	.295	.048*	[-.045; .421]
Los profesionales deben recibir formación – nivel 3	.246	.106	.391	.003**	[.038; .454]
Los profesionales deben recibir formación – nivel 4	.326	.108	.418	.000***	[.114; .537]
Los profesionales deben recibir formación – nivel 5	.349	.100	-.104	.000***	[.153; .544]
Las herramientas perpetúan sesgos- nivel 2	-.087	.119	-.127	.624	[-.320; .146]
Las herramientas perpetúan sesgos- nivel 3	-.106	.102	-.046	.801	[-.306; .094]
Las herramientas perpetúan sesgos- nivel 4	-.038	.104	-.126	.426	[-.242; .166]
Las herramientas perpetúan sesgos- nivel 5	-.105	.099	.239	.685	[-.300; .090]
Las herramientas se pueden manipular – nivel 2	.199	.147	.121	.258	[-.090; .488]
Las herramientas se pueden manipular – nivel 3	.101	.131	.163	.969	[-.156; .358]
Las herramientas se pueden manipular – nivel 4	.136	.132	.013	.755	[-.122; .394]
Las herramientas se pueden manipular – nivel 5	.011	.129	.195	.723	[-.242; .263]
Es necesario regularlas legalmente -nivel 2	.162	.188	.380	.266	[-.206; .530]
Es necesario regularlas legalmente -nivel 3	.316	.165	.395	.543	[-.008; .641]
Es necesario regularlas legalmente -nivel 4	.329	.166	.384	.621	[.004; .654]
Es necesario regularlas legalmente -nivel 5	.320	.160	.148	.240	[.006; .633]
Las herramientas son objetivas- nivel 2	.123	.092	.236	.000***	[-.058; .304]
Las herramientas son objetivas- nivel 3	.197	.083	.227	.000***	[.034; .359]
Las herramientas son objetivas- nivel 4	.189	.095	.564	.000***	[.002; .376]
Las herramientas son objetivas- nivel 5	.470	.101	.237	.000***	[.272; .669]
Mas justas que los humanos – nivel 2	.197	.073	.483	.000***	[.054; .340]
Mas justas que los humanos – nivel 3	.402	.071	.942	.000***	[.263; .541]
Mas justas que los humanos – nivel 4	.785	.089	1.230	.000***	[.610; .959]
Mas justas que los humanos – nivel 5	1.030	.112	.041	.000***	[.807; 1.245]
Mas imparciales que los jueces – nivel 2	.034	.101	.370	.243	[-.164; .232]
Mas imparciales que los jueces – nivel 3	.308	.092	.433	.065.	[.128; .489]
Mas imparciales que los jueces – nivel 4	.361	.097	.694	.011*	[.170; .552]
Mas imparciales que los jueces – nivel 5	.578	.100	-.040	.000***	[.381; .775]

[a] Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

4.3.2. Modelos de regresión lineal de herramientas algorítmicas autónomas (Human-out-of-the-Loop).

El presente modelo sigue la misma estructura general que el anterior, aunque los patrones de significación y magnitud de los efectos difieren entre los escenarios supervisados HITL y autónomos HOTL.

Los resultados del modelo muestran que la aceptación de herramientas algorítmicas autónomas depende principalmente de un conjunto específico de actitudes hacia estas tecnologías, más que de características sociodemográficas. En particular, los incrementos más fuertes en la aceptación aparecen entre quienes están familiarizados con este tipo de sistemas, así como aquellos que consideran que debe recibirse formación. Igualmente, aquellos que consideran que estas herramientas producen decisiones objetivas, más justas o imparciales que las emitidas por operadores humanos. Estos predictores, especialmente en sus niveles más altos de acuerdo, se consolidan como los efectos más robustos del modelo, lo que sugiere que la disposición a adoptar sistemas totalmente automatizados se fundamenta en percepciones positivas sobre su fiabilidad, coherencia y familiaridad.

Por el contrario, las variables sociodemográficas (sexo y grupos de edad) no presentan asociaciones estadísticamente significativas ($p > .05$ en todos los casos), y del mismo modo tampoco resultan significativas las creencias sobre la perpetuación de sesgos o la posibilidad de manipulación. Todo ello indica que las diferencias en la aceptación no se explican por factores individuales ni por percepciones de riesgo, sino por evaluaciones valorativas del rendimiento comparado entre humanos y algoritmos.

Entre los predictores significativos, la familiaridad con el uso de algoritmos presenta asociaciones positivas tanto en los niveles iniciales como en los más altos (nivel 2: $B = 0.118$, $p = .035$; nivel 5: $B = 0.362$, $p = .002$), lo que sugiere que tanto la exposición básica como el dominio avanzado de estas herramientas incrementan la disposición a aceptarlas. Asimismo, la convicción de que los profesionales deben recibir formación especializada muestra efectos significativos en los niveles 3 ($B = 0.246$, $p = .020$), 4 ($B = 0.398$, $p < .001$) y 5 ($B = 0.349$, $p < .001$), evidenciando que una mayor

demanda de capacitación se asocia consistentemente con un mayor apoyo a la automatización algorítmica. La creencia de que los algoritmos son objetivos se consolida también como uno de los predictores más robustos, con asociaciones positivas en los niveles 3 (B = 0.197, p = .018), 4 (B = 0.189, p = .047) y 5 (B = 0.470, p < .001). Del mismo modo, la percepción de que las decisiones producidas por estas herramientas son más justas que las humanas presentan uno de los gradientes más intensos del modelo, con efectos significativos en los niveles 2 (B = 0.197, p = .006), 3 (B = 0.402, p < .001), 4 (B = 0.785, p < .001) y 5 (B = 1.31, p < .001). Finalmente, la creencia de que los algoritmos son más imparciales que los jueces también muestra asociaciones significativas en niveles altos (nivel 3: B = 0.308, p < .001; nivel 4: B = 0.361, p < .001; nivel 5: B = 0.578, p < .001), reforzando la idea de que la percepción de superioridad moral y decisoria respecto al juicio humano constituye el principal motor de aceptación de la automatización total.

El modelo mostró un R^2 de 0.405 y un R^2 ajustado de 0.385, lo que indica que aproximadamente el 38% de la varianza de la variable dependiente queda explicada por los predictores incluidos. El error estándar residual fue de 0.653 y la prueba global del modelo fue significativa ($F(36, 1060) = 20.06$, $p < 0.001$).

Tabla 23.

Resultados del análisis de regresión lineal múltiple para la aceptación de herramientas autónomas (Human-out-of-the-Loop).

	B	SE B	β	pª	IC 95%
Intercepto	1.63	.190	NA	.000***	[1.271; 1.982]
Sexo nivel 2 - mujer	-.067	.041	-.044	.103	[-0.149; 0.014]
Grupo edad (30-44)	-.057	.060	-.073	.342	[-0.174; 0.06]
Grupo edad (45-54)	-.087	.064	-.111	.184	[-0.212; 0.041]
Grupo edad (55-64)	.012	.070	.011	.778	[-0.118; 0.157]
Estar familiarizado con el uso de algoritmos – nivel 2	.118	.056	.164	.035*	[0.008; 0.229]
Estar familiarizado con el uso de algoritmos – nivel 3	.103	.062	.194	.071.	[-0.009; 0.214]
Estar familiarizado con el uso de algoritmos – nivel 4	.163	.097	.289	.071.	[-0.014; 0.341]
Estar familiarizado con el uso de algoritmos – nivel 5	.362	.121	.487	.002**	[0.127; 0.597]
Los profesionales deben recibir formación – nivel 2	.188	.118	.221	.113	[-0.045; 0.421]
Los profesionales deben recibir formación – nivel 3	.246	.106	.295	.020*	[0.038; 0.454]
Los profesionales deben recibir formación – nivel 4	.326	.108	.403	.002**	[0.114; 0.537]
Los profesionales deben recibir formación – nivel 5	.349	.100	.426	.000***	[0.153; 0.544]
Las herramientas perpetúan sesgos- nivel 2	-.087	.118	-.109	.465	[-0.32; 0.146]
Las herramientas perpetúan sesgos- nivel 3	-.106	.102	-.134	.299	[-0.306; 0.094]
Las herramientas perpetúan sesgos- nivel 4	-.038	.104	-.048	.716	[-0.242; 0.166]
Las herramientas perpetúan sesgos- nivel 5	-.105	.099	-.122	.290	[-0.3; 0.09]
Las herramientas se pueden manipular – nivel 2	.199	.147	.221	.176	[-0.09; 0.488]
Las herramientas se pueden manipular – nivel 3	.101	.131	.121	.443	[-0.156; 0.358]
Las herramientas se pueden manipular – nivel 4	.136	.131	.163	.301	[-0.122; 0.394]
Las herramientas se pueden manipular – nivel 5	.012	.129	.009	.935	[-0.242; 0.263]
Es necesario regularlas legalmente -nivel 2	.162	.187	.208	.388	[-0.206; 0.53]
Es necesario regularlas legalmente -nivel 3	.316	.165	.381	.055.	[-0.008; 0.641]
Es necesario regularlas legalmente -nivel 4	.329	.165	.404	.047*	[0.004; 0.654]
Es necesario regularlas legalmente -nivel 5	.320	.160	.405	.045*	[0.006; 0.633]
Las herramientas son objetivas- nivel 2	.123	.092	.150	.182	[-0.058; 0.304]
Las herramientas son objetivas- nivel 3	.197	.083	.236	.017*	[0.034; 0.359]
Las herramientas son objetivas- nivel 4	.189	.095	.224	.045*	[0.002; 0.376]
Las herramientas son objetivas- nivel 5	.470	.101	.557	.000***	[0.272; 0.669]
Mas justas que los humanos – nivel 2	.197	.073	.230	.006**	[0.054; 0.34]
Mas justas que los humanos – nivel 3	.402	.071	.469	.000***	[0.263; 0.541]
Mas justas que los humanos – nivel 4	.785	.089	.927	.000***	[0.61; 0.959]
Mas justas que los humanos – nivel 5	1.03	.111	1.23	.000***	[0.807; 1.245]
Mas imparciales que los jueces – nivel 2	.034	.101	.044	.713	[-0.164; 0.232]
Mas imparciales que los jueces – nivel 3	.308	.091	.374	.000***	[0.128; 0.489]
Mas imparciales que los jueces – nivel 4	.361	.097	.434	.000***	[0.17; 0.552]
Mas imparciales que los jueces – nivel 5	.578	.100	.702	.000***	[0.381; 0.775]

ª Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

4.4. Arboles de decisión.

Con el objetivo de explorar los factores asociados a la aceptación de las herramientas algorítmicas en el sistema de justicia penal, se aplicaron árboles de decisión como técnica de análisis no paramétrica. Tal como se ha detallado en la sección metodológica, se desarrollaron cuatro modelos diferenciados en función del tipo de variables y del tipo de implementación de herramientas algorítmicas (autónoma o con supervisión humana).

A continuación, se presentan los resultados obtenidos para cada uno de estos árboles de forma individual, con el fin de identificar las variables que explican de manera jerárquica la aceptación en cada escenario analizado.

4.4.1. Variables sociodemográficas y aceptación de herramientas algorítmicas con supervisión humana.

En primer lugar, se analizó las variables sociodemográficas en relación con la aceptación de herramientas algorítmicas supervisadas. La variable dependiente se presenta en una escala ordinal de cinco categorías, que van desde "Totalmente en desacuerdo" hasta "Totalmente de acuerdo" donde se les pregunta si estarían de acuerdo en que se utilizaran herramientas algorítmicas.

En este caso concreto se observa que la aceptación de las herramientas supervisadas (human-in-the-loop) está condicionada principalmente por características sociodemográficas, siendo el sexo el factor que inicia la primera división del árbol. Esta primera división indica que existen diferencias significativas entre hombres y mujeres en la forma en que valoran este tipo de herramientas: mientras que los hombres muestran niveles ligeramente inferiores de aceptación, las mujeres presentan una distribución más inclinada hacia posiciones de acuerdo, lo que evidencia que el género actúa como punto de partida para comprender la tendencia general. En una segunda rama, dentro del grupo de mujeres, la edad opera como modulador adicional. Las mujeres de 38 años o menos tienden a situarse en posiciones más neutrales o moderadamente favorables, mientras que las que tienen más de 38 años muestran una inclinación algo mayor hacia el acuerdo o desacuerdo explícito. Finalmente, entre las mujeres de 38 años o menos, la familiaridad con los

algoritmos jurídicos introduce un nuevo nivel de diferenciación. Quienes no están familiarizadas con estas herramientas tienden a adoptar posiciones más neutrales o incluso algo más escépticas, mientras que quienes sí poseen familiaridad muestran los porcentajes más altos de acuerdo y mayor apertura hacia la adopción de sistemas supervisados. Este patrón refuerza la idea de que la experiencia o cercanía previa con tecnologías jurídicas contribuye a generar confianza y a mejorar la predisposición hacia su aceptación.

El nodo raíz (Nodo 0), que agrupa la totalidad de la muestra (N = 1100), indica que la mayor parte de los participantes se sitúan en una posición neutral respecto a la aceptación de la IA con supervisión humana ("Ni de acuerdo ni en desacuerdo", 48.5%), seguida por una distribución más reducida entre las categorías de acuerdo y en desacuerdo.

La primera variable predictora que divide el árbol es el sexo del encuestado, resultando significativo con un valor ajustado de p = .031 (χ^2 = 4.655, gl = 1). Esta variable origina dos ramas principales:

- Nodo 1: Hombre que se muestran mayoritariamente neutrales (49.5%), con un 27.4% distribuidos entre las categorías de acuerdo. Esta rama no se subdivide más, lo que indica que ese grupo es un nodo terminal, y su composición de respuestas se considera suficientemente explicada sin más particiones.
- Nodo 2: Mujeres cuya distribución también es mayoritariamente neutral (47.5%), pero permite una subdivisión posterior significativa.

En la rama masculina no se identificaron predictores adicionales significativos, por lo que el nodo queda terminal, aunque aceptaban más las herramientas que las mujeres. En el caso de la rama femenina, la segunda variable introducida es la edad, con un valor ajustado de p = .020 (χ^2 = 9.375, gl = 1), generando dos nuevos nodos:

- Nodo 3: Mujeres de 38 años o menos. Este grupo muestra una actitud más favorable hacia la IA, con un 25.1% en las categorías de acuerdo. Este nodo se divide posteriormente según el nivel de familiarización con los algoritmos jurídicos.

- Nodo 4: Las mujeres mayores de 38 años muestran una distribución más ambivalente, con un leve aumento desacuerdo en comparación con las mujeres jóvenes, aunque la mayoría sigue situada en posiciones neutrales.

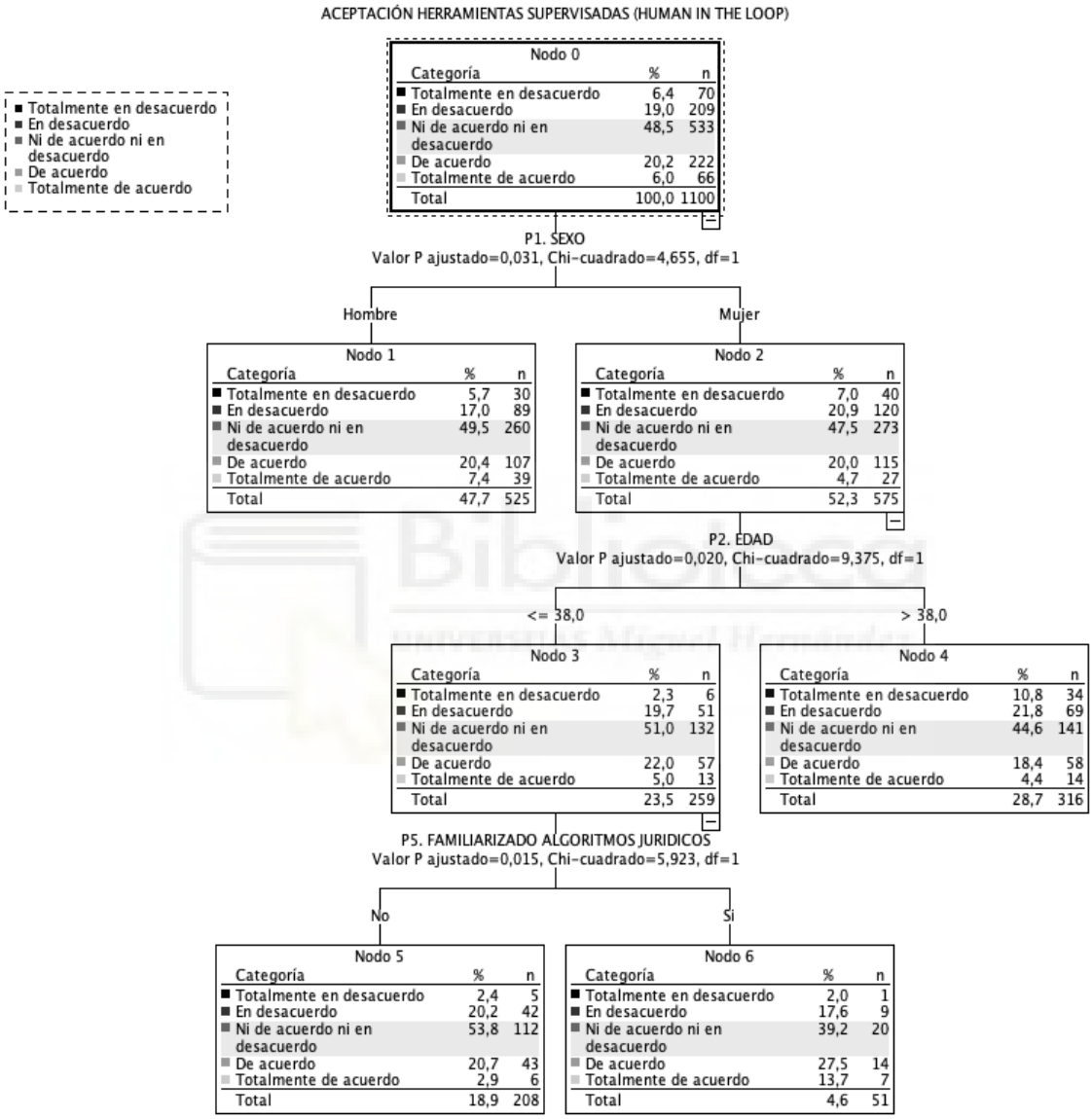
En el Nodo 3, la variable de familiaridad con algoritmos en el ámbito judicial desarrolla una nueva bifurcación significativa ($p = .015$; $\chi^2 = 5.923$, $gl = 1$):

- Nodo 5: Mujeres jóvenes no familiarizadas con algoritmos jurídicos (N = 208), donde predomina la actitud neutral (53.8%), y las posturas de acuerdo alcanzan solo un 11.5%.
- Nodo 6: Mujeres jóvenes familiarizadas con dichos algoritmos (N = 51), quienes presentan la actitud más favorable del conjunto del árbol, con un 43.2% en las categorías de acuerdo ("De acuerdo" y "Totalmente de acuerdo").

En síntesis, el modelo indica que el nivel de aceptación general de la IA supervisada se ve influido significativamente por el sexo, la edad y la familiarización previa con algoritmos jurídicos. El perfil más favorable hacia la adopción de la IA corresponde a mujeres jóvenes (≤ 38 años) con experiencia previa o familiarización con herramientas algorítmicas en el ámbito jurídico. Por el contrario, el perfil más reticente se encuentra entre los hombres. Estos resultados se muestran de la Figura 12:

Figura 12.

Árbol de decisión de las variables sociodemográficas y la aceptación de herramientas algorítmicas con supervisión humana.



4.4.2. Creencias y aceptación de herramientas algorítmicas con supervisión humana.

En segundo lugar, se analizó la aceptación de herramientas algorítmicas con supervisión humana en función de variables actitudinales y de experiencia previa. Del mismo modo, la variable dependiente se presenta en una escala ordinal de cinco categorías.

En este caso concreto se muestra que la aceptación de las herramientas algorítmicas en el sistema de justicia está fundamentalmente asociada a percepciones positivas sobre su justicia, objetividad e imparcialidad. La primera división se produce según el grado en que los encuestados creen que los algoritmos son más justos que la intuición humana, lo que indica que esta creencia constituye el núcleo de la actitud favorable hacia su adopción. Entre quienes no comparten esta idea, la aceptación es muy baja, reforzando la importancia de la confianza en la equidad del sistema automatizado. En ramas posteriores, la percepción de objetividad e imparcialidad del algoritmo actúa como modulador adicional: incluso entre individuos inicialmente escépticos, (agrupación de totalmente en desacuerdo y en desacuerdo) aquellos que consideran posible la imparcialidad algorítmica tienden a mostrar una aceptación relativamente mayor. Finalmente, en los perfiles moderadamente favorables, la aceptación aumenta especialmente cuando se reconoce la necesidad de capacitación de jueces y abogados, lo que sugiere que la confianza en la implementación responsable y supervisada también contribuye a mejorar la predisposición hacia estas herramientas.

De manera más concreta, encontramos el nodo raíz (Nodo 0), que agrupa la totalidad de la muestra (N = 1100), y como puede observarse, la mayor parte de los participantes se sitúan en una posición neutral respecto a la aceptación general de herramientas algorítmicas predictivas (“Ni de acuerdo ni en desacuerdo”, 48.5%). La primera variable predictora que divide el árbol es la variable “las herramientas algorítmicas son más justas en comparación con la intuición humana” resultando altamente significativa con un valor ajustado de p = .000 (χ^2 =307.757, gl = 4). Esta variable origina cinco ramas principales:

- Nodo 1: Participantes que están totalmente en desacuerdo con la afirmación de que las herramientas algorítmicas son más justas que la intuición humana. Este grupo manifiesta un patrón de rechazo hacia las herramientas algorítmicas supervisadas, con la mayoría de las respuestas en las categorías "En desacuerdo" (38,5%) o "Totalmente en desacuerdo" (19,9%).
- Nodo 2: Participantes con baja aceptación con la afirmación de que las herramientas algorítmicas son más justas que la intuición humana, se muestran mayoritariamente en desacuerdo respecto a su uso general (51%).
- Nodo 3: Participantes con aceptación intermedia o neutral con la afirmación, presenta actitudes más variadas, con predominio de la neutralidad y ciertas inclinaciones tanto hacia el acuerdo como hacia el desacuerdo.
- Nodo 4: Participantes con aceptación moderada de la afirmación manifiestan una mayor apertura hacia el uso de herramientas algorítmicas supervisadas, aunque todavía existe una proporción importante de respuestas neutrales.
- Nodo 5: Participantes con alta aceptación de la afirmación de que las herramientas algorítmicas son más justas que la intuición humana, presentan el perfil más favorable en relación con su uso predictivo. En este grupo representa el perfil más favorable del modelo, alcanzando un 72.3% en las categorías "De acuerdo" o "Totalmente de acuerdo", aunque es solamente un 5,9%de la muestra.

La segunda variable introducida en el modelo afecta a los participantes con baja aceptación del uso de herramientas algorítmicas (Nodo 1) y corresponde al nivel de aceptación de la afirmación "las herramientas algorítmicas utilizadas en el sistema de justicia son objetivas en sus evaluaciones", con un valor ajustado de $p = .000$ ($\chi^2 = 33.980$, $gl = 1$). Esta bifurcación genera dos nodos:

- Nodo 6: Participantes que no están de acuerdo con la afirmación de que las herramientas algorítmicas son objetivas muestran una actitud fuertemente negativa hacia su uso predictivo, con más de un 75% en las categorías de totalmente desacuerdo y desacuerdo.
- Nodo 7: Participantes que se encuentran más de acuerdo con la afirmación muestran un leve aumento de su aceptación, aunque el desacuerdo y el

totalmente desacuerdo continúa siendo mayoritario (45.2%).

En la rama de aceptación intermedia (Nodo 2), el árbol se ramifica según la variable "las herramientas algorítmicas son más imparciales que los jueces humanos en la toma de decisiones", que también resulta altamente significativa ($p = .000$; $\chi^2 = 32.632$, $gl = 1$), generando tres nuevos nodos:

- Nodo 8: Participantes que no están de acuerdo con la afirmación muestran una actitud de aceptación más crítica, con un 41,5% de respuestas en la categoría de desacuerdo.
- Nodo 9: Participantes con postura neutral ante la valoración del riesgo. En este grupo predomina también la neutralidad respecto a la aceptación general (63.6%).

De este último nodo (nodo 9), se bifurca en dos nuevos nodos en función de la variable "los jueces y abogados deben recibir capacitación sobre el posible uso de herramientas algorítmicas en el sistema de justicia" ($p = .007$; $\chi^2 = 9.742$, $gl = 1$):

- Nodo 12: Participantes que no están de acuerdo con la afirmación muestran una actitud de aceptación neutral, con un 56,1% de respuestas en la categoría ni acuerdo ni desacuerdo.
- Nodo 13: Participantes que se muestran de acuerdo con la afirmación, aunque sigue distribuyéndose mayoritariamente en el grupo neutral, muestran una mayor aceptación (33.3%).

Finalmente, en la rama correspondiente a quienes presentan neutralidad ante la afirmación de "las herramientas algorítmicas son más justas en comparación con la intuición humana" (Nodo 3), la variable predictora que introduce una nueva división si están familiarizados con las herramientas algorítmicas, con un valor de $p = .010$ ($\chi^2 = 6.670$, $gl = 1$):

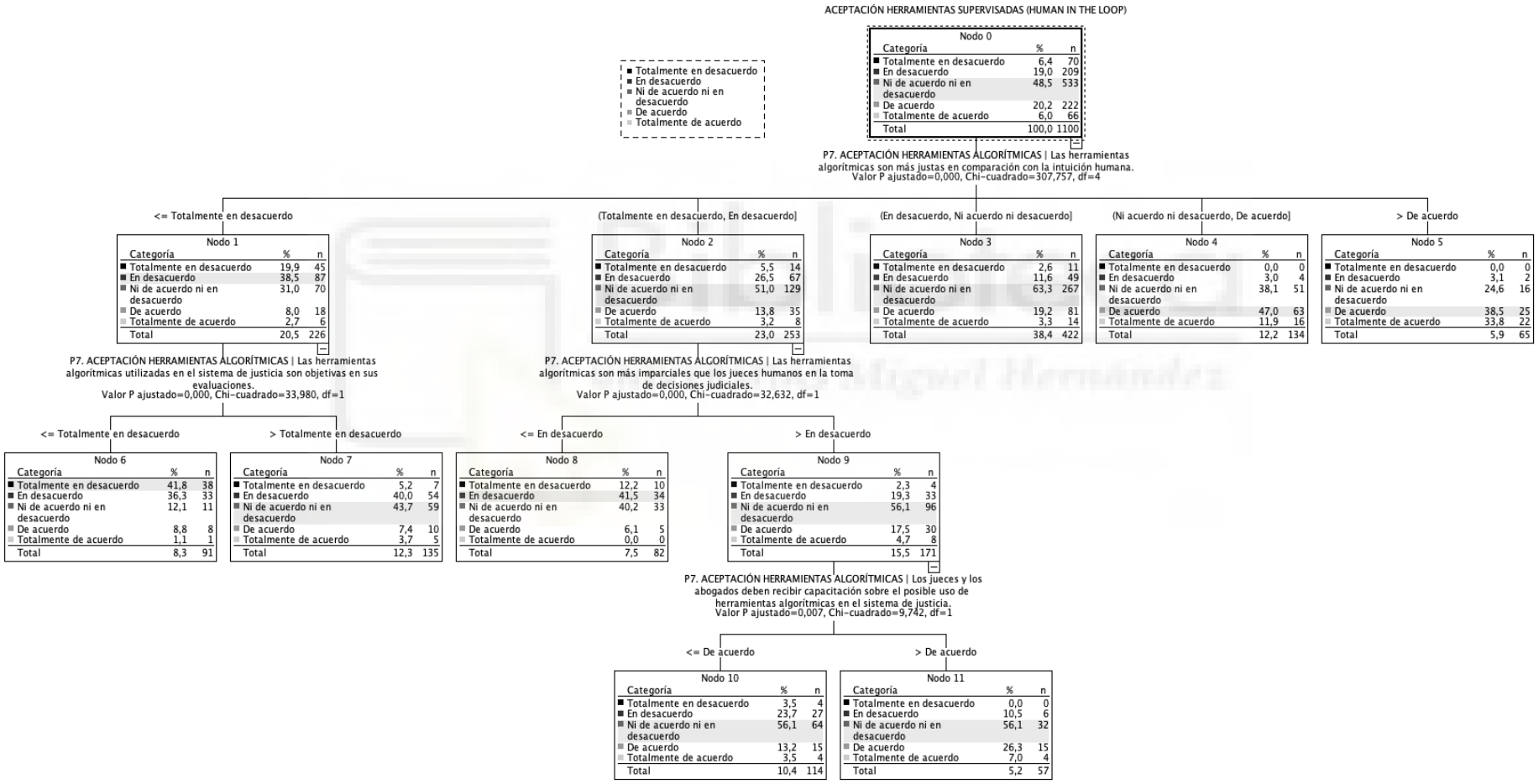
- Nodo 10: Participantes no familiarizados con herramientas algorítmicas mantienen una actitud general positiva, aunque con una mayor tendencia a respuestas neutrales 64,3% en las categorías de acuerdo).
- Nodo 11: Participantes familiarizados con algoritmos jurídico muestran una

mayor aceptación, un 36,3% en las categorías “De acuerdo” o “Totalmente de acuerdo”, aunque predomina la neutralidad (56,4%).

Estos resultados se muestran en la Figuera 13:

Figura 13.

Árbol de decisión de las variables de actitudes de la aceptación de herramientas algorítmicas con supervisión humana – (HITL).

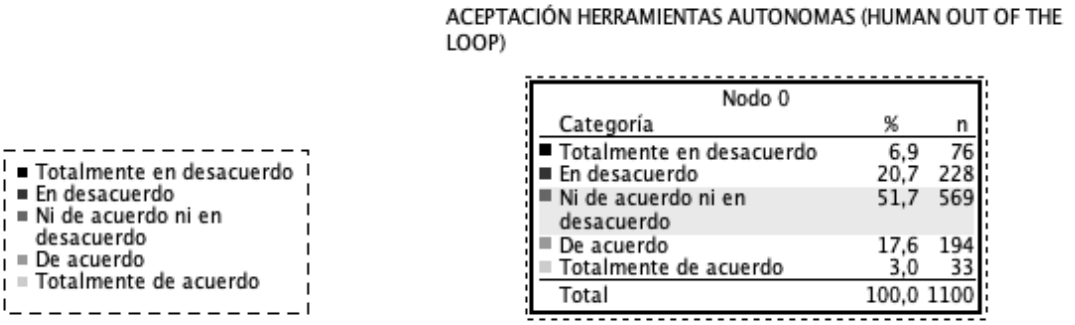


4.4.3. Variables sociodemográficas y aceptación de herramientas algorítmicas autónomas.

Los resultados del análisis entre variables sociodemográficas y la aceptación de herramientas algorítmicas autónomas indica que ninguna de las variables independientes incluidas en el modelo presentó una asociación estadísticamente significativa con la variable dependiente, de acuerdo con los criterios del algoritmo CHAID. Es decir, no se identificaron divisiones que permitieran mejorar la clasificación de los casos en función de las respuestas de aceptación.

La ausencia de ramificaciones en el árbol puede deberse a diferentes factores, como la falta de relación significativa entre las variables analizadas o una distribución homogénea de las respuestas. En este caso, la distribución general de las respuestas mostró una elevada concentración en la categoría "ni de acuerdo ni en desacuerdo" (51,7 %), lo cual podría haber limitado la capacidad del modelo para detectar patrones diferenciadores. Así se muestra en la Figura 14:

Figura 14. Árbol de decisión de las variables sociodemográficas y la aceptación de herramientas algorítmicas con autónomas (HOTL).



4.4.4. Actitudes y aceptación de herramientas algorítmicas autónomas.

Finalmente, se analizó la aceptación general de herramientas algorítmicas autónomas en el ámbito penal en función de variables actitudinales. En este caso, el árbol nuestra que la aceptación de herramientas autónomas en el sistema de justicia (Human out of the loop) depende principalmente de las creencias previas sobre la justicia, objetividad e imparcialidad de los algoritmos. La primera división se produce según el grado en que las personas piensan que "las herramientas algorítmicas son más justas que la intuición humana". Esta variable funciona como el núcleo estructural de la actitud hacia la autonomía algorítmica: quienes no están de acuerdo o están totalmente en desacuerdo conforman un grupo mayoritario con niveles de aceptación bajos, lo que confirma que la percepción de justicia es un prerrequisito fundamental para confiar en un sistema automatizado sin intervención humana. Entre quienes sí consideran que los algoritmos pueden ser más justos que la intuición humana, el árbol muestra subdivisiones que revelan matices importantes. En las ramas donde predominan posiciones neutrales o poco favorables, emergen nuevas divisiones basadas en la creencia de que "los algoritmos son más objetivos en sus evaluaciones". Esta segunda capa de percepciones funciona como un modulador significativo: incluso entre los perfiles inicialmente escépticos, aquellos que reconocen cierto grado de objetividad algorítmica tienden a exhibir mayores niveles de aceptación, lo que evidencia que la confianza en las propiedades técnicas del algoritmo puede compensar parcialmente la falta de confianza inicial. En las ramas donde existe ya una predisposición más favorable, la aceptación aumenta todavía más cuando las personas están "de acuerdo en que jueces y abogados deberían recibir capacitación sobre el uso de estas herramientas". Este patrón sugiere que, aun cuando se reconoce la calidad técnica del algoritmo, la confianza en una implementación responsable, supervisada y profesionalizada actúa como un tercer pilar que fortalece la disposición a aceptar sistemas autónomos en procesos judiciales. Esto indica que la aceptación no es dicotómica, sino el resultado acumulativo de tres pilares: (1) justicia, (2) objetividad e imparcialidad y (3) confianza en la capacitación humana para la correcta implementación. Estos tres elementos interactúan y determinan la predisposición a permitir que los algoritmos operen sin intervención humana en el sistema judicial.

El nodo raíz (Nodo 0), que agrupa la totalidad de la muestra (N = 1100), muestra que la mayoría de los participantes se sitúan en una posición neutral respecto a la aceptación de herramientas algorítmicas en este ámbito ("Ni de acuerdo ni en desacuerdo", 51,7%). La primera variable predictora que divide el árbol es la afirmación "las herramientas algorítmicas son más justas que la intuición humana", la cual resulta altamente significativa con un valor ajustado de p = .000 (χ^2 = 375.438, gl = 4). Esta variable origina tres ramas principales:

- Nodo 1: Participantes que están totalmente en desacuerdo con la afirmación muestran una menor aceptación hacia las herramientas algorítmicas autónomas, concentrándose en las categorías de "En desacuerdo" (38,5%) o "Totalmente en desacuerdo" (21,7%).
- Nodo 2: Participantes con baja grado de acuerdo con la afirmación se muestran neutrales respecto a su uso general (51,8%).
- Nodo 3: Participantes con aceptación intermedia o neutral con la afirmación, presenta actitudes más variadas, con predominio de la neutralidad (69,4) y ciertas inclinaciones tanto hacia el acuerdo (14,4%) como hacia el desacuerdo (13%).
- Nodo 4: Participantes con aceptación moderada de la afirmación manifiestan una mayor aceptación (50,7% de acuerdo) hacia el uso de herramientas algorítmicas autónomas, aunque todavía existe una proporción importante de respuestas neutrales (42,5%).
- Nodo 5: Participantes con alta aceptación de la afirmación de que las herramientas algorítmicas son más justas que la intuición humana, presentan el perfil más favorable en relación con su uso predictivo. Del mismo modo que en el árbol de aceptación de herramientas algorítmicas supervisadas, este grupo representa el perfil más favorable del modelo, alcanzando un 72.3% en las categorías "De acuerdo" o "Totalmente de acuerdo", aunque es solamente un 5,9%de la muestra.

En el Nodo 1, la variable que genera una nueva bifurcación es la afirmación "las herramientas algorítmicas utilizadas en el sistema de justicia son objetivas en sus evaluaciones", con un valor ajustado de p = .000 (χ^2 = 26.756, gl = 1). Esta división

genera dos nodos adicionales:

- Nodo 6: Participantes que no están de acuerdo con la afirmación presentan una actitud general claramente negativa hacia la IA en este ámbito (37,4% en desacuerdo y 39,6% totalmente en desacuerdo).
- Nodo 7: Participantes que se muestran más de acuerdo con la afirmación muestran una actitud neutral especto a su aceptación (42,2%)

En el Nodo 2, el árbol se ramifica a partir de la variable de que "las herramientas algorítmicas son más imparciales que los jueces humanos en la toma de decisiones" con p = .000 (χ^2 = 22.021, gl = 1). De esta división surgen dos perfiles diferenciados:

- Nodo 8: Participantes que no están de acuerdo con la afirmación presentan un mayor rechazo (47,6% en desacuerdo).
- Nodo 9: Participantes que consideran que las HA son más imparciales que los jueces se muestran neutrales en su aceptación (59,1%).

Finalmente, en el Nodo 9, correspondiente a quienes creen que las herramientas algorítmicas son más justas que la intuición humana, la variable predictora que introduce una nueva división es "los jueces y abogados deben recibir capacitación sobre el posible uso de HA en el sistema de justicia", con p = .002 (χ^2 = 12.008, gl=1):

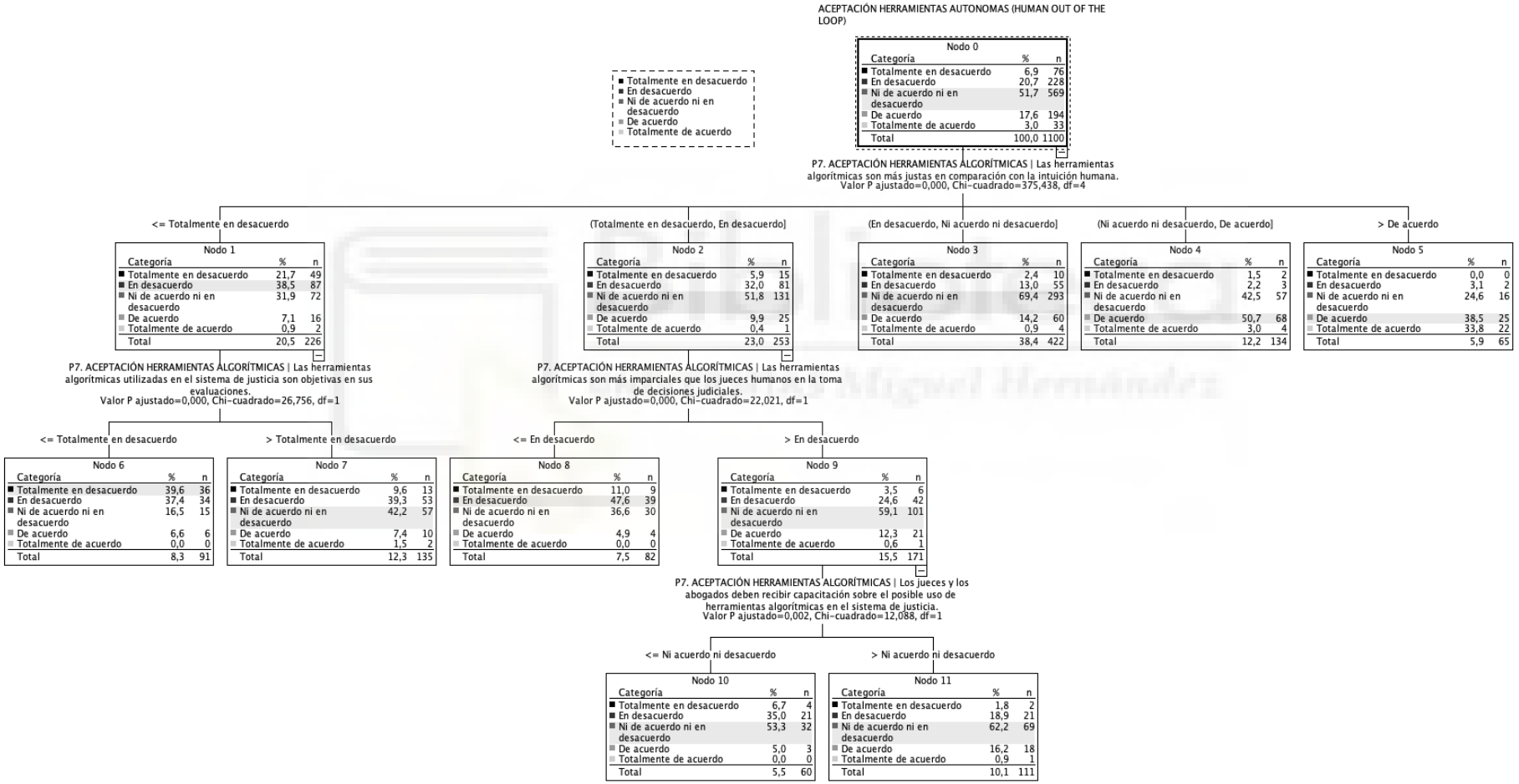
- Nodo 10: Participantes que no se muestran de acuerdo con que los profesionales deban recibir formación mantienen una postura neutral generalizada (53,3%).
- Nodo 11: Participantes que consideran que debe formarse a los profesionales presentan una predisposición neutral (62,2%), aunque con una mayor proporción en las posiciones de aceptar las HA (16,2%).

Este último modelo sobre la aceptación de herramientas algorítmicas autónomas sugiere que la aceptación de herramientas algorítmicas autónomas en el ámbito penal está fuertemente determinada por la percepción de imparcialidad y objetividad atribuida a estas tecnologías. Asimismo, la actitud frente a la formación de operadores jurídicos aparece como un factor modulador relevante entre quienes presentan predisposición a su aceptación. El perfil más favorable se configura en

torno a quienes creen en la justicia algorítmica, aceptan su imparcialidad y valoran la formación profesional en esta materia. Estos resultados se exponen en la Figura 15:

Figura 15.

Árbol de decisión de las variables actitudinales de la aceptación de herramientas algorítmicas autónomas – (HOTL).



5. Discusión y conclusiones.

Los resultados obtenidos en el presente estudio permiten aportar evidencia empírica al debate sobre la aceptación social de las herramientas algorítmicas aplicadas a la justicia penal. A partir de las hipótesis planteadas, se observa un patrón complejo en el que las actitudes hacia la inteligencia artificial judicial están mediadas por percepciones de imparcialidad, objetividad y justicia algorítmica, mientras que los factores sociodemográficos presentan un papel marginal. Este apartado discute los hallazgos en relación con cada hipótesis, los contrasta con la literatura previa, identifica limitaciones metodológicas y plantea las principales implicaciones prácticas.

La primera hipótesis planteaba que la percepción de imparcialidad en las decisiones emitidas con apoyo de herramientas algorítmicas, ya fueran supervisadas o autónomas, se asociaría positivamente con actitudes favorables hacia su utilización en el sistema judicial. Los resultados confirman de manera robusta esta afirmación. Tanto los modelos de regresión como los árboles de decisión muestran que la creencia en la imparcialidad y objetividad de la IA constituye uno de los predictores más relevantes de la aceptación, especialmente en los niveles de acuerdo más elevados. Estos hallazgos coinciden con investigaciones previas que han destacado que la percepción de imparcialidad es un factor central para legitimar el uso de sistemas automatizados en justicia (Wang, 2023; Završnik, 2020). De hecho, estudios experimentales con jueces y población general han mostrado que la confianza en los algoritmos aumenta significativamente cuando son percibidos como neutrales en comparación con la intuición humana (Green & Chen, 2019; Yalcin et al., 2023), especialmente en contextos donde los algoritmos se aplican a tareas técnicas o rutinarias, pero generan resistencias cuando las decisiones implican dimensiones éticas o emocionales complejas.

En relación con la segunda hipótesis, que preveía una asociación positiva entre la creencia de que las herramientas algorítmicas pueden reducir los errores judiciales humanos y su aceptación social, los datos la confirman parcialmente. La percepción de que los algoritmos son más justos y objetivos que los jueces o que la intuición

humana se relaciona con mayores niveles de aceptación, especialmente en el caso de las herramientas autónomas. Coincidiendo con estos resultados, Choung, David y Ross (2022) evidencian que la confianza no opera de forma directa, sino mediando en la percepción de utilidad y justicia, distinción que permite comprender por qué la percepción de imparcialidad puede resultar más influyente que la de simple eficacia.

La tercera hipótesis, que establecía una relación negativa entre la percepción de sesgo algorítmico y la aceptación social, recibe un apoyo más débil en los datos. Aunque los análisis de correlación muestran asociaciones negativas entre estas percepciones y la aceptación de la IA, los modelos de regresión no identifican efectos significativos. Ello sugiere que, si bien existe conciencia social sobre la posibilidad de sesgos, esta percepción no constituye un factor determinante en la decisión de aceptar o rechazar las herramientas. Este hallazgo contrasta parcialmente con la literatura crítica, que ha advertido de manera reiterada sobre el riesgo de replicar y amplificar desigualdades sociales en el uso de algoritmos judiciales (Eubanks, 2018; Noble, 2018). Una posible explicación es que la ciudadanía perciba los sesgos como un problema abstracto o técnico, sin una traslación inmediata a su propia experiencia, lo que reduce su impacto en la aceptación práctica. En esta línea, Kelly, Kaye y Oviedo-Trespalacios (2023) destacan que la resistencia hacia la IA en justicia no responde únicamente a características técnicas, sino al valor simbólico y relacional del contacto humano en el proceso judicial.

En cuanto a la cuarta hipótesis, que proponía una asociación positiva entre el nivel de conocimientos tecnológicos y la aceptación, los resultados no confirman esta relación en el caso de las herramientas autónomas, pero sí muestran un efecto limitado en las herramientas supervisadas. Los participantes con familiarización previa con algoritmos manifiestan una mayor disposición a aceptar la IA cuando esta se encuentra bajo control humano, lo que sugiere que la experiencia tecnológica incrementa la confianza en escenarios donde la autonomía algorítmica está limitada. Estos resultados se alinean con estudios sobre adopción tecnológica que resaltan la relevancia del conocimiento previo y la autoeficacia tecnológica para reducir la llamada "aversión algorítmica" (Dietvorst, Simmons & Massey, 2015; Lin & Hsieh,

2007). Además, investigaciones recientes subrayan que factores como la alfabetización en IA (Schiavo, Businaro & Zancanaro, 2024) y la percepción de competencia tecnológica (Bergdahl et al., 2023) influyen positivamente en la disposición a utilizar estas herramientas, especialmente si se perciben como respetuosas con la autonomía del usuario.

La quinta hipótesis, que planteaba una influencia significativa de factores sociodemográficos como el sexo, la edad o la familiaridad previa en la aceptación de la IA judicial, encuentra un apoyo parcial. Si bien las variables sociodemográficas no resultan significativas en los modelos de regresión para herramientas autónomas, los árboles de decisión sí muestran ciertos patrones diferenciados en el caso de la supervisión humana. En particular, las mujeres jóvenes con experiencia en algoritmos jurídicos constituyen el perfil más favorable a la aceptación. No obstante, en términos generales, el peso de las variables sociodemográficas es menor en comparación con las creencias sobre imparcialidad y justicia algorítmica. Estos hallazgos son consistentes con estudios que destacan la importancia de factores actitudinales y contextuales por encima de las características demográficas en la aceptación de tecnologías complejas (Kozak & Fel, 2024; Zouridis, Van Eck & Bovens, 2020).

Los resultados de este estudio refuerzan que la supervisión humana es un componente esencial para la legitimidad social de los algoritmos judiciales, y muestran además que la creencia en su objetividad, imparcialidad y mayor justicia respecto a los operadores humanos constituye el principal predictor de actitudes favorables hacia su uso. Este patrón coincide con la literatura que subraya que la legitimidad de la justicia algorítmica depende de su capacidad para ser explicada, auditada y controlada (Zalnieriute et al., 2019; Doshi-Velez & Kim, 2017). Al mismo tiempo, los hallazgos aportan un matiz al debate sobre los sesgos algorítmicos. Aunque la literatura crítica destaca su gravedad (Angwin et al., 2016), su influencia en la percepción ciudadana parece menos determinante de lo que suele asumirse. Esta divergencia sugiere un desfase entre el debate académico y la conciencia pública acerca del alcance de estos sesgos, lo que a su vez pone de relieve la necesidad de fortalecer la alfabetización digital y jurídica. Finalmente, la escasa

relevancia de los factores sociodemográficos en la mayoría de los modelos indica que la aceptación de la IA judicial no se encuentra segmentada de forma consistente por variables como edad o sexo, sino que se ancla principalmente en creencias cognitivas compartidas. Si bien esto contrasta con algunos estudios clásicos sobre adopción tecnológica que sí identifican diferencias generacionales (Venkatesh et al., 2003), se alinea con investigaciones recientes que apuntan hacia un efecto decreciente de estas variables en tecnologías ya consolidadas (Straub, 2021).

Desde una perspectiva teórica, el estudio contribuye a la literatura sobre justicia algorítmica al evidenciar que las actitudes hacia la IA no dependen tanto de características sociodemográficas como de percepciones cognitivas y normativas sobre imparcialidad, objetividad y legitimidad. De este modo, refuerza los planteamientos de la Teoría del Comportamiento Planeado (Ajzen, 1991) y el Modelo de Aceptación Tecnológica (Davis, 1989), que otorgan un papel central a las actitudes y creencias en los procesos de adopción tecnológica. En el plano práctico, los resultados sugieren que la implementación de IA en justicia penal debe orientarse hacia modelos de apoyo al juicio humano, con mecanismos de supervisión robustos, transparencia en el funcionamiento y formación especializada para los operadores jurídicos. Estas medidas no solo incrementarían la aceptación social, sino que también contribuirían a garantizar un uso ético y legítimo de estas tecnologías en línea con los estándares europeos (European Commission, 2021).

Finalmente, aunque los esfuerzos gubernamentales recientes se han centrado de manera prioritaria en el diseño de marcos legales, estrategias regulatorias y lineamientos éticos para el uso de la inteligencia artificial, los resultados de este estudio sugieren que la mera existencia de normas no incrementa la aceptación social de las herramientas algorítmicas. La regulación aparece como una condición necesaria, pero claramente insuficiente. En cambio, la aceptación se relaciona de forma más estrecha con creencias sustantivas sobre la objetividad, justicia e imparcialidad de estas herramientas, así como con la convicción de que su uso debe ir acompañado de formación específica y alfabetización técnica entre los profesionales del sistema judicial. Por ello, las políticas públicas deberían complementar el enfoque regulatorio con estrategias orientadas a fortalecer el

conocimiento, la comprensión y la capacitación práctica, de modo que la ciudadanía pueda evaluar el valor añadido real de la inteligencia artificial en justicia y desarrollar una confianza informada en su funcionamiento.

Estudio 3. La aceptación social del uso de inteligencia artificial en la toma de decisiones judiciales y penitenciarias: un estudio experimental.

1. Justificación.

La incorporación de la inteligencia artificial en los sistemas judiciales y penitenciarios constituye uno de los desafíos más relevantes para las ciencias sociales y jurídicas contemporáneas. En los últimos años, la digitalización y la automatización de la toma de decisiones han pasado de ser herramientas auxiliares para ocupar un papel cada vez más central en procesos judiciales y de gestión penitenciaria (Goodman & Flaxman, 2017; Yeung, 2018). Esta expansión tecnológica plantea preguntas esenciales sobre la aceptación, la transparencia y la equidad de los sistemas de justicia, especialmente cuando las decisiones afectan directamente los derechos fundamentales de las personas (Mittelstadt et al., 2016). Desde la perspectiva criminológica y jurídica, la confianza ciudadana en las decisiones judiciales es un pilar de la legitimidad institucional (Tyler, 2006). Sin embargo, la introducción de herramientas algorítmicas en estos contextos puede alterar la percepción pública sobre la imparcialidad y la humanidad de las decisiones, especialmente cuando los sistemas automatizados adquieren un alto grado de autonomía (Eubanks, 2018; Angwin et al., 2016). Por ello, resulta necesario analizar empíricamente cómo la sociedad percibe y acepta el uso de la IA en decisiones judiciales y penitenciarias, identificando los factores que incrementan o reducen esa aceptación.

En este sentido, el presente estudio se justifica por tres razones fundamentales:

Primero, porque existe un déficit de evidencia empírica sobre las actitudes ciudadanas hacia la inteligencia artificial en el sistema de justicia penal, especialmente en el contexto español. Aunque abundan estudios sobre la eficiencia o precisión técnica de los algoritmos (Hildebrandt, 2020; Selbst et al., 2019), son escasas las investigaciones que exploran la dimensión social de su aceptación.

Segundo, porque la aceptación social es un elemento indispensable para la sostenibilidad de cualquier innovación tecnológica en el sector público. La literatura muestra que la adopción de nuevas herramientas depende no sólo de su eficacia,

sino también de su compatibilidad con valores democráticos como la transparencia, la rendición de cuentas y la equidad (Lepri et al., 2018; Zuboff, 2019).

Tercero, porque el estudio permite profundizar en el papel del factor humano dentro del proceso de toma de decisiones en el sistema de justicia penal, y determinar cuáles son las preferencias ciudadanas y profesionales acerca de en qué momento del proceso debe (o no) introducirse (paradigma *Human-X-the-Loop)* (Rahwan et al., 2019).

Este estudio ofrece evidencia empírica sobre cómo los ciudadanos valoran el grado de autonomía de las herramientas y la proporcionalidad de sus resultados, aportando una base científica para el diseño de políticas públicas y marcos regulatorios, que hasta ahora entienden al factor humano como un elemento de supervisión. Además, sus hallazgos enriquecen el debate académico sobre la gobernanza algorítmica y el paradigma Human-X-the-Loop, permitiendo analizar el modelo de justicia digital que la sociedad busca consolidar.

2. Objetivos.

El presente estudio tiene como objetivo evaluar la aceptación social del uso de herramientas algorítmicas aplicadas en los ámbitos judicial y penitenciario. En particular, se examina de qué manera los diferentes niveles de autonomía de la inteligencia artificial y la adecuación de las sanciones propuestas con el código penal influyen en el grado de aceptación de estas.

Con el fin de alcanzar este objetivo general, se han definido los siguientes objetivos específicos:

OE1. Analizar la influencia que ejerce el nivel de autonomía de las herramientas algorítmicas en la aceptación de sus decisiones dentro del ámbito judicial y penitenciario.

OE2. Analizar el impacto que tiene la adecuación de las sanciones propuestas por las herramientas algorítmicas en la aceptación en los contextos judicial y penitenciario.

A partir de estos objetivos se formulan las siguientes hipótesis de investigación:

H1. La aceptación de las decisiones disminuye a medida que aumenta el grado de autonomía en el proceso.

H2. La aceptación de las decisiones es menor cuando las sanciones propuestas resultan inadecuadas al caso.

H3. La aceptación social de las herramientas algorítmicas disminuye conforme aumenta su autonomía en el proceso.

H4. La aceptación social es menor cuando las sanciones propuestas resultan inadecuadas al caso.

3. Metodología.

3.1. Diseño del experimento y variables.

Para la realización de presente estudió se decidió emplear una metodología experimental con el objetivo de analizar la aceptación social del uso de herramientas algorítmicas en el ámbito judicial y penitenciario. En concreto, se evaluó cómo el nivel de autonomía de la inteligencia artificial, la coherencia de las sanciones propuestas y el perfil de quienes diseñan dichas herramientas influyen en la percepción de legitimidad y aceptación por parte de la ciudadanía. Para ello, se empleó la técnica de casos-escenario (vignette methodology), utilizada en ciencias sociales para investigar juicios morales y legales en contextos simulados (Aguinis & Bradley, 2014; Bieneck, 2009). Esta metodología permitió presentar a los participantes situaciones hipotéticas estandarizadas, en las que se manipulaban sistemáticamente las condiciones experimentales, preservando la validez ecológica del diseño.

En total, se elaboraron 10 escenarios experimentales, resultantes de la combinación de dos factores principales: (1) el grado de intervención humana en la decisión, que oscilaba desde una resolución adoptada exclusivamente por un juez hasta un modelo de inteligencia artificial completamente autónomo, pasando por sistemas algorítmicos con diferentes niveles de supervisión y origen de diseño; y (2) la proporcionalidad de la sanción propuesta, clasificada como adecuada (proporcional con el marco penal y las características del caso) o inadecuada (desproporcional respecto al marco penal o a la situación presentada).

De esta forma, se configuró un diseño factorial 5 × 2, en el que los cinco niveles de autonomía se combinaron con dos niveles de proporcionalidad, dando lugar a diez grupos experimentales (G1–G10). Cada participante fue asignado aleatoriamente a uno solo de los escenarios, con el objetivo de evitar sesgos derivados de la exposición múltiple y garantizar la independencia de las observaciones.

Las condiciones experimentales se distribuyeron de la siguiente manera:

- G1 y G2: decisiones adoptadas por una herramienta de IA completamente

autónoma diseñada por expertos humanos.

- G3 y G4: decisiones tomadas por una herramienta de IA autónoma, desarrollada exclusivamente a partir de modelos matemáticos, sin intervención directa de expertos en su diseño.
- G5 y G6: sistema de IA creado por expertos humanos que no resolvía de manera independiente, sino que ofrecía recomendaciones al operador humano.
- G7 y G8: sistema de IA desarrollado a partir de modelos matemáticos con funcionamiento asistido, proporcionando únicamente apoyo a la decisión final.
- G9 y G10: condiciones de control, en las que las resoluciones fueron tomadas de manera exclusiva por jueces humanos, sin asistencia algorítmica.

En cada uno de estos pares de grupos, los casos se diferenciaron en función de la proporcionalidad de la sanción propuesta. Así, los grupos G1, G3, G5, G7 y G9 representaron escenarios con sanciones proporcionales, es decir, ajustadas y adecuadas a la gravedad del caso planteado y al marco penal aplicable. Por el contrario, los grupos G2, G4, G6, G8 y G10 reflejaron situaciones con sanciones desproporcionadas, caracterizadas por ser inadecuadas en relación con las características específicas del caso simulado.

Para la medición de las variables dependientes se diseñó un cuestionario estructurado ad hoc, que incluía los escenarios experimentales y escalas de evaluación específicas. Las variables analizadas fueron: (1) la aceptación de la decisión algorítmica, entendida como el grado de conformidad del participante con la resolución presentada en el escenario; y (2) la aceptación general del uso de la inteligencia artificial en el ámbito judicial y penitenciario, concebida como un indicador de la actitud global hacia estas tecnologías. Ambas variables se evaluaron mediante escalas tipo Likert de 5 puntos, con opciones de respuesta que oscilaron entre 1 (totalmente en desacuerdo) y 5 (totalmente de acuerdo).

En la Tabla 24 se pueden observar los diferentes grupos experimentales:

Tabla 24.

Distribución de los grupos experimentales en función de las condiciones del estudio.

			SANCIÓN	
			Proporcional	**Desproporcional**
USO DE IA	**TOTAL**	IA DISEÑO HUMANO	G1	G2
		IA DISEÑO MAQUINA	G3	G4
	PARCIAL	IA DISEÑO HUMANO	G5	G6
		IA DISEÑO MAQUINA	G7	G8
	SIN IA	HUMANOS	G9	G10

3.2. Participantes.

Con el fin de establecer el número de participantes requeridos para el estudio, se llevó a cabo un análisis de potencia a priori mediante el software G*Power 3.1 (Faul et al., 2007). El cálculo se efectuó considerando un análisis de varianza (ANOVA) de un factor con diez grupos, asumiendo un tamaño del efecto de f = 0.15, un nivel de significación de $\alpha = .05$ y una potencia estadística del 90% ($1 - \beta = 0.90$). Los resultados indicaron que era necesario contar con una muestra mínima de 890 participantes para detectar diferencias significativas entre los grupos con la potencia especificada. Finalmente, la muestra del estudio estuvo compuesta por 1.100 participantes residentes en España, distribuidos equitativamente entre las condiciones experimentales, asignándose 110 personas aleatoriamente a cada una de las diez condiciones del diseño[13]. La distribución de los grupos fue la siguiente (Tabla 25):

[13] Véase el apartado de muestra en las consideraciones previas del capítulo 6.

Tabla 25.

Distribución de los participantes por sexo y edad en cada grupo experimental.

Casos	Sexo				Edad
	Mujeres		Hombres		
	(%)	n	(%)	n	Media (DT)
G1	51.8	57	48.2	53	43.16 (12.20)
G2	50	55	50	55	42.67 (11.25)
G3	55.5	61	44.5	49	43.02 (11.70)
G4	53.6	59	46.4	51	42.62 (11.70)
G5	49.1	54	50.9	56	42.65 (12.43)
G6	54.5	60	45.5	50	42.92 (13.37)
G7	56.4	62	43.6	48	44.66 (11.45)
G8	51.8	57	48.2	53	43.28 (12.92)
G9	47.3	52	52.7	58	43.72 (12.12)
G10	52.7	58	47.3	52	42.96 (11.45)

3.3. Análisis de datos.

El análisis de datos se llevó a cabo combinando el uso de IBM SPSS Statistics (versión 29) y RStudio (versión 2024.12), en función de la naturaleza de las variables y de los supuestos estadísticos asociados. El procedimiento analítico incluyó tanto análisis descriptivos como contrastes de hipótesis no paramétricos y pruebas post-hoc, con el objetivo de evaluar las diferencias en los niveles de aceptación social de decisiones judiciales y penitenciarias bajo distintos escenarios experimentales.

En una primera fase, se efectuaron análisis descriptivos de las variables dependientes en cada una de las condiciones experimentales. Para ello se calcularon medidas de tendencia central (media y mediana) y de dispersión (desviación típica y rango intercuartílico). Asimismo, los resultados se representaron gráficamente mediante diagramas de caja y gráficos de línea que reflejaban las medias agrupadas por condición experimental.

A continuación, se comprobó si las variables cumplían los supuestos básicos necesarios para la aplicación de pruebas paramétricas, concretamente los de normalidad, homogeneidad e independencia, dado que es la misma muestra para todos los estudios, se observó que las variables no cumplían adecuadamente los

supuestos requeridos para la aplicación de técnicas paramétricas. Por este motivo, se optó por emplear análisis no paramétricos equivalentes, más apropiados para este tipo de datos y para las características de la muestra.

Para comparar los diez grupos experimentales, derivados de la combinación de los niveles de automatización y la proporcionalidad de la sanción, se aplicó la prueba de Kruskal–Wallis, considerada la alternativa no paramétrica al ANOVA de un factor. Este análisis se ejecutó tanto en SPSS como en R. En una segunda fase, con el fin de explorar diferencias específicas entre grupos, se utilizó la prueba de Dunn con corrección de Bonferroni para las comparaciones post-hoc, empleando el paquete *dunn.test* en R. Este procedimiento permitió identificar diferencias pareadas entre condiciones experimentales, controlando el error tipo I asociado a la realización de múltiples contrastes.

Finalmente, se calcularon medidas de tamaño del efecto aproximadas (ε^2) a partir del estadístico de Kruskal–Wallis, con el objetivo de valorar la magnitud de las diferencias encontradas y complementar así la interpretación de los resultados inferenciales.

4. Resultados.

Los resultados se presentan de manera diferenciada según los tres ámbitos de análisis: prisiones, judicial y aceptación general del uso de herramientas. En cada uno de ellos se describen, en primer lugar, los valores obtenidos en las medidas de aceptación y su distribución estadística. Posteriormente, se incluyen los resultados de las pruebas estadísticas aplicadas para detectar diferencias entre grupos experimentales, junto con los tamaños del efecto correspondientes.

4.1. Aceptación social en el ámbito penitenciario.

En el contexto penitenciario, las valoraciones de aceptación de las decisiones se situaron alrededor del punto medio de la escala (entre 2,63 y 3,15 en una escala de 1 a 5) (Figura 16 y 17). Esto significa que, en general, los participantes no se mostraron ni a favor ni en contra de las decisiones: las respuestas se ubican en una posición neutral. Al analizar las respuestas de los distintos grupos experimentales (G1 a G10), se observó que la mayoría coincidió en esta tendencia: la mediana se mantuvo en 3, lo que indica una posición intermedia constante, y la dispersión de respuestas fue bastante similar en todos los casos.

Figura 16.

Distribución de las puntuaciones medias en el ámbito de prisiones para cada una de las condiciones experimentales.

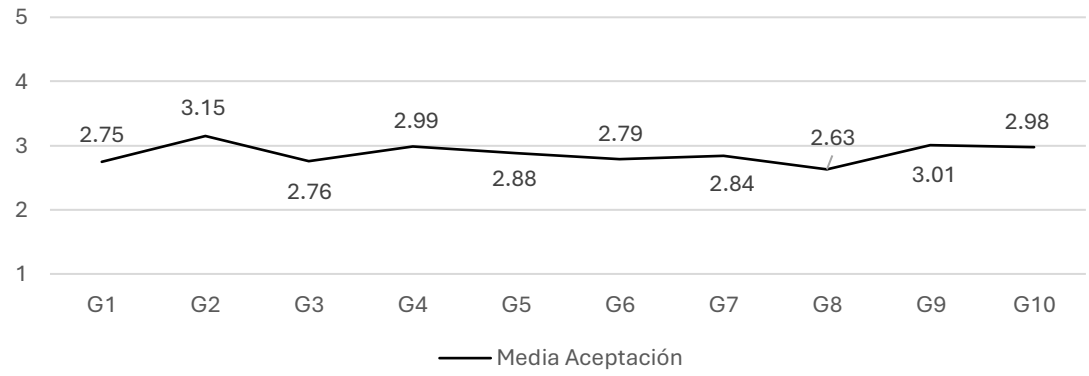
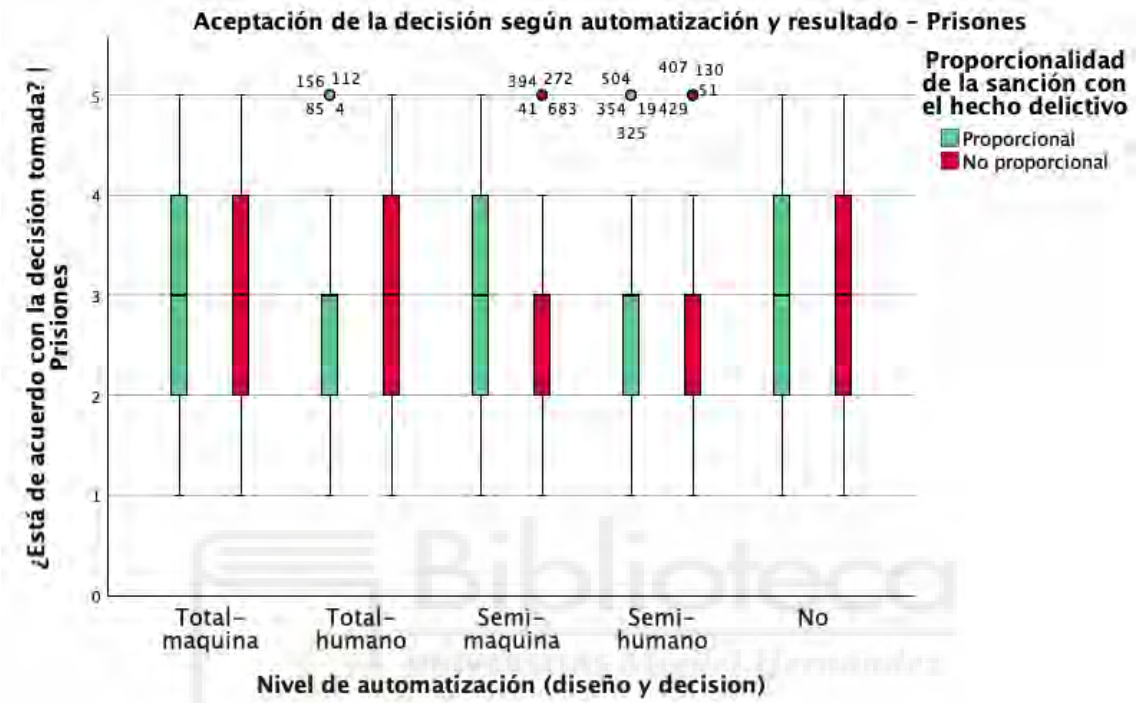


Figura 17.

Aceptación de la decisión en el ámbito penitenciario en función del nivel de automatización y la proporcionalidad de la sanción.



Dado que los datos no cumplían el supuesto de normalidad, se aplicó la prueba no paramétrica de Kruskal–Wallis para la comparación de los diez grupos experimentales. Los resultados no mostraron diferencias estadísticamente significativas en el nivel de aceptación entre los grupos ($\chi^2 = 16.46$, gl = 9, p = .058). Sin embargo, considerando la proximidad del valor al umbral de significación, se llevaron a cabo comparaciones post hoc mediante la prueba de Dunn con corrección de Bonferroni. Estas comparaciones no evidenciaron diferencias significativas en la mayoría de los casos (p > .05), excepto en la comparación entre el grupo G2 (*decisiones adoptadas por una herramienta de IA completamente autónoma diseñada por expertos humanos, con sanción desproporcionada*) que mostró una diferencia con el grupo G8 (*decisiones tomadas por un sistema de IA desarrollado a partir de modelos matemáticos con funcionamiento asistido, proporcionando únicamente apoyo a la decisión final, con sanción desproporcionada*), donde se obtuvo un valor p ajustado de .024. El tamaño del efecto se calculó a partir del estadístico de Kruskal–Wallis

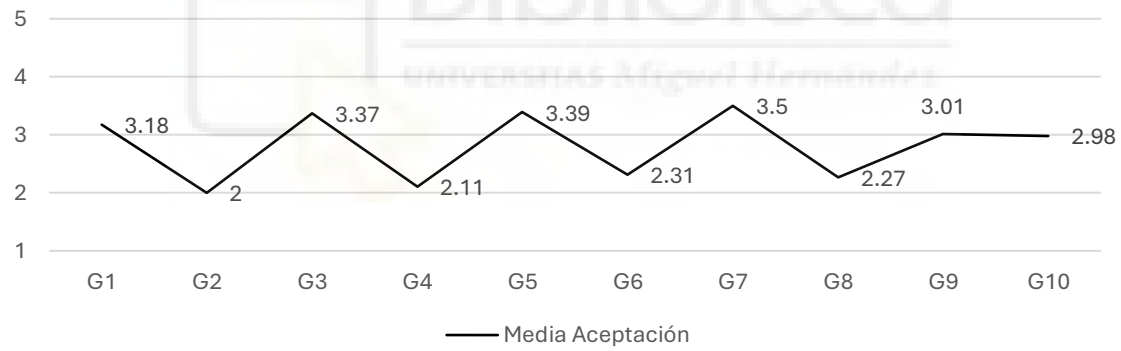
utilizando la fórmula de epsilon cuadrado (ε^2). El valor obtenido fue $\varepsilon^2 = 0.007$, lo que corresponde a un efecto de magnitud pequeña y refleja la reducida variabilidad entre los grupos en el nivel de aceptación.

4.2. Aceptación social en el ámbito judicial.

En el ámbito judicial, la Figura 18 presenta las medias de aceptación de la decisión tomada en función de los diez grupos experimentales, las puntuaciones medias oscilan entre 2,10 y 3,50, situándose en general próximas a la neutralidad. Esta variabilidad moderada refleja diferencias en la valoración de las decisiones según las condiciones experimentales, aunque sin alejarse significativamente de una posición de neutralidad.

Figura 18.

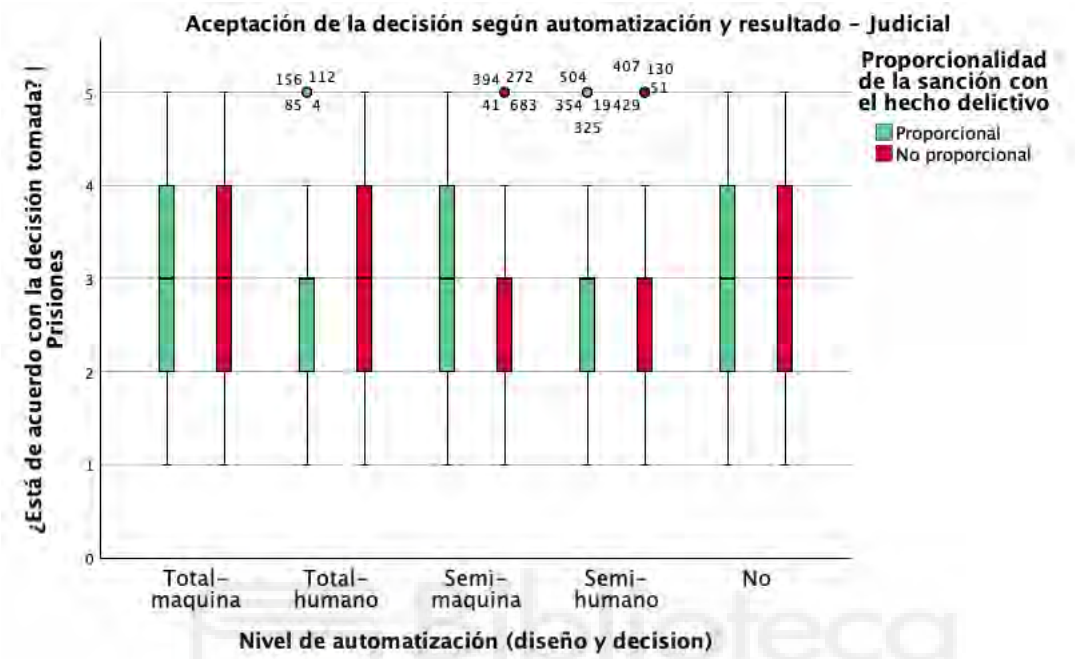
Media de aceptación de la decisión en el ámbito judicial por grupo experimental.



La Figura 18 permite observar de manera más detallada la distribución de respuestas en función de las variables independientes. Los resultados indican que las decisiones percibidas como proporcionales al hecho delictivo reciben, en todos los niveles de automatización, puntuaciones de aceptación más altas que aquellas consideradas desproporcionales. Las decisiones proporcionales tienden a situarse entre los valores 3 y 4, mientras que las desproporcionales se mantienen alrededor de 2 a 3.

Figura 19.

Aceptación de la decisión en el ámbito judicial según nivel de automatización y proporcionalidad de la sanción.



Los resultados obtenidos de la prueba no paramétrica de Kruskal–Wallis muestran un efecto significativo del grupo sobre el nivel de aceptación de la decisión, (χ^2 = 216.68, gl = 9, p < .001), lo que indica que al menos un grupo difiere significativamente del resto en términos de su valoración de justicia o acuerdo con la decisión tomada. Para identificar qué grupos presentaban diferencias específicas, se realizaron comparaciones múltiples utilizando el procedimiento de corrección de Bonferroni. Los resultados revelaron diferencias estadísticamente significativas en múltiples combinaciones de grupos, especialmente entre aquellos con distintas condiciones de proporcionalidad de la sanción y niveles de automatización. En concreto:

Tabla 26.

Comparaciones post-hoc entre grupos experimentales en el ámbito judicial.

Casos	χ^2	ES	P
G1 vs G10	5.22	.154	< .001
G1 vs G2	6.07	.166	< .001
G1 vs G4	5.54	.159	< .001
G1 vs G6	4.44	.142	< .001
G1 vs G8	4.62	.145	< .001
G2 vs G3	-7.05	.179	< .001
G2 vs G5	-7.09	.180	< .001
G2 vs G7	-7.63	.186	< .001
G2 vs G9	-8.70	.199	< .001
G3 vs G4	6.51	.172	< .001
G3 vs G6	5.42	.157	< .001
G3 vs G8	5.60	.160	< .001
G3 vs G10	-6.19	.168	< .001
G4 vs G5	-6.56	.173	< .001
G4 vs G7	-7.09	.180	< .001
G4 vs G9	-8.17	.193	< .001
G5 vs G6	5.47	.158	< .001
G5 vs G8	5.64	.160	< .001
G5 vs G10	-6.24	.168	< .001
G6 vs G7	-6.00	.165	< .001
G6 vs G9	-7.07	.179	< .001
G7 vs G8	6.18	.168	< .001
G7 vs G10	-6.77	.175	< .001
G8 vs G9	-7.25	.182	< .001
G9 vs G10	-7.85	.189	< .001

El tamaño del efecto, calculado mediante epsilon cuadrado (ε^2), fue de 0.191, lo que indica que aproximadamente el 19 % de la variabilidad en las puntuaciones de aceptación se explicó por las condiciones experimentales consideradas, correspondientes al nivel de automatización y a la congruencia de la sanción.

4.3. Aceptación social general del uso de nuevas herramientas.

El análisis descriptivo indicó que las puntuaciones de aceptación se mantuvieron estables entre los diferentes grupos experimentales. Las medias se situaron en un rango de 2,42 a 3,34 en una escala de 1 a 5, sin registrarse variaciones relevantes en función del nivel de automatización ni de la congruencia de la sanción. Los diagramas de caja confirmaron esta uniformidad, al mostrar medianas constantes en 3 y rangos intercuartílicos semejantes en todos los grupos. La presencia de valores atípicos fue aislada y no modificó la tendencia general observada (Figura 20 y 21).

Figura 20.

Media de aceptación general del uso de herramientas por grupo experimental.

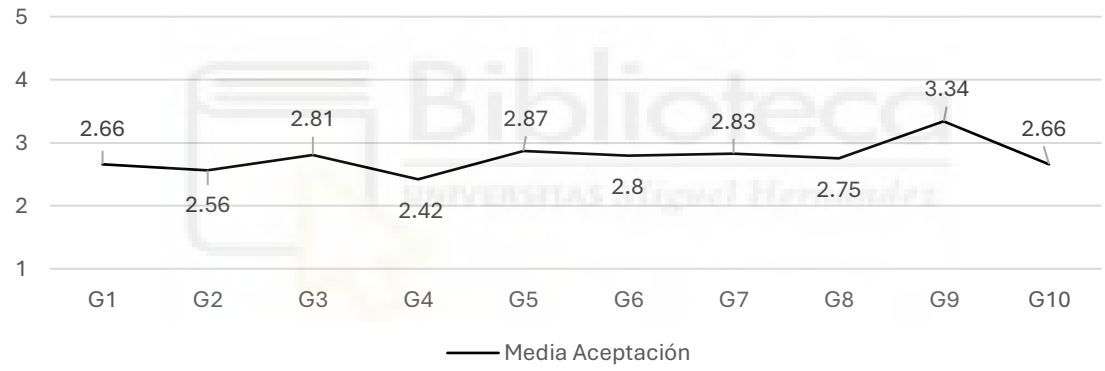
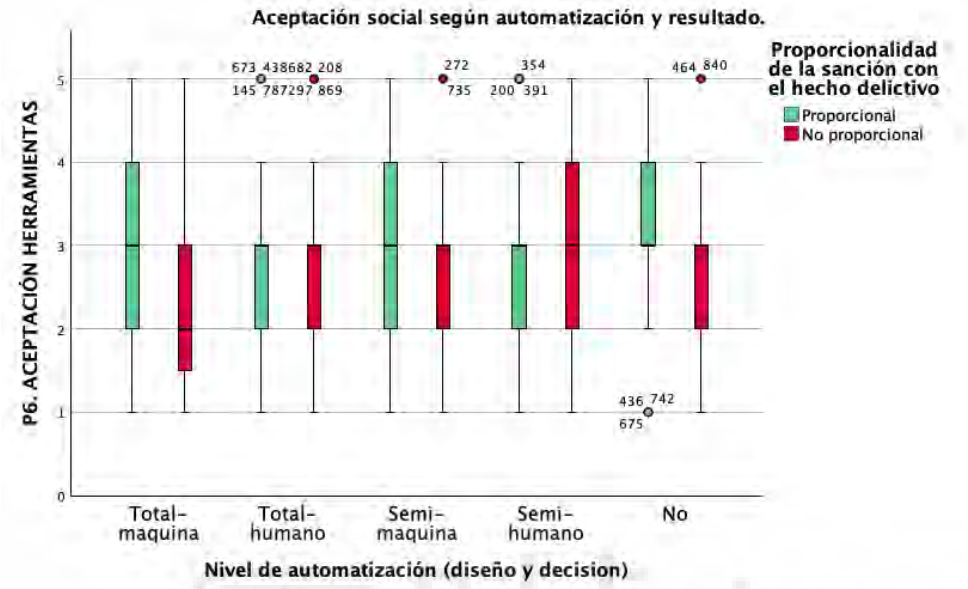


Figura 21.

Aceptación social del uso de herramientas según nivel de automatización y proporcionalidad de la sanción.



Con el objetivo de examinar posibles diferencias entre los grupos experimentales en relación con la variable dependiente, se realizó una prueba no paramétrica de Kruskal–Wallis, dado que los datos no cumplían con el supuesto de normalidad. El análisis reveló diferencias estadísticamente significativas entre los diez grupos, χ^2 = 36.64, gl = 9, p < .001, lo que sugiere que al menos uno de los grupos difiere de los demás en sus puntuaciones. Para identificar las diferencias específicas entre grupos, se llevaron a cabo comparaciones múltiples post hoc utilizando el procedimiento de corrección de Bonferroni. Los resultados mostraron que el grupo G9 presentó diferencias estadísticamente significativas con respecto a varios otros grupos (tabla 27):

Tabla 27.

Comparaciones post hoc entre grupos experimentales en la aceptación general del uso de herramientas.

Casos	χ^2	**ES**	**P**
G9 vs G1	-3.83	.131	0.0029
G9 vs G10	-3.96	.134	0.0017
G9 vs G2	-4.50	.143	< .001
G9 vs G4	-3.12	.119	0.0407
G9 vs G8	-3.22	.120	0.0287

El tamaño del efecto (ε^2) fue de .029, lo que representa un efecto pequeño. Esto sugiere que, aunque las diferencias entre grupos alcanzan significación estadística, el impacto real de la manipulación experimental sobre la variable dependiente es limitado.

5. Discusión y conclusiones.

Los resultados del estudio permiten extraer varias conclusiones relevantes sobre la aceptación social del uso de inteligencia artificial en contextos judiciales y penitenciarios, aportando evidencia empírica que dialoga de manera directa con el marco teórico sobre legitimidad institucional, gobernanza algorítmica y el paradigma Human-X-the-Loop. En conjunto, los hallazgos muestran que la ciudadanía mantiene posiciones matizadas y dependientes del contexto frente al uso de herramientas algorítmicas, lo que respalda la necesidad de una incorporación cuidadosa y normativamente guiada de la IA en el sistema de justicia penal.

En primer lugar, los datos indican que la aceptación de decisiones adoptadas en el ámbito penitenciario se mantiene en niveles cercanos a la neutralidad, sin diferencias significativas entre condiciones experimentales. Este patrón sugiere que, en contextos de menor visibilidad pública y donde las decisiones no se perciben como directamente vinculadas a garantías procesales, el grado de autonomía algorítmica no constituye un factor determinante en la valoración ciudadana. En consecuencia, las hipótesis H1 y H2 no encuentran apoyo en el ámbito penitenciario, pues ni el aumento de autonomía ni la desproporcionalidad de la sanción alteran sustancialmente la aceptación.

En este punto, resulta especialmente relevante el comportamiento del Grupo 2, en el que se presentaba una sanción desproporcional (esto es, la denegación de una libertad que sí podría ser concedida). Los participantes aceptaron esta decisión en mayor medida cuando provenía de un algoritmo que cuando era atribuida a un juez humano. Este hallazgo resulta particularmente llamativo si se lo compara con la tradición garantista del derecho penal, especialmente con la fórmula de Blackstone, según la cual "*better that ten guilty persons escape than that one innocent suffer*" (Blackstone, Commentaries on the Laws of England, 1765; véase también su difusión contemporánea en LawInfo, 2023). Esta máxima, pilar del pensamiento penal liberal, expresa la prioridad moral de evitar errores que perjudiquen a inocentes incluso a costa de reducir la eficiencia punitiva (Laudan, 2006; Volokh, 1997). Sin embargo, en este estudio se observa el fenómeno inverso: cuando el perjuicio injusto se atribuye a un algoritmo, la ciudadanía reduce su aversión al

error punitivo, aceptando con mayor facilidad una resolución que, desde una perspectiva blackstoniana, sería difícilmente justificable. Esta mayor tolerancia hacia un "falso positivo" penal cuando proviene de una máquina sugiere que el algoritmo opera como un amortiguador moral que diluye la imputación de injusticia, coherente con investigaciones que muestran que las personas tienden a atribuir menos responsabilidad moral a sistemas automáticos que a agentes humanos (Bryson, 2018; Bigman & Gray, 2018).

En segundo lugar, los resultados obtenidos en el ámbito judicial sí muestran patrones claramente diferenciados y consistentes con el marco teórico. A diferencia de lo observado en el contexto penitenciario, la ciudadanía se muestra mucho más sensible tanto al grado de autonomía algorítmica como a la proporcionalidad de la sanción. En este escenario las decisiones autónomas y especialmente aquellas que resultan desproporcionales reciben niveles significativamente menores de aceptación. Estos resultados respaldan de manera clara las hipótesis H1 y H2: el aumento de autonomía algorítmica y la desproporcionalidad de las sanciones reduce la aceptación de la decisión. De acuerdo con la literatura sobre legitimidad y justicia procedimental, la falta de intervención humana visible genera una percepción de frialdad, deshumanización y ausencia de juicio contextual (Tyler, 2006; Yeung, 2018), lo cual parece activar un rechazo más marcado que en contextos penitenciarios. Además, la penalización simbólica hacia las sanciones desproporcionadas sugiere que, en el imaginario ciudadano, los algoritmos deben operar bajo estándares especialmente estrictos cuando sus decisiones tienen consecuencias jurídicas serias. Este hallazgo es consistente con trabajos que indican que las personas aplican estándares morales más exigentes a los sistemas automatizados que a los humanos (Bigman & Gray, 2018).

En tercer lugar, los resultados relativos a la aceptación general del uso de herramientas algorítmicas muestran un patrón más estable y menos polarizado. La aceptación se mantiene en niveles moderados, sin variaciones pronunciadas entre condiciones, lo que sugiere que la ciudadanía no rechaza de manera frontal la incorporación de IA en el sistema judicial, pero tampoco la abraza sin reservas. La existencia de diferencias significativas en torno al grupo de control (G9), donde la

decisión la adoptan exclusivamente jueces humanos, indica que el uso de IA no desplaza al factor humano como referente de legitimidad. Más bien opera como un complemento cuyo valor depende de cómo se articula, de qué decisiones afecta y de qué garantías lo acompañan.

En este caso, en relación con las hipótesis 3 y 4, relativas a la aceptación global del uso de herramientas algorítmicas en el sistema de justicia penal , los resultados muestran que las variaciones en el grado de autonomía no producen cambios sustantivos en dicha valoración. Esto sugiere que la actitud ciudadana hacia la IA, entendida como una tecnología de carácter institucional, se mantiene relativamente estable y no se ve alterada por incrementos experimentales en el nivel de control algorítmico. De forma similar, la proporcionalidad de las sanciones tampoco modifica de manera significativa esta percepción global, pese a su evidente peso en la evaluación de decisiones judiciales concretas. Esta estabilidad apunta a que la aceptación general del uso de IA opera en un plano conceptual más amplio que el juicio sobre casos específicos, y está guiada por expectativas estructurales respecto al funcionamiento del sistema de justicia, al grado de confianza institucional y a la percepción de utilidad pública atribuida a estas tecnologías.

Los resultados plantean un escenario en el que la ciudadanía no rechaza la automatización, pero se trata de un fenómeno profundamente contextual y dependiente del tipo de decisión implicada. Mientras que en el ámbito penitenciario predomina una respuesta neutral y poco sensible a las variaciones experimentales, en el ámbito judicial emergen patrones claros que evidencian la importancia del control humano y la proporcionalidad como pilares de legitimidad. Por su parte, la valoración global del uso de herramientas algorítmicas se muestra más estable y desconectada del juicio sobre casos concretos.

PARTE III. Transferencia a la práctica profesional.

El presente capitulo se centra en el desarrollo empresarial enmarcado en el convenio de colaboración entre la Universidad Miguel Hernández de Elche (UMH) y la empresa Plus Ethics. Su finalidad es explicar la transferencia de los resultados de la presente tesis, los cuales se han materializado en la creación de una herramienta orientada a evaluar la aceptabilidad del uso de la inteligencia artificial en la administración pública, como pueden ser la educación, la sanidad, o como es el caso de la presente tesis, el sistema de justicia penal.

1. Identificación de necesidades del mercado.

En este apartado se describe el contexto que justifica el desarrollo de la herramienta y las razones que motivan su necesidad. El análisis se centra en la rápida evolución y el despliegue creciente de herramientas basadas en inteligencia artificial en el ámbito de las administraciones públicas. Asimismo, se establecen los objetivos del desarrollo industrial.

1.1. Justificación.

Desde una perspectiva más amplia que la del sistema de justicia penal analizada en la tesis, la rápida incorporación de tecnologías basadas en inteligencia artificial está produciendo transformaciones profundas en la estructura, los procedimientos y la cultura organizativa de la mayoría de las administraciones públicas. A escala comparada, la OECD (2024) constata que dos tercios de los países de la organización (alrededor del 67%) ya emplean IA en el diseño y prestación de servicios públicos, lo que indica un estadio de despliegue más allá del pilotaje inicial. En el contexto europeo, el estudio del *Joint Research Centre* (JRC) (Grimmelikhuijsen & Tangi, 2024) con directivos públicos muestra que el 52% declara al menos un proyecto de inteligencia artificial plenamente adoptado en su organización y que entre el 63% y el 52% sitúan sus administraciones en fase de planificación o implementación de proyectos, sobre todo en servicios y operaciones internas, y en menor medida en decisión de políticas. Estas evidencias sugieren que la velocidad de incorporación tecnológica avanza, con frecuencia, a un ritmo superior al de la capacitación y

adaptación organizativa del personal público, lo que refuerza la necesidad de instrumentos específicos para medir la aceptación entre profesionales del sector público.

En este contexto de expansión acelerada, el marco normativo vigente, especialmente el Reglamento (UE) 2024/1689, busca garantizar la conformidad legal y los requisitos de seguridad, transparencia y control humano. Es por ello, que actualmente existen múltiples herramientas orientadas a verificar el cumplimiento de dicho reglamento[14]. Estas herramientas, generalmente basadas en listas de verificación, permiten evaluar si una solución de inteligencia artificial cumple con las disposiciones establecidas. No obstante, el cumplimiento legal y técnico es una condición necesaria, pero no suficiente para garantizar el éxito de la implantación de una tecnología dentro de una organización. Superada la fase de verificación normativa, la cuestión clave pasa a ser cómo será percibida y aceptada la herramienta por las personas que deben emplearla. La experiencia muestra que algunas soluciones tecnológicas son adoptadas con facilidad, mientras que otras generan resistencia o rechazo, incluso cuando cumplen plenamente con las exigencias legales (Madan & Ashok, 2023; Rekunenko et al., 2025).

Como se ha mostrado en la tesis, estas diferencias en la adopción se explican por factores relacionados con la aceptación profesional, como la percepción de utilidad, la facilidad de uso, la confianza en los sistemas o la transparencia en su funcionamiento. Cuando las tecnologías son percibidas como opacas, complejas o poco comprensibles, la resistencia a su uso aumenta, especialmente en sectores sensibles donde la aceptación depende de la imparcialidad y la transparencia, como el ámbito judicial. Por tanto, resulta necesario disponer de metodologías que vayan más allá del mero cumplimiento normativo y que integren la dimensión humana en los procesos de evaluación e implantación de la inteligencia artificial. La aceptación de la transformación digital en la Administración Pública no depende únicamente

[14] Véase, por ejemplo: https://artificialintelligenceact.eu/es/evaluacion/comprobador-del-cumplimiento-de-la-ley-de-ai-de-la-ue/

de la conformidad legal, sino también de la confianza y disposición de los profesionales que la implementan.

En respuesta a esta necesidad, se propone el desarrollo de una herramienta, diseñada para evaluar empíricamente el grado de aceptación profesional de las tecnologías basadas en inteligencia artificial dentro de la Administración Pública. Esta herramienta ofrece un marco sistemático para medir la percepción y disposición de los profesionales antes de la implantación de una tecnología, permitiendo identificar resistencias, áreas de mejora y condiciones facilitadoras para una su adopción.

1.2. Análisis DAFO.

Con el objetivo de completar la justificación del desarrollo industrial de la herramienta y ofrecer una visión de su posicionamiento en el contexto actual, se presenta a continuación un análisis DAFO (Debilidades, Amenazas, Fortalezas y Oportunidades). Su finalidad es ofrecer una visión sintética y estratégica que facilite la toma de decisiones futuras en relación con la explotación, consolidación y expansión de la herramienta en distintos contextos administrativos.

En cuanto a las fortalezas, la organización dispone de personal con formación y experiencia en ética aplicada y evaluación social, así como en metodologías de análisis relacionadas con la adopción tecnológica. La colaboración entre la Universidad Miguel Hernández y Plus Ethics aporta una base metodológica y académica que favorece la coherencia científica del proyecto. Además, el equipo cuenta con la capacidad técnica necesaria para el cálculo de los diferentes índices, así como la elaboración de recomendaciones para la toma de decisiones en las administraciones públicas.

Entre las debilidades, se observa la ausencia de un equipo interno de desarrollo de software, lo que implica depender de colaboradores externos para la implementación tecnológica. Asimismo, el modelo de negocio de la entidad, centrado principalmente en la consultoría ética en proyectos europeos, requeriría de una adaptación hacia un formato híbrido que integre servicios de asesoramiento y soluciones digitales.

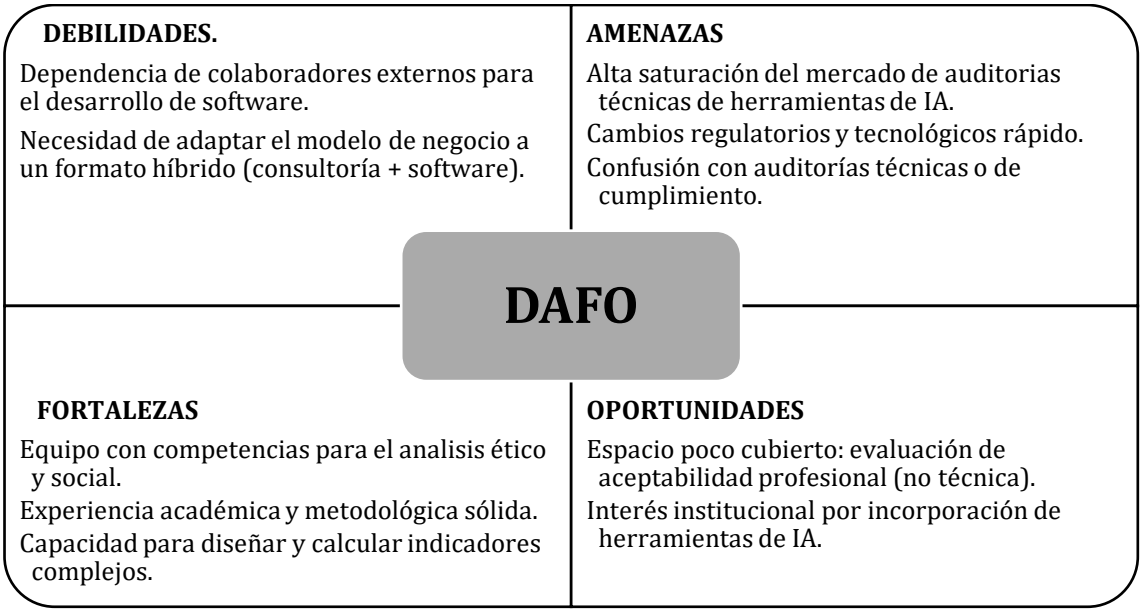
En relación con las oportunidades, el análisis identifica un espacio de aplicación con escasa cobertura actual: la evaluación de la aceptabilidad profesional de las herramientas basadas en inteligencia artificial. A diferencia de las auditorías técnicas o de cumplimiento normativo, este enfoque se orienta hacia la percepción y disposición de los profesionales, lo que permite abordar dimensiones complementarias en los procesos de implantación tecnológica. Además, el interés institucional creciente por la ética en la inteligencia artificial y la disponibilidad de programas públicos de apoyo a la innovación digital pueden favorecer la viabilidad de este tipo de iniciativas.

Por último, las amenazas se vinculan con la elevada oferta de herramientas de auditoría técnica de herramientas de IA, especialmente en primeras fases de desarrollo. También se identifican riesgos asociados a la rápida evolución normativa y tecnológica, que puede exigir actualizaciones continuas, y a la posible confusión conceptual entre auditorías técnicas y evaluaciones de aceptabilidad profesional.

La Figura 22 expone de forma resumida los principales elementos identificados en el análisis DAFO:

Figura 22.

Análisis DAFO..



1.3. Objetivos.

El objetivo principal de este capítulo es proporcionar a las administraciones públicas una herramienta metodológica que permita evaluar el nivel de aceptación del uso de las tecnologías basadas en inteligencia artificial por parte de sus profesionales. De este modo, el capítulo tiene como propósito presentar el diseño, la fundamentación teórica y la aplicación práctica de dicha herramienta, orientada a fortalecer los procesos de adopción responsable e informada de la IA en el sector público.

A través de la herramienta desarrollada, se persiguen los siguientes objetivos específicos:

OE1. Describir y analizar el nivel de aceptación del uso de herramientas de inteligencia artificial entre los profesionales de las administraciones públicas.

OE2. Formular recomendaciones orientadas a mejorar la aceptabilidad del uso de la IA en los procesos de diseño, despliegue y aplicación dentro del ámbito público.

OE3. Identificar y caracterizar aquellos casos de uso que resulten no aceptables o problemáticos desde la perspectiva de los profesionales del sector público.

2. Estructura de la herramienta.

En este apartado se presenta la estructura general de la herramienta desarrollada, detallando cómo ha sido desarrollado el instrumento de evaluación. Se explica la justificación que sustenta su diseño y su aplicabilidad más allá del marco de la tesis doctoral, trasladando los resultados obtenidos en un recurso operativo transferible a contextos reales. Asimismo, se describen los principales módulos que la componen: el Módulo 0, dedicado a la definición del caso de uso, en el que se delimita el contexto y la herramienta objeto de análisis; y el Módulo 1, correspondiente al cuestionario de aceptación profesional, que recoge y analiza las percepciones de los profesionales a través de un conjunto de ítems que permiten determinar el nivel de aceptación de uso.

2.1. Bases científicas para la creación de la herramienta.

El desarrollo de la herramienta se sustenta en un marco conceptual que integra las principales teorías sobre la aceptación tecnológica, orientadas a comprender cómo las personas y los colectivos profesionales asimilan y utilizan las innovaciones basadas en inteligencia artificial. En este contexto, se ha constatado que la incorporación de sistemas de IA en las Administraciones Públicas está generando transformaciones estructurales en los procesos, las funciones y los modelos organizativos institucionales. Este proceso, inscrito en la denominada *Cuarta Revolución Industrial* (Schwab, 2016), trasciende el ámbito meramente técnico y supone una reconfiguración más amplia de la gestión pública, en la que las interacciones entre tecnología, competencias profesionales y aceptación social adquieren un papel determinante (Aguilar, 2021; Hildebrandt, 2015; Susskind, 2019).

En esta línea, diversos estudios han mostrado que la confianza, la percepción de utilidad, la facilidad de uso y la transparencia en el funcionamiento de los sistemas son variables decisivas para su adopción (Davis, 1989; Venkatesh et al., 2003; Correia, Pereira & Bilhim, 2024). Cuando las tecnologías son percibidas como opacas o excesivamente complejas, la disposición de los profesionales a utilizarlas disminuye, especialmente en ámbitos donde la aceptación del usos depende de la

imparcialidad, la transparencia y la rendición de cuentas (van den Bos, 2001; Završnik, 2021). Estos factores sitúan la aceptación profesional como un elemento clave para comprender la dinámica real de la digitalización en el sector público.

La herramienta se apoya teóricamente en los principales modelos de aceptación tecnológica, que proporcionan un marco explicativo sobre las actitudes y comportamientos asociados al uso de nuevas tecnologías. La Teoría de la Acción Razonada (Fishbein & Ajzen, 1975) y la Teoría del Comportamiento Planeado (Ajzen, 1991) plantean que la intención de uso depende de la actitud hacia la tecnología, de las normas sociales percibidas y del control conductual subjetivo. El Modelo de Aceptación de la Tecnología (TAM) (Davis, 1989) y su desarrollo posterior en la Teoría Unificada de Aceptación y Uso de la Tecnología (UTAUT) (Venkatesh et al., 2003) destacan el papel de la utilidad percibida, la facilidad de uso y las condiciones organizativas como predictores directos de la adopción. De forma complementaria, el *Te*chnology *Readiness Index* (Parasuraman, 2000) incorpora la predisposición individual hacia la innovación tecnológica, mientras que la Teoría de la Difusión de Innovaciones (Rogers, 2003) explica la adopción en función de las características percibidas de la innovación, como su compatibilidad con los valores existentes o su complejidad. Finalmente, se ha considerado incluir elemento relacionados con el tecnoestrés y la carga cognitiva asociada al uso de sistemas digitales complejos, factores cada vez más reconocidos en la literatura sobre gestión pública y bienestar laboral (Tarafdar, Cooper & Stich, 2019; Salanova, Llorens & Cifre, 2013). Estos aspectos afectan tanto la disposición a adoptar nuevas herramientas como la percepción de autoeficacia y control. En el contexto de la Administración Pública, donde la implementación de la inteligencia artificial suele implicar cambios en rutinas, responsabilidades y procesos de supervisión, estos elementos adquieren relevancia. Su inclusión en la herramienta nos permite analizar posibles resistencias derivadas de la sobreexposición o del sentimiento de falta de competencia digital, aportando así una visión más completa de los factores que inciden en la aceptación profesional.

En este sentido, se ha incorporado la Escala de Creadores de Tecnoestrés (TCS) validada en España por Arenas, Sanclemente, Terán-Tinedo y Di Marco (2023), la

cual mide el nivel de estrés derivado del uso de las tecnologías en el trabajo. La versión española consta de 18 ítems distribuidos en cinco dimensiones: tecno-sobrecarga, tecno-invasión, tecno-complejidad, tecno-inseguridad y tecno-incertidumbre. La inclusión de este instrumento posibilita evaluar cómo la exposición a entornos digitales complejos puede influir en el bienestar y en la predisposición de los empleados públicos a integrar nuevas tecnologías en su práctica profesional.

Aunque la herramienta se desarrolla como un paso posterior a la adecuación del caso de uso conforme al Reglamento Europeo de Inteligencia Artificial (UE 2024/1689), resulta pertinente tener también en consideración el marco ético y normativo en su fundamentación, ya que permite vincular el cumplimiento técnico y jurídico con la aceptación profesional. Los principios de transparencia, supervisión humana, equidad y responsabilidad pueden traducirse en percepciones de confianza, control y comprensión, variables que influyen en la predisposición profesional a utilizar sistemas basados en inteligencia artificial.

La integración de estos enfoques responde a un propósito analítico: comprender cómo las actitudes individuales interactúan en los procesos de adopción tecnológica dentro de la Administración Pública. Existen algunos instrumentos de evaluación, como el Algorithmic Impact Assessment (AIA) canadiense o la herramienta DUTUS (Castro-Toledo & Gómez-Bellvís, 2026), que incluyen aspectos como el impacto social positivo, que valora la capacidad de la tecnología para contribuir de manera constructiva a la sociedad.

Esta fundamentación teórica constituye el marco conceptual sobre el que se articula la arquitectura metodológica. A partir de esta base teórica, la herramienta se estructura en dos módulos sucesivos que permiten, en primer lugar, determinar el caso de uso de concreto del que va a analizarse la aceptación, y en segundo lugar, la evaluación de esta.

2.2. Arquitectura de la herramienta.

2.2.1. Descripción general.

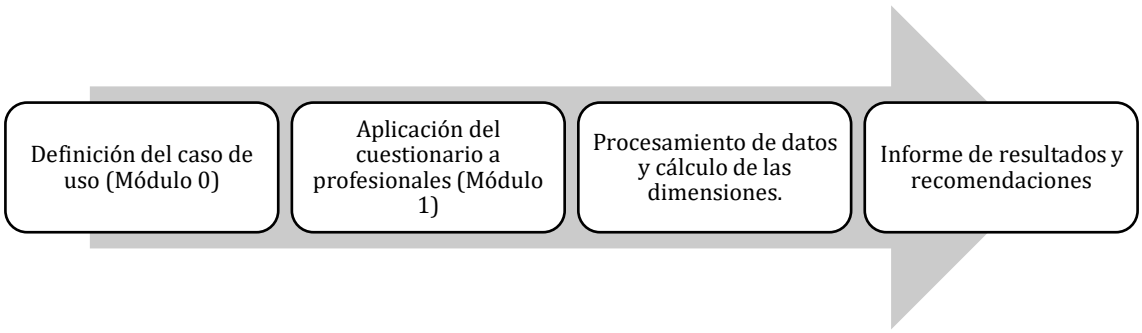
La arquitectura de la herramienta se estructura en torno a un modelo modular y secuencial, diseñado para facilitar tanto la recopilación estructurada de información contextual (a través de la definición del caso de uso) como la evaluación de las percepciones y actitudes de los profesionales. Cada módulo cumple una función específica, en concreto:

Figura 23. Resumen de la arquitectura.

Módulo	Denominación	Función principal
Módulo 0	Definición del caso de uso	Recopilar la información necesaria sobre el proyecto o iniciativa a evaluar, permitiendo configurar los parámetros de análisis y establecer la unidad de estudio.
Módulo 1	Evaluación de percepciones y actitudes	Recolectar información a partir de un cuestionario estructurado, dirigido a los profesionales que harán uso de la herramienta, para calcular los resultados por dimensiones-
Análisis y reporte	Procesamiento automático de resultados y cálculo del índice por dimensión.	Generar indicadores y recomendaciones específicas en función de las dimensiones analizadas en el módulo 1.

Por tanto, el flujo de funcionamiento de la herramienta sería:

Figura 24. Flujo de funcionamiento.



2.2.2. Fases de la metodología.

Fase I – Definición del caso de uso (Módulo 0).

El Módulo 0 representa la fase inicial del proceso de evaluación y tiene como objetivo contextualizar el proyecto o tecnología a analizar, definiendo las condiciones de partida y los parámetros necesarios para interpretar los resultados obtenidos en las fases posteriores de la herramienta. En el ámbito de la innovación tecnológica, el concepto de use case se ha consolidado como una unidad analítica esencial para delimitar escenarios de aplicación, identificar actores implicados y establecer los objetivos específicos de la intervención (Jacobson, 1993; Carroll, 2003). Un caso de uso se entiende como un escenario concreto en el que se pretende implementar una solución algorítmica o de IA, precisando los usuarios finales que interactuarán con ella, los procesos organizativos o sociales en los que intervendrá y los objetivos que persigue (Alexander & Maiden, 2005).

Su cumplimentación se realiza conjuntamente entre el desarrollador del proyecto o herramienta y el responsable designado por la administración pública interesada en su implantación.

En esta fase se define el caso de uso, que constituye la unidad básica de análisis de la herramienta y establece el marco que permitirá comprender las características que serán evaluadas en el Módulo 1.

Para la recogida de información se ha diseñado una plantilla que caracteriza al proyecto o tecnología objeto de evaluación. Puede cumplimentarse en formato

digital o físico y constituye la información de referencia a partir de la cual se articulan los módulos posteriores. En este sentido, por cada caso de uso que tenga la herramienta que se pretende analizar debe cumplimentarse una plantilla.

Plantilla para cumplimentar del caso de uso. Información general		
Campo	**Alcance**	**Descripción**
Nombre del proyecto o herramienta de IA	Denominación oficial o interna	
Entidad promotora / responsable	Organismo público que impulsa el proyecto.	
Área o servicio de aplicación	Sector o ámbito donde se usa la herramienta.	
Etapa actual del proyecto	Explica la fase actual de la herramienta Piloto / Implementación / En funcionamiento	
Persona de contacto / responsable técnico	Nombre, cargo, correo electrónico	
Socios o entidades colaboradoras	Indicar otros centros que desarrollan o apoyan la herramienta.	
Descripción del sistema		
Campo	**Descripción / Ejemplo**	**A completar**
Objetivo general	Explicar qué función cumple la herramienta.	
Principales funcionalidades	Describir las tareas o procesos para los que ha sido desarrollada.	
Tecnología base empleada	Tipo de técnica de IA usada	
Usuarios previstos	Perfiles que usarán la herramienta	
Ámbito de aplicación	En qué sector se va a implantar	
Entorno, Funcionamiento y Gobernanza.		
Datos y funcionamiento técnico		
Campo	**Descripción**	**A completar**
Tipo y origen de los datos empleados	Especifica si los datos son administrativos, textuales, de imagen, etc., y si provienen de fuentes internas, abiertas o de terceros.	
Procesos de validación técnica o control de calidad realizados	Describe las pruebas de precisión, validaciones cruzadas, revisiones de errores o indicadores de fiabilidad aplicados.	
Procedimientos de actualización o revisión de los modelos	Indica cada cuánto se actualizan los algoritmos o bases de datos, quién lo hace y bajo qué protocolo.	
Nivel de automatización y tecnología base empleada	Especifica si el sistema asiste decisiones, automatiza parcial o totalmente, y qué tipo de tecnología IA usa.	
Gobernanza		
Campo	**Descripción**	**A completar**
Mecanismos de control institucional	Qué órganos o figuras supervisan el sistema desde el punto de vista técnico, ético o funcional.	

Canales de coordinación y comunicación entre usuarios implicados	Describe cómo se comunican las incidencias o decisiones.
Procesos de auditoría, evaluación o revisión previstos	Indica si existen auditorías técnicas, revisiones éticas, evaluaciones de impacto algorítmico o de protección de datos.
Participación de usuarios o grupos de interés	Explica si se recaba feedback del personal, ciudadanía o expertos externos para mejorar la herramienta.
Medidas para prevenir sesgos y garantizar la equidad	Métodos o controles aplicados para evitar discriminación o resultados injustos.
Acciones o materiales destinados a la transparencia del sistema	Documentos o canales que explican el funcionamiento de la IA (guías, informes, web pública).

Fase II –Evaluación de percepciones y actitudes (Módulo 1).

El Módulo 1 constituye la segunda fase del proceso metodológico de la herramienta y tiene como finalidad evaluar el nivel de aceptación profesional hacia una herramienta o sistema de inteligencia artificial en el contexto de la Administración Pública.

Este módulo se aplica una vez definido el caso de uso en la Fase I y permite recoger información sistemática sobre las percepciones, actitudes y experiencias de los profesionales que interactúan o interactuarán con la tecnología evaluada. Su objetivo es determinar el nivel de aceptación profesional mediante un conjunto de indicadores agrupados en ocho dimensiones analíticas, que permiten calcular un Índice de Aceptación Profesional (IAP).

a) Procedimiento.

Una vez recogida la información del Módulo 0 se da comienzo a la fase de evaluación, centrada en recoger las percepciones y actitudes de los profesionales implicados en el uso o implementación de la herramienta descrita en el caso de uso.

La persona responsable de la administración seleccionará a los profesionales que utilizan o utilizarán la herramienta de inteligencia artificial para cumplimentar el cuestionario. Esta selección se realiza procurando incluir perfiles representativos de los distintos niveles jerárquicos, áreas funcionales y grados de experiencia tecnológica presentes en la organización. De este modo, se busca garantizar una

visión lo más amplia posible de las percepciones profesionales. Una vez seleccionados los participantes, se les informa sobre los objetivos del estudio y las condiciones de confidencialidad.

El cuestionario se cumplimenta en formato digital a través de la plataforma habilitada por Plus Ethics, que permite registrar de manera automática las respuestas y almacenar la información. Cada participante recibe un enlace de acceso individual, que debe cumplimentar de manera autónoma e independiente. En caso de duda sobre el significado de alguno de los ítems o sobre el procedimiento de respuesta, el participante puede contactar con el representante designado por la administración, quien está encargado de ofrecer las aclaraciones necesarias sobre la interpretación de las preguntas o del formato de respuesta. En aquellos casos en que no sea posible el uso de medios digitales, se facilita un formato físico equivalente. Las respuestas recogidas en papel son digitalizadas posteriormente, respetando el mismo esquema de codificación empleado por la plataforma, con el fin de mantener la homogeneidad del conjunto de datos.

La aplicación del cuestionario se desarrolla dentro del periodo previamente establecido en la planificación de la evaluación. Durante este tiempo, la persona responsable realiza un seguimiento del grado de participación y resuelve las posibles incidencias técnicas o logísticas que puedan surgir. El proceso concluye cuando todos los cuestionarios han sido cumplimentados y enviados a través de la plataforma. Este cierre marca la finalización de la fase de recogida de información y permite dar paso a la etapa siguiente, en la que se procederá al análisis y procesamiento de los datos recopilados.

b) Dimensiones evaluadas.

El cuestionario del Módulo 1 se organiza en ocho dimensiones analíticas, cada una orientada a evaluar un aspecto específico de la aceptación profesional de las herramientas basadas en inteligencia artificial. En concreto:

Tabla 28.

Resumen de las dimensiones incluidas en el cuestionario.

Dimensión	Aspecto que evalúa	N.º de ítems
1. Fiabilidad de la herramienta	Seguridad percibida en el uso de la herramienta y en la corrección de sus resultados.	3
2.Transparencia y explicabilidad	Grado en que los usuarios comprenden el funcionamiento del sistema y perciben claridad en sus decisiones.	7
3. Supervisión institucional y participación ciudadana	Existencia de mecanismos de control, supervisión y participación en el uso de la tecnología.	5
4. Utilidad percibida y facilidad de uso	Valoración de la herramienta en términos de eficacia, eficiencia y facilidad de aprendizaje.	4
5. Imparcialidad	Percepción sobre la equidad, ausencia de sesgos y objetividad en los resultados generados.	3
6. Responsabilidad y autonomía percibida	Sensación de control y mantenimiento del juicio profesional frente a la automatización.	4
7. Intención de uso y recomendación.	Voluntad de utilizar la herramienta tecnológica y recomendar su uso a otros.	3
8. Tecnoestrés	Nivel de sobrecarga o tensión percibida derivada del uso de sistemas digitales complejos.	18
9. Variables sociodemográficas	Datos descriptivos para contextualizar el perfil del participante (edad, nivel educativo, experiencia, rol).	4

c) Cuestionario.

El instrumento se compone de una serie de ítems organizados en diferentes dimensiones, cada una orientada a medir un aspecto específico del proceso de aceptación profesional. Las respuestas se recogen mediante una escala tipo Likert de 1 (totalmente en desacuerdo) a 5 (totalmente de acuerdo), incluida la escala de tecnoestrés.

A continuación, se presenta la versión completa del cuestionario, organizada según las ocho dimensiones analíticas definidas en el marco metodológico.

Tabla 29.

Cuestionario completo.

Dimensión	Items
Fiabilidad de la herramienta. (Basado en Davis, 1989; Venkatesh et al., 2003; Lee & See, 2004)	Me siento seguro/a utilizando esta herramienta en mis tareas. Confío en que los resultados que ofrece la herramienta son correctos y útiles. Me preocupa que esta herramienta tome decisiones sin suficiente control humano. (ítem invertido) Conozco los procedimientos para detectar y corregir errores de la herramienta.
Transparencia y explicabilidad. (Basado en Doshi-Velez & Kim, 2017; HLEG-UE, 2019; AI Act, 2024; UNESCO, 2021)	Entiendo cómo funciona la herramienta. He recibido la formación necesaria para comprender su funcionamiento. Esta herramienta me ofrece explicaciones claras y comprensibles sobre cómo llega a sus resultados. Me sentiría más cómodo/a utilizando esta herramienta si recibiera información adicional o ejemplos prácticos sobre su funcionamiento. No usaría esta herramienta si no pudiera comprender cómo toma sus decisiones. (ítem invertido). Puedo acceder fácilmente a información sobre quién es responsable de su funcionamiento. Tengo información sobre qué datos utiliza la herramienta y con qué finalidad. Me preocuparía utilizar esta herramienta si no estuviera claro quién puede acceder a los datos que procesa. (ítem invertido)
Supervisión institucional y participación ciudadana. (Basado en European Commission, 2024; HLEG-UE, 2019; Consejo de Europa, 2023; Zouridis et al., 2020)	Considero que el uso de esta herramienta está respaldado por controles institucionales adecuados. Los usuarios finales participamos en la mejora de esta herramienta. Me preocupa que esta herramienta pueda usarse sin suficiente supervisión o control democrático. *(ítem invertido)* Las personas responsables tienen la capacidad real de revisar, corregir o anular las decisiones sugeridas por la herramienta. Dispongo del tiempo y la información necesaria para cuestionar los resultados de la herramienta cuando lo considero adecuado.
Utilidad percibida y facilidad de uso. (Basado en Davis, 1989; Venkatesh, 2003; Parasuraman, 2000; Rogers, 2003)	Esta herramienta me ayuda a realizar mi trabajo de forma más eficiente. La considero fácil de aprender y de usar en mi contexto. Mi organización me proporciona la formación necesaria para usarla de forma responsable y eficaz. Existen políticas internas claras sobre cuándo y cómo debe utilizarse esta herramienta.
Imparcialidad. (Basado en AI Act, 2024; Buolamwini & Gebru, 2018; Draws et al., 2021)	Percibo que esta herramienta toma decisiones de forma justa y sin discriminaciones. Me preocupa que los resultados de la herramienta estén sesgados por factores como género, edad o etnia. *(ítem invertido)* Considero que la herramienta contribuye a decisiones más objetivas.

Dimensión	Items
Responsabilidad y autonomía percibida (Lee & See, 2004; European Commission, 2024)	Siento que mantengo el control final sobre las decisiones que toma esta herramienta. El uso de esta herramienta complementa, pero no sustituye, mi juicio profesional. Me preocupa que la herramienta reduzca mi capacidad de decidir. *(ítem invertido)* Confío en que la supervisión humana evita que la herramienta tome decisiones automáticas inadecuadas.
Intención de uso y recomendación. (Davis, 1989; Venkatesh et al., 2003)	Recomendaría el uso de esta herramienta a otros profesionales de mi organización. Integraría esta herramienta de forma habitual en mis tareas cuando esté disponible. La herramienta mejora mi capacidad para ofrecer un servicio de mayor calidad a la ciudadanía.
Tecnoestrés. Arenas, Sanclemente, Terán-Tinedo & Di Marco (2023).	***Tecno-sobrecarga (TC1)*** Estas tecnologías me obligan a trabajar mucho más rápido. Estas tecnologías me obligan a hacer más trabajo del que puedo manejar. Estas tecnologías me obligan a tener horarios de trabajo muy ajustados. ***Tecno-invasión (TC2)*** Tengo que estar en contacto con mi trabajo, incluso durante mis vacaciones, debido a estas tecnologías. Tengo que sacrificar mis vacaciones y tiempo de mi fin de semana para mantenerme al día en nuevas tecnologías. Siento que mi vida personal está siendo invadida por estas tecnologías. ***Tecno-complejidad (TC3)*** No sé lo suficiente sobre estas tecnologías para hacer mi trabajo satisfactoriamente. Necesito mucho tiempo para entender y utilizar nuevas tecnologías. No encuentro tiempo suficiente para estudiar y mejorar mis habilidades tecnológicas. Creo que el nuevo personal de esta organización sabe más sobre tecnología informática que yo. A menudo me parece demasiado complicado entender y usar nuevas tecnologías. ***Tecno-inseguridad (TC4)*** Tengo que actualizar constantemente mis habilidades para evitar ser reemplazado. Me siento amenazado por compañeros de trabajo con habilidades tecnológicas más actualizadas. No comparto mi conocimiento con mis compañeros de trabajo por miedo a ser reemplazado. Siento que hay menos intercambio de conocimiento entre compañeros de trabajo por miedo a ser reemplazados. ***Tecno-incertidumbre (TC5)*** Hay cambios constantes en el software de los ordenadores de nuestra organización. Hay cambios constantes en el hardware de los ordenadores de nuestra organización.

Dimensión	Items
	Hay actualizaciones frecuentes en las redes informáticas de nuestra organización.
Variables sociodemográficas.	Edad, nivel educativo. Nivel de experiencia en el uso de tecnologías digitales (básico, intermedio, avanzado). Rol con el que se utiliza la herramienta (profesional del sector, ciudadano, responsable institucional, etc.).

Fase III – Análisis y reporte.

La tercera fase corresponde al análisis y la presentación de los resultados obtenidos tras la aplicación del cuestionario de percepciones y actitudes. Su finalidad es analizar los datos recogidos para evaluar el nivel de aceptación profesional de la herramienta de inteligencia artificial en estudio. Esta fase comprende el cálculo de los indicadores por dimensión y la elaboración del informe de resultados.

El proceso de análisis se desarrolla mediante diferentes etapas: en primer lugar, se realiza la depuración de la base de datos con el fin de eliminar registros incompletos, inconsistentes o duplicados. Los ítems formulados en sentido inverso se recodifican antes del procesamiento para mantener la coherencia interna del instrumento. A continuación, se calculan estadísticos descriptivos para cada ítem y para cada una de las dimensiones del cuestionario.

Una vez analizados los ítems individualmente, se procede al cálculo de los índices por dimensión, los cuales reflejan de manera independiente el nivel de aceptación asociado a cada ámbito analizado (fiabilidad, transparencia, supervisión, utilidad, imparcialidad, responsabilidad y tecnoestrés). Cada índice se obtiene como la media aritmética de las puntuaciones de los ítems que integran la dimensión correspondiente, transformada a una escala de 1 a 5. La obtención de estos índices permite identificar áreas específicas de fortaleza y de mejora dentro del proceso de adopción de la tecnología. De este modo, una organización puede conocer no solo el grado general de aceptación, sino también en qué aspectos concretos es necesario intervenir antes de su implementación o durante su uso operativo.

Los índices por dimensión se expresan en una escala continua de 1 a 5 puntos y se interpretan según los intervalos definidos en la siguiente tabla:

Tabla 30.

Escala de interpretación de las puntuaciones de los índices.

Índice dimensión	Nivel de cumplimiento de la dimensión	Interpretación general
1,00 – 2,00	Bajo	Nivel de aceptación bajo; predominan percepciones negativas y se rechaza su inclusión en la administración.
2,01 – 3,00	Medio-bajo	Nivel de aceptación medio-bajo; existen actitudes ambivalentes y dudas que limitan la disposición a adoptar la herramienta.
3,01 – 4,00	Medio-alto	Nivel de aceptación medio-alto; se observa una actitud favorable con algunos aspectos susceptibles de mejora antes de la implementación plena.
4,01 – 5,00	Alto	Nivel de aceptación alto; se evidencia confianza general y disposición positiva hacia el uso de la herramienta.

Con base en estos intervalos, se establece un criterio de interpretación aplicable a los índices por dimensión. Si alguna de las dimensiones presenta un índice inferior al nivel medio-alto, la herramienta se considera no validada para su implantación y se recomienda posponer su despliegue hasta que se adopten medidas correctivas o de mejora que eleven los niveles de aceptación a los rangos deseables.

Una vez finalizado el análisis, los responsables de Plus Ethics elaboran el informe de resultados y lo remiten al responsable institucional designado. El documento presenta las puntuaciones por dimensiones, pudiendo incluir representaciones gráficas de los resultados. Asimismo, el informe incorpora recomendaciones específicas sobre las principales fortalezas y los aspectos de mejora identificados en cada dimensión, con el propósito de facilitar la comprensión y el aprovechamiento práctico de los resultados. De esta manera, el documento cumple una doble función: ofrecer una evaluación detallada por dimensiones y servir como herramienta de apoyo a la toma de decisiones en las administraciones públicas, permitiendo elaborar planes de mejora orientados a la adopción responsable de tecnologías basadas en inteligencia artificial en la Administración Pública..

a) Interpretación de los resultados para la toma de decisiones.

La interpretación de los resultados se orienta a facilitar la lectura y el uso práctico de los indicadores por dimensión, que constituyen el núcleo del análisis de aceptabilidad profesional. Cada dimensión evaluada refleja un aspecto específico de la relación entre los profesionales y la herramienta de inteligencia artificial analizada, por ello, para simplificar la comunicación de los resultados y facilitar la toma de decisiones, las puntuaciones obtenidas en cada dimensión se interpretan mediante un sistema tipo semáforo, que permite visualizar de manera intuitiva el grado de aceptación alcanzado:

Figura 25.

Escala de interpretación de los resultados..

Aceptación alta o consolidada: 4,01 – 5,00.
Indica valoraciones positivas y confianza generalizada en la dimensión evaluada.
Las condiciones son adecuadas para mantener o replicar el funcionamiento observado, sin requerir intervenciones inmediatas.

Aceptación media o intermedia: 3,01 – 4,00.
Señala una percepción favorable, aunque con aspectos susceptibles de mejora. Se recomienda aplicar medidas de refuerzo específicas (por ejemplo, formación, clarificación de procedimientos o comunicación interna) antes de avanzar en nuevas fases de implementación.

Aceptación baja o deficiente: 1,00 - 3,00.
Refleja desconfianza, rechazo o percepciones críticas por parte del personal. En estas dimensiones, no se recomienda avanzar en la implantación hasta que se adopten medidas correctivas y se repita la evaluación.

Los resultados de cada dimensión se representarán visualmente mediante gráficos de araña, en los que cada eje corresponde a una dimensión evaluada. La forma y extensión del polígono permiten identificar visualmente el equilibrio entre los

distintos aspectos de la aceptación profesional.

A continuación, se muestran varios ejemplos de la devolución de resultados, representados mediante gráficos de araña que ilustran distintos niveles de aceptación por dimensión, como se muestra en los siguientes ejemplos:

Figura 26.

Gráfico de ejemplo 1 de resultados.

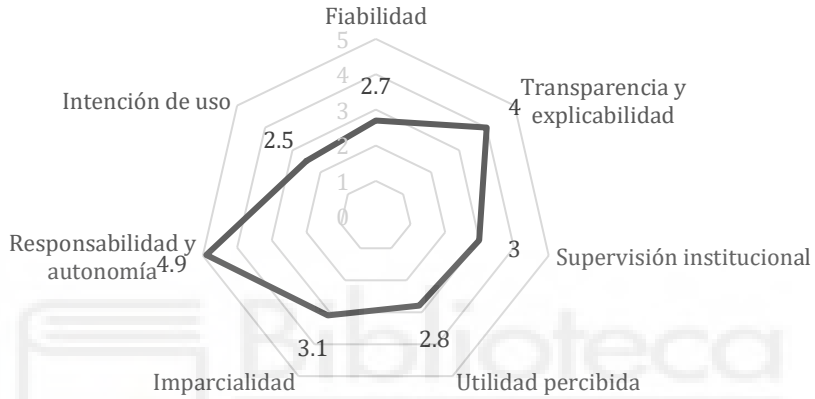


Figura 27.

Gráfico de ejemplo 2 de resultados.

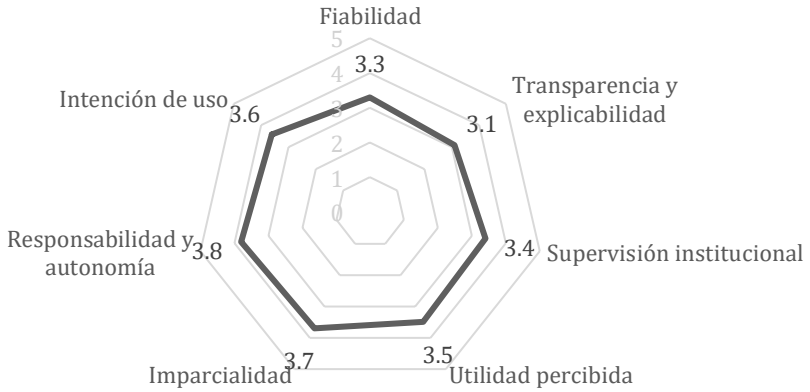


Figura 28.

Gráfico de ejemplo 3 de resultados.

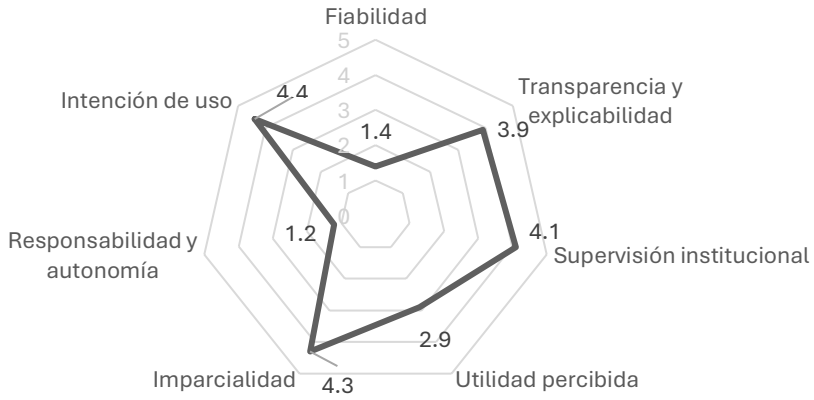
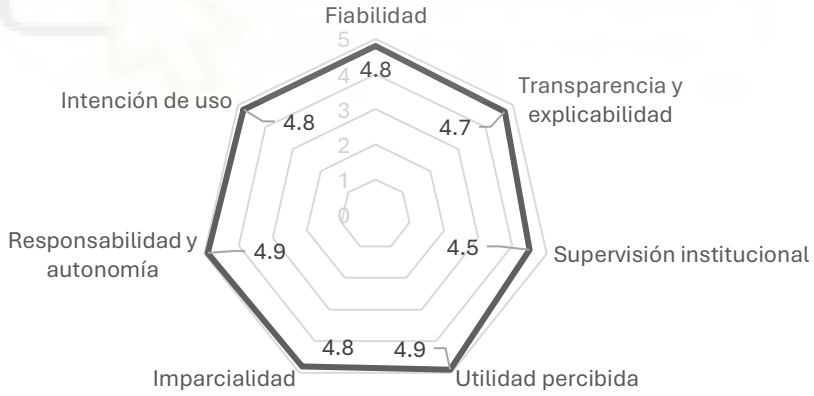


Figura 29.

Gráfico de ejemplo 4 de resultados.



La lectura de los gráficos de ejemplo permite observar cómo se distribuyen las puntuaciones obtenidas en cada dimensión evaluada y cómo estas configuran distintos escenarios de aceptabilidad. En la Figura 26 se aprecian áreas con puntuaciones altas, pero también dimensiones que requieren atención, situando la implantación de la herramienta en un escenario condicionado a mejoras específicas antes de una adopción plena. En la Figura 27, el polígono muestra un perfil más

equilibrado, con valores moderados en la mayoría de las dimensiones, lo que sugiere un nivel de aceptación suficiente pero aún dependiente de ajustes concretos para garantizar una implementación aceptada. La Figura 28 muestra niveles muy bajos en algunas dimensiones, en este tipo de caso, la recomendación es de no implantación y trabajar para mejorar las dimensiones más bajas, requiriendo de una nueva evaluación al finalizar los ajustes. Por último, la Figura 29 presenta un polígono amplio y homogéneo, este escenario plantearía el ideal, caracterizado por puntuaciones altas en todas las dimensiones, lo que indica un escenario de implantación plenamente favorable, en el que la herramienta sería aceptada con escasas objeciones por los profesionales. En conjunto, estos ejemplos ilustran cómo los gráficos permiten identificar de forma intuitiva fortalezas, áreas críticas y condiciones necesarias para la implementación informada de herramientas de inteligencia artificial en el entorno profesional.

3. Proceso de funcionamiento.

En este apartado se presenta la secuencia de acciones que conforman el proceso de aplicación de la herramienta. Su propósito es ofrecer una descripción estructurada del funcionamiento del servicio, identificando los actores involucrados en su gestión y los procedimientos que aseguran su adecuado desarrollo. La fase de explotación describe las formas en que la herramienta puede ponerse a disposición de las administraciones públicas y las condiciones bajo las cuales se gestiona su uso, mantenimiento y actualización. Este apartado tiene como finalidad presentar las alternativas existentes para el acceso y utilización del sistema, considerando tanto la sostenibilidad técnica del servicio como las necesidades operativas de las entidades usuarias.

3.1. Alojamiento.

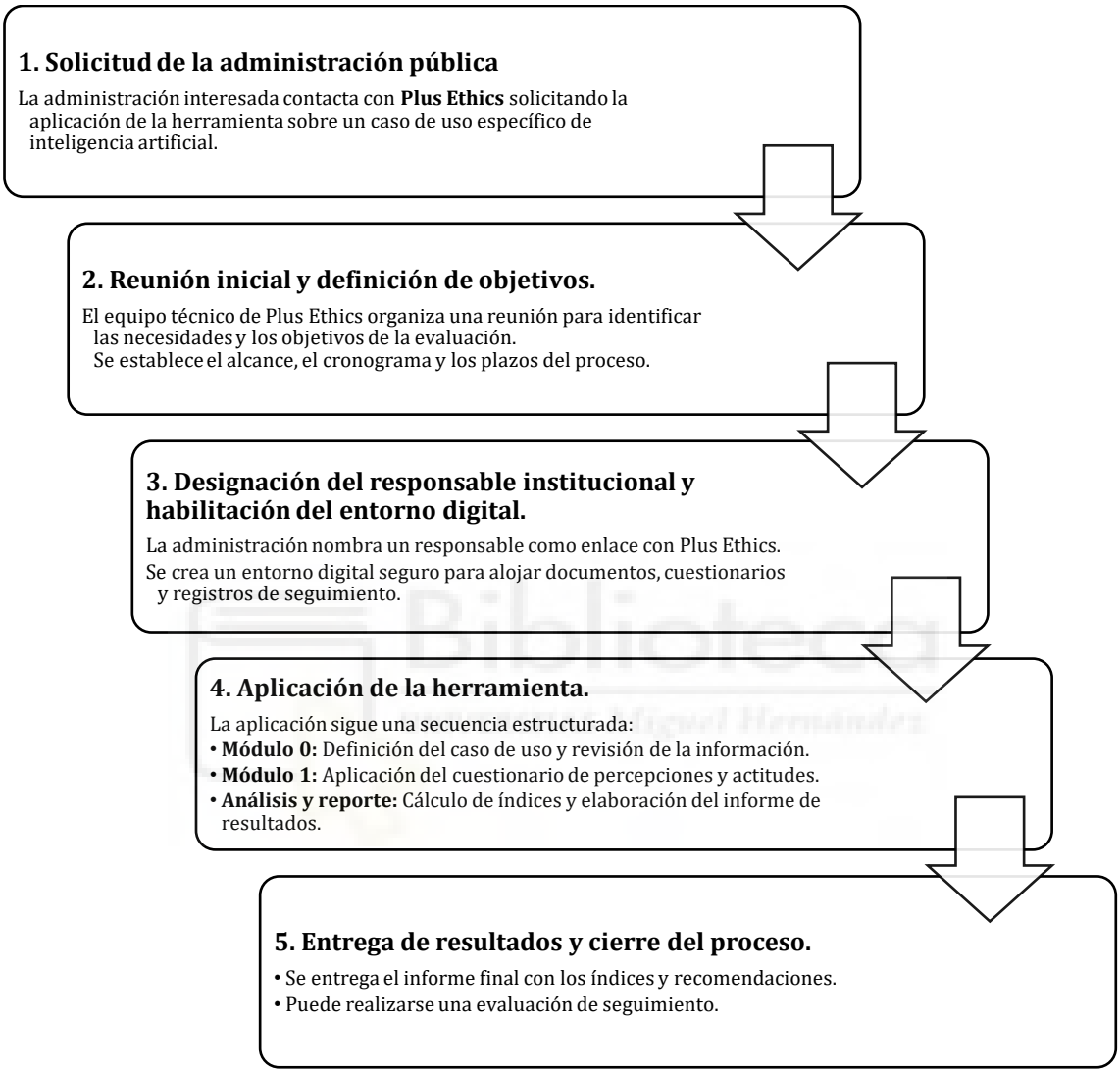
El alojamiento de la herramienta se llevará a cabo en la plataforma web gestionada por Plus Ethics, en la que se integran los distintos módulos que conforman su arquitectura funcional. Plus Ethics será responsable de la administración técnica del sistema, de la gestión de usuarios, de la actualización de versiones y de la adopción de las medidas necesarias para garantizar la seguridad y confidencialidad de los datos.

El entorno web se organiza en tres áreas diferenciadas: un área institucional de acceso, destinada a las administraciones públicas registradas; un área de aplicación, en la que se ejecutan los módulos 0 y 1 y se gestionan las evaluaciones; y un área de gestión y soporte, utilizada por el equipo de Plus Ethics para el seguimiento técnico del servicio. Cada administración usuaria dispondrá de credenciales específicas y de un espacio desde el que podrá configurar sus evaluaciones, supervisar su desarrollo y consultar, descargar o archivar los informes generados.

3.2. Lógica de uso.

La lógica de uso de la herramienta describe el flujo operativo completo que sigue una administración pública desde el momento en que solicita la aplicación de la herramienta hasta la finalización del proceso de evaluación. Este esquema refleja la

secuencia funcional de las interacciones entre la administración y Plus Ethics, así como las principales tareas asociadas a cada fase. En concreto, se desarrollaría de la siguiente manera:



A continuación, se presenta una descripción pormenorizada de las fases mencionadas que conforman el funcionamiento operativo de la herramienta.

El procedimiento se inicia cuando una administración pública contacta con Plus Ethics para solicitar la aplicación de la herramienta sobre un caso de uso específico de inteligencia artificial. Este primer contacto puede realizarse mediante comunicación institucional formal o a través del formulario disponible en la página web de la empresa. Una vez recibida la solicitud, el equipo técnico organiza una reunión inicial de diagnóstico en la que se identifican las necesidades de la

organización, el tipo de sistema de IA a evaluar, su estado de desarrollo y el nivel de madurez tecnológica. A partir de esta información, se definen el alcance del servicio, los objetivos específicos, el cronograma y los plazos de ejecución.

Tras la aceptación formal, la administración designa a una persona responsable institucional, que actuará como enlace permanente entre ambas partes. Una vez designado este perfil, se activa la planificación conjunta del trabajo y se habilita un entorno de colaboración dentro de la plataforma gestionada por Plus Ethics. En este espacio se alojan los documentos de referencia, la plantilla del caso de uso, el cuestionario, los registros de seguimiento y los informes generados a lo largo del proceso. Este entorno constituye el canal oficial de comunicación y supervisión.

La ejecución práctica de la herramienta se desarrolla de acuerdo con la siguiente secuencia lógica:

1. Aplicación (definida en la arquitectura de la herramienta):
    - Definición del caso de uso (Módulo 0).
      Cumplimentación de la plantilla de uso del proyecto o sistema de IA que se evaluará.
    - Aplicación del cuestionario (Módulo 1).
      Cumplimentación del cuestionario de percepciones y actitudes en relación con la plantilla de uso.
    - Análisis y reporte (Fase III del Módulo 1).
      Se procesan los datos para calcular los índices por dimensión y se elaboran las recomendaciones de mejora antes de su despliegue operativo.
    - Entrega de resultados.
      Plus Ethics elabora un informe técnico que incluye los resultados globales y por dimensión, representaciones gráficas y un resumen interpretativo con recomendaciones específicas sobre aspectos de mejora.

Una vez entregados los resultados, se archivan todos los datos en un repositorio accesible únicamente a la administración titular del caso. En este punto, Plus Ethics puede ofrecer la posibilidad de realizar una evaluación de seguimiento en fechas

posteriores para comprobar la evolución del nivel de aceptación tras la aplicación de las recomendaciones o reuniones de seguimiento para su correcta adaptación.

3.3. Tareas asociadas a cada responsable.

El funcionamiento de la herramienta implica la participación coordinada de varios responsables, cada uno con responsabilidades específicas dentro del proceso de evaluación. En concreto, intervienen principalmente cuatro figuras:

1. Plus Ethics, entidad responsable del desarrollo, mantenimiento, actualización y supervisión técnica de la herramienta.
2. La persona responsable institucional, designada por la administración pública usuaria.
3. Los profesionales evaluados, que aportan la información sobre las percepciones y actitudes hacia el sistema de IA.
4. El equipo técnico o responsable del sistema de IA evaluado, encargado de facilitar la información técnica y funcional necesaria durante la definición del caso de uso.

A continuación, se describen las tareas específicas de cada actor.

Actor	Rol principal	Funciones principales
Plus Ethics	Equipo responsable del desarrollo, alojamiento y mantenimiento de la herramienta.	– Recibe la solicitud de evaluación y organiza la reunión inicial con la administración. – Asesora en la definición del caso de uso (Módulo 0). – Configura la herramienta y el cuestionario (Módulo 1) y sus actualizaciones. – Ofrece soporte técnico durante la aplicación. – Procesa los datos, calcula los índices e interpreta resultados. – Elabora y entrega el informe final y, si se solicita, realiza el seguimiento posterior.
Responsable institucional (Administración Pública)	Persona designada por la administración para coordinar el proceso y actuar como enlace con Plus Ethics.	– Participa en la reunión inicial y facilita la información sobre el contexto y el sistema de IA. – Supervisa la correcta cumplimentación del caso de uso y valida la información aportada. – Selecciona a los profesionales participantes, asegurando diversidad de perfiles. – Informa a los participantes sobre los objetivos del proceso, su carácter confidencial y el modo de acceso al cuestionario. – Supervisa el desarrollo de la aplicación y gestiona incidencias o consultas. – Recibe el informe final, coordina la difusión de los resultados y, si procede, solicita una evaluación de seguimiento.
Participantes / Profesionales evaluados	Usuarios o personal que aporta su opinión sobre la herramienta mediante el cuestionario.	– Cumplimentan las preguntas de forma autónoma y confidencial. – Consultan dudas con el responsable institucional si es necesario. – Envían el cuestionario dentro del plazo establecido.
Equipo técnico del sistema de IA	Personal técnico que proporciona información detallada sobre el sistema evaluado.	– Aporta la documentación técnica (arquitectura, datos, auditorías, control de calidad). – Colabora en la descripción del caso de uso junto con la administración – Participa en la implementación de mejoras si el informe lo recomienda.

3.4. Explotación.

La fase de explotación define las condiciones bajo las cuales la herramienta se pone a disposición de las administraciones públicas y regula su uso operativo, mantenimiento y actualización. El servicio será gestionado por Plus Ethics, entidad responsable del alojamiento, supervisión técnica y evolución metodológica del sistema. Asimismo, la empresa podrá incorporar actualizaciones periódicas que integren mejoras funcionales, ajustes normativos o adaptaciones derivadas de la experiencia de uso y de los nuevos requerimientos institucionales.

Con el fin de atender a distintos contextos administrativos y modelos de gestión, la explotación de la herramienta podrá realizarse bajo diferentes modalidades de licencia, que determinan el tipo de acceso, el alcance temporal del uso y el grado de autonomía de la administración usuaria. A continuación, se detallan las principales modalidades previstas junto con sus ventajas y limitaciones desde la perspectiva tanto de Plus Ethics como de las administraciones públicas.

Tabla 31.

Modalidades de explotación de la herramienta.

Licencia abierta - La administración accede de manera libre o gratuita a la herramienta.			
Pro		**Contra**	
Plus Ethics	**Adm. Pública.**	**Plus Ethics**	**Adm. Pública.**
– Aumenta la visibilidad de la empresa. – Permite feedback de usuarios.	– Acceso sin coste directo. – Adaptación propia de la herramienta. – Acceso rápido y sin intermediarios.	– Escasa rentabilidad directa. – Falta de control en versiones derivadas.	– Limitado soporte técnico. – Ausencia de garantía de actualizaciones.
Licencia por uso *(token)* - La administración pública paga por cada proceso de evaluación o conjunto de evaluaciones, recibiendo acceso temporal a la herramienta.			
Pro		**Contra**	
Plus Ethics	**Adm. Pública.**	**Plus Ethics**	**Adm. Pública.**

– Proporciona ingresos continuos vinculados al uso real del sistema. – Mantiene el control técnico y metodológico de la herramienta. – Permite adaptar los recursos a la demanda de las administraciones.	– Ofrece flexibilidad presupuestaria, pagando solo por las evaluaciones realizadas. – Garantiza soporte técnico y actualizaciones – Evita la necesidad de infraestructura propia o mantenimiento interno.	– Requiere gestión administrativa constante (tokens, facturación, asistencia). – Dependencia del número de clientes activos para la sostenibilidad de la herramienta. – Requiere mantener un sistema de actualizaciones.	– Dependencia del proveedor para cada acceso o nueva evaluación. – Mayor coste si se realizan evaluaciones frecuentes. – Menor autonomía de adaptación de la herramienta.

Licencia adquirida - la administración pública adquiere la herramienta de forma indefinida, con derecho a uso continuo.

Pro		**Contra**	
Plus Ethics	**Adm. Pública.**	**Plus Ethics**	**Adm. Pública.**
– Ingresos directos. – Mantiene la relación a través de servicios de actualización.	– Mayor rentabilidad si se realizan evaluaciones frecuentes. – Facilita la integración en entornos institucionales propios.	– Requiere mantener un sistema de actualizaciones. – Menor flujo de ingresos recurrentes.	– Mayor inversión inicial. – Costes adicionales por la implementación de actualizaciones.

3.5. Resumen: modelo CANVAS.

PARTE IV. DISCUSIÓN Y CONCLUSIONES.

1. RECAPITULACIONES DE LOS ESTUDIOS EMPÍRICOS Y SUS LIMITACIONES.

La presente tesis doctoral se ha desarrollado con el propósito de analizar cómo los procesos de digitalización, algoritmización e incorporación de la inteligencia artificial están transformando, de manera irreversible, las dinámicas humanas que sustentan el sistema de justicia penal. La transformación tecnológica ya no constituye una posibilidad futura, sino una realidad que ha irrumpido para quedarse. Sin embargo, su integración efectiva exige una comprensión profunda de las condiciones humanas, sociales y profesionales en las que se desarrolla. Esta necesidad de comprensión responde a una premisa fundamental: las innovaciones tecnológicas no son entidades autónomas ni neutras, sino construcciones atravesadas por los valores, las decisiones y los significados de quienes las diseñan e implementan. En consecuencia, la aceptación e incorporación de estas herramientas no dependen únicamente de sus capacidades técnicas, sino de cómo los actores institucionales y sociales las interpretan, adoptan y reconfiguran dentro de sus contextos específicos. El verdadero eje de la transformación reside, pues, en el factor humano: en la capacidad de las personas para integrar, cuestionar y dar sentido a la tecnología dentro de sus prácticas y estructuras cotidianas.

Estas consideraciones constituyen la base del desarrollo empírico llevado a cabo a lo largo de la tesis. Si bien los capítulos 4, 5 y 6 incluyen en sus respectivas secciones de conclusiones los principales resultados y limitaciones de cada estudio, la diversidad de objetivos y metodologías hace necesario ofrecer una síntesis integradora que posteriormente permita extraer conclusiones generales sobre el papel del factor humano en la transformación tecnológica del sistema de justicia penal.

En primer lugar, la revisión sistemática sobre sesgos humanos y algorítmicos puso de manifiesto que la toma de decisiones judiciales no es completamente objetiva y que la subjetividad que caracteriza al juicio humano también está empezando a analizarse en el uso de herramientas digitales. En los últimos años, la preocupación por los sesgos y la equidad en la justicia se ha extendido al ámbito algorítmico, donde

los estudios muestran que, aunque estas tecnologías pueden reducir ciertos errores humanos, también pueden reproducir o incluso amplificar los prejuicios presentes en los datos y en su diseño. De este modo, la revisión pone de manifiesto que las preocupaciones que tradicionalmente se habían dirigido exclusivamente hacia las limitaciones y sesgos del juicio humano no han desaparecido, sino que se han ampliado hacia las tecnologías que pretenden complementarlo o sustituirlo, trasladándose ahora al terreno de la digitalización judicial y generando nuevas preguntas sobre cómo garantizar decisiones justas, transparentes y respetuosas con los derechos fundamentales en entornos cada vez más automatizados.

En segundo lugar, los estudios cualitativos realizados con operadores jurídicos y cuerpos policiales mostraron que la digitalización es percibida como un proceso necesario e irreversible, pero también generador de tensiones y resistencias. Los participantes reconocen los beneficios de las tecnologías, como la agilización de procedimientos o la mejora en la gestión de la información, pero expresan preocupación por la pérdida de autonomía profesional, la falta de formación especializada y la deshumanización del proceso judicial. Estos resultados reflejan que la aceptación de la innovación tecnológica depende tanto de factores organizativos y culturales como de la percepción de legitimidad y control por parte de los profesionales implicados.

En tercer lugar, los estudios cuantitativos sobre las precepciones, actitudes y aceptación social de la inteligencia artificial en la justicia evidenciaron una actitud de aceptación condicionada. La ciudadanía valora positivamente la eficiencia y la rapidez que la tecnología puede aportar, pero exige que las decisiones sigan sometidas a una supervisión humana efectiva y que los procesos sean transparentes y explicables. En otras palabras, la confianza pública en la justicia digital depende no solo de la fiabilidad técnica de las herramientas, sino también de la percepción de que las personas siguen teniendo el control sobre sus decisiones.

Sin embargo, no debemos abordar estos resultados como apartados diferenciados que no se relacionan entre sí, sino como elementos interdependientes de un mismo fenómeno complejo: la aceptación social y profesional de la inteligencia artificial en el sistema de justicia penal. Los estudios empíricos proporcionan aproximaciones

complementarias que, al integrarse, permiten construir una explicación de las dinámicas cognitivas, éticas y contextuales que condicionan la actitud hacia la justicia algorítmica.

Finalmente, dado que esta tesis analiza empíricamente el papel del factor humano en la transformación digital del sistema de justicia penal, presenta algunas limitaciones derivadas del diseño metodológico y del contexto empírico:

En primer lugar, los grupos nominales realizados con operadores judiciales y penitenciarios proporcionan información cualitativa de gran valor, pero su tamaño reducido y su carácter no representativo del conjunto del sistema penal español condicionan el alcance de las conclusiones, algo habitual en investigaciones de naturaleza exploratoria. A ello se suma la dificultad inherente para acceder a muestras amplias de profesionales judiciales. Estas barreras hacen que la participación sea limitada y condicionan la posibilidad de obtener una representación más amplia del colectivo.

En segundo lugar, el rápido avance de los procesos de digitalización, algoritmización e inteligencia artificial implica que algunos resultados puedan verse afectados por la evolución tecnológica y regulatoria, una circunstancia inherente al estudio de fenómenos en constante transformación. Asimismo, aunque la tesis profundiza en la cuestión de los sesgos, su análisis empírico depende en buena medida de la disponibilidad de datos y del acceso a herramientas realmente empleadas en el ámbito penal, lo cual puede limitar la observación directa de ciertos procesos.

Por último, la medición de actitudes y percepciones en los estudios experimentales implica abordar constructos complejos y dinámicos, que pueden variar en función de factores legislativos, mediáticos o institucionales. A ello se suma que el uso de escenarios narrativos, aun siendo adecuado para investigar decisiones en contextos sensibles, puede activar interpretaciones diversas entre los participantes. Estas consideraciones no invalidan los hallazgos, pero invitan a contextualizarlos dentro de las características naturales de los métodos empleados y del objeto de estudio.

2. CONCLUSIONES GENERALES.

La transformación del sistema de justicia penal ha generado un escenario inédito en el que los avances tecnológicos conviven con nuevas exigencias sociales, institucionales y éticas. A lo largo de esta tesis se han analizado los efectos de la digitalización, la algoritmización y la inteligencia artificial sobre el trabajo de los operadores jurídicos y las percepciones ciudadanas A partir de este recorrido analítico, la presente investigación nos permite extraer las siguientes conclusiones generales:

CONCLUSIÓN 1. El proceso de transformación tecnológica 4.0 del sistema de justicia penal ha modificado de forma profunda la manera en que las sociedades actuales conciben el ejercicio de la autoridad, la toma de decisiones y la generación de legitimidad institucional. El sistema de justicia penal no ha quedado al margen de este cambio. La digitalización y la inteligencia artificial se han convertido en elementos que tensionan sus fundamentos y, a la vez, abren posibilidades para mejorar la eficacia, la confianza y el trabajo judicial. En este marco, los resultados de la investigación muestran que la incorporación de estas tecnologías no es solo una innovación instrumental, sino un fenómeno que redefine la arquitectura del quehacer judicial, policial y penitenciario, introduciendo nuevas modalidades de evaluación del riesgo, de acceso a la información y de estructuración del juicio profesional (Floridi y Cowls, 2019; Pasquale, 2015).

CONCLUSIÓN 2. La digitalización, la algoritmización y el uso de inteligencia artificial en el sistema penal se perciben como procesos ambivalentes, con capacidad para aportar mejoras sustanciales, pero también para generar riesgos si no cuentan con una supervisión sólida y con marcos de gobernanza adecuados. Tanto la ciudadanía como los operadores jurídicos coinciden en que el futuro de la justicia no reside en la automatización, sino en la ampliación de las capacidades humanas mediante tecnologías que funcionen como apoyo. La inteligencia artificial se valora por su potencial para mejorar el análisis de información, reducir ciertos errores y aportar coherencia en tareas complejas. Sin embargo, su aceptación depende de que no sustituya el

juicio experto, respete la autonomía decisoria y se someta a controles éticos y democráticos claros (O Neil, 2016).

CONCLUSIÓN 3. Conforme a los estudios realizados se puede decir que existe una preferencia entre los operadores jurídicos y por la propia ciudadanía por los modelos de colaboración entre personas y máquinas, apostando por modelos Human-In-the-Loop. Los operadores jurídicos señalan la necesidad de que estas herramientas optimicen tareas técnicas sin deshumanizar los procesos y destacan requisitos como la transparencia algorítmica, la formación especializada, la mitigación de sesgos y la preservación de un control humano efectivo. La ciudadanía, por su parte, asocia su aceptación no a factores sociodemográficos, sino a percepciones de imparcialidad, objetividad y proporcionalidad, mostrando una predisposición positiva solo cuando existen garantías de supervisión humana, límites funcionales claros y marcos regulatorios sólidos. En conjunto, estas percepciones indican que la aceptación de la justicia digital dependerá menos del despliegue técnico que de su integración en un modelo de gobernanza alineado con las demandas sociales.

CONCLUSIÓN 4. Existe una correspondencia entre los resultados obtenidos y los principios que, con posterioridad, quedaron plasmados en el Reglamento de Inteligencia Artificial de la Unión Europea. La insistencia de los participantes en contar con una supervisión humana significativa refleja uno de los pilares centrales del marco para sistemas de alto riesgo. Sin embargo, los estudios también ponen de relieve un punto aún sin resolver: cómo debe concretarse esa supervisión en cuanto a las fases del proceso, su alcance, las responsabilidades involucradas y las garantías procedimentales. Del mismo modo, las inquietudes sobre sesgos y discriminación coinciden con las obligaciones de gestión de riesgos y calidad de datos previstas tanto en el reglamento como en estándares internacionales, como los de la OECD (2019) y la UNESCO (2021), aunque sigue habiendo cierta incertidumbre entre los operadores respecto a los mecanismos prácticos para identificarlos y corregirlos.

CONCLUSION 5. Los operadores jurídicos y penitenciarios coinciden en que la formación en las nuevas tecnologías como los algoritmos o la inteligencia artificial es un requisito indispensable. Sin competencias digitales no puede existir una supervisión humana real. La ciudadanía, por su parte, demanda alfabetización tecnológica básica para comprender qué sistemas usa la justicia penal, por qué se aplican, sus límites y sus riesgos. Sin este doble proceso formativo, profesional y social, la digitalización no podrá implementarse de forma legítima ni responsable.

CONCLUSIÓN 6. La transformación tecnológica no reduce la importancia del componente humano en la administración de justicia; al contrario, incrementa la complejidad del juicio profesional y redefine las exigencias propias de la función jurisdiccional. La incorporación de tecnologías algorítmicas añade nuevas capas de información, métricas y criterios de análisis que exigen no solo entender su funcionamiento básico, sino también interpretar críticamente sus recomendaciones y reconocer sus limitaciones. Esto demanda que los operadores jurídicos desarrollen competencias adicionales: identificar posibles sesgos en los datos, evaluar la solidez metodológica de los modelos, contextualizar los resultados dentro del marco normativo aplicable y, sobre todo, contrastar las salidas automatizadas con el conocimiento jurídico y la experiencia práctica acumulada. Lejos de permitir la delegación de decisiones sensibles en sistemas digitales, estas tecnologías deben emplearse como herramientas de apoyo que amplían la información disponible sin sustituir el razonamiento humano. Así, la inteligencia artificial no reemplaza al profesional, sino que lo orienta hacia funciones de supervisión, análisis crítico y control de los sistemas utilizados. El profesional pasa a ser garante de que la tecnología se aplique conforme a los principios de proporcionalidad, imparcialidad y respeto de los derechos fundamentales. La transformación tecnológica no desplaza la centralidad humana, sino que la refuerza: solo a partir de un juicio experto, consciente y adecuadamente formado puede la justicia digital operar con fiabilidad.

CONCLUSIÓN 7. La integración de nuevas tecnologías en el sistema de justicia

penal evidencia que la aceptación de cualquier innovación depende en gran medida del nivel de comprensión y participación de la ciudadanía. Aunque las personas no interactúen directamente con los sistemas tecnológicos, sus decisiones y efectos inciden sobre derechos, oportunidades y condiciones, lo que hace indispensable que la población disponga de información accesible para interpretar el funcionamiento, alcance y límites de las herramientas. En contextos caracterizados por una brecha de conocimiento, tienden a surgir posturas sociales polarizadas que dificultan su adopción. Por ello, la construcción de confianza pública requiere no solo marcos técnicos y regulatorios sólidos, sino también procesos de alfabetización social y mecanismos de participación que permitan a la ciudadanía entender el uso de estas herramientas.

CONCLUSION 8. La incorporación de tecnologías basadas en inteligencia artificial en las administraciones públicas pone de manifiesto que la evaluación normativa y técnica, aunque imprescindible, no garantiza por sí sola la adopción efectiva de estas herramientas. La capacidad de una organización para integrar una tecnología depende, de manera determinante, de la aceptación y disposición de quienes deben utilizarla en su práctica cotidiana. Los datos comparados muestran que la velocidad del despliegue tecnológico supera con frecuencia la adaptación organizativa y la capacitación del personal, generando brechas entre la conformidad legal de los sistemas y su uso real. De ahí que resulte necesario complementar los marcos regulatorios con instrumentos que permitan comprender cómo perciben los profesionales estas tecnologías, qué niveles de confianza generan, qué barreras encuentran y bajo qué condiciones están dispuestos a incorporarlas. La transformación digital no es solo un proceso técnico, sino también humano y organizativo: sin una base de aceptación profesional, incluso las soluciones más robustas desde el punto de vista jurídico y ético pueden fracasar en la práctica.

CONCLUSIÓN 9. La incorporación de nuevas tecnologías en el sistema de justicia penal se produce en un contexto donde las personas no perciben

todos los ámbitos del sistema del mismo modo ni atribuyen el mismo riesgo a cada tipo de decisión. Estas diferencias muestran que no es posible aplicar soluciones uniformes a todos los espacios en los que interviene la inteligencia artificial. Las políticas públicas necesitan adaptarse a las características propias de cada ámbito y al impacto que sus decisiones pueden tener sobre la ciudadanía. Esto exige combinar consideraciones técnicas, organizativas e institucionales para que la regulación y la práctica se ajusten a las necesidades reales del sistema y a las expectativas sociales sobre su funcionamiento. Un enfoque así facilita que la gobernanza algorítmica se mantenga alineada con los principios del Estado de Derecho y con las demandas de quienes participan o se ven afectados por estas tecnologías.

En última instancia, la transformación digital del sistema de justicia penal, impulsada por la digitalización, la algoritmización y la inteligencia artificial, revela un escenario complejo en el que la innovación tecnológica convive con profundas exigencias sociales, éticas e institucionales. La investigación muestra que estas herramientas no solo reconfiguran los modos de producir información, evaluar riesgos o fundamentar decisiones, sino que introducen desafíos para la legitimidad democrática y el ejercicio de la autoridad. Tanto operadores jurídicos como ciudadanía coinciden en que el futuro de la justicia no depende de la automatización, sino de la ampliación de las capacidades humanas mediante tecnologías de apoyo, integradas bajo modelos colaborativos y supervisión significativa. La aceptación social de estos sistemas se vincula a percepciones de imparcialidad, transparencia, proporcionalidad y control humano, lo que evidencia que su aceptación no se sostiene únicamente en su rendimiento técnico. Los retos identificados subrayan que la transformación 4.0 solo será aceptada si se orienta hacia un modelo que potencie lo humano, limite los riesgos y consolide una justicia digital compatible con los principios del Estado de derecho. En definitiva, la digitalización del sistema penal no debe concebirse como un reemplazo de la labor humana, sino como una oportunidad para fortalecerla y garantizar que la tecnología permanezca, siempre, al servicio de la ciudadanía.

3. GENERAL CONCLUSIONS.

The transformation of the criminal justice system has produced an unprecedented context in which technological advances intersect with emerging social, institutional, and ethical demands. Throughout this thesis, the effects of digitalization, algorithmic processes, and artificial intelligence on the work of legal practitioners and on public perceptions have been examined. Building on this analytical trajectory, the present research allows us to draw the following overarching conclusions:

CONCLUSION 1. The technological transformation 4.0 of the criminal justice system has profoundly reshaped the ways in which contemporary societies conceive the exercise of authority, decision-making processes, and the production of institutional legitimacy. The criminal justice system has not remained unaffected by this shift. Digitalization and artificial intelligence have become elements that place pressure on its foundational principles while simultaneously opening possibilities to enhance effectiveness, public trust, and judicial performance. In this context, the findings of the research show that the incorporation of these technologies is not merely an instrumental innovation, but a phenomenon that redefines the architecture of judicial, policing, and correctional work. It introduces new modalities of risk assessment, information access, and the structuring of professional judgement (Floridi & Cowls, 2019; Pasquale, 2015).

CONCLUSION 2. Digitalization, algorithmic processes, and the use of artificial intelligence in the criminal justice system are perceived as ambivalent developments: they hold the potential to deliver significant improvements while also generating risks if not supported by strong oversight and appropriate governance frameworks. Both citizens and legal practitioners agree that the future of justice does not lie in automation, but rather in the expansion of human capabilities through technologies that operate as forms of support. Artificial intelligence is valued for its potential to improve information analysis, reduce certain errors, and bring greater coherence to complex tasks. However, its acceptance depends on ensuring that it does not

replace expert judgement, that it respects decisional autonomy, and that it remains subject to clear ethical and democratic safeguards (O'Neil, 2016).

CONCLUSION 3. The studies conducted indicate a clear preference among legal practitioners and the general public for collaborative human–machine models, favouring Human-in-the-Loop approaches. Legal practitioners emphasize the need for these tools to optimise technical tasks without dehumanizing procedures, and they highlight requirements such as algorithmic transparency, specialized training, bias mitigation, and the preservation of effective human control. The public, for its part, grounds its acceptance not in sociodemographic factors but in perceptions of impartiality, objectivity, and proportionality, expressing a positive predisposition only when there are guarantees of human oversight, clearly defined functional limits, and robust regulatory frameworks. Taken together, these perceptions suggest that the acceptance of digital justice will depend less on its technical deployment and more on its integration within a governance model aligned with social expectations.

CONCLUSION 4. A correspondence can be observed between the results obtained and the principles later codified in the European Union Artificial Intelligence Act. Participants' insistence on meaningful human oversight reflects one of the central pillars of the regulatory framework for high-risk systems. However, the studies also highlight an unresolved issue: how such oversight should be operationalized in terms of process stages, scope, responsibilities, and procedural safeguards. Likewise, concerns regarding bias and discrimination align with the risk-management and data-quality obligations set out both in the regulation and in international standards, such as those of the OECD (2019) and UNESCO (2021). Nonetheless, some degree of uncertainty persists among practitioners regarding the practical mechanisms for identifying and correcting such issues.

CONCLUSION 5. Legal and correctional practitioners agree that training in new technologies - such as algorithms and artificial intelligence - is an essential requirement. Without digital competencies, genuine human

oversight cannot exist. The public, for its part, calls for basic technological literacy in order to understand which systems are used in the criminal justice process, why they are applied, and what their limitations and risks are. Without this dual process of professional and societal education, digitalization cannot be implemented in a legitimate or responsible manner.

CONCLUSION 6. Technological transformation does not diminish the importance of the human component in the administration of justice; on the contrary, it increases the complexity of professional judgement and redefines the demands inherent to judicial functions. The incorporation of algorithmic technologies introduces new layers of information, metrics, and analytical criteria that require not only an understanding of their basic functioning but also a critical interpretation of their recommendations and awareness of their limitations. This requires legal practitioners to develop additional competencies: identifying potential data biases, evaluating the methodological soundness of models, contextualizing results within the applicable legal framework, and, above all, contrasting automated outputs with legal knowledge and accumulated practical experience. Far from enabling the delegation of sensitive decisions to digital systems, these technologies must be used as support tools that expand the available information without replacing human reasoning. Artificial intelligence does not supplant the professional; rather, it orients their work toward functions of oversight, critical analysis, and control of the systems employed. The professional becomes the guarantor that technology is applied in accordance with the principles of proportionality, impartiality, and respect for fundamental rights. Technological transformation does not displace human centrality; it reinforces it. Only through expert, conscious, and properly trained judgement can digital justice operate reliably.

CONCLUSION 7. The integration of new technologies into the criminal justice system demonstrates that the acceptance of any innovation largely depends on the public's level of understanding and engagement. Even when individuals do not interact directly with technological systems, their

decisions and effects shape rights, opportunities, and conditions. This makes it essential for the population to have access to information that enables them to interpret how these tools function, as well as their scope and limitations. In contexts marked by knowledge gaps, polarized social attitudes tend to emerge, complicating their adoption. For this reason, building public trust requires not only robust technical and regulatory frameworks but also processes of social literacy and participatory mechanisms that allow citizens to understand the use of these tools.

CONCLUSION 8. The incorporation of artificial intelligence–based technologies into public administrations shows that normative and technical evaluation, although indispensable, does not by itself guarantee the effective adoption of such tools. An organization's capacity to integrate a technology depends fundamentally on the acceptance and willingness of those who must use it in their daily work. Comparative data indicate that the pace of technological deployment often outstrips organizational adaptation and staff training, generating gaps between the legal compliance of systems and their actual use. Consequently, regulatory frameworks must be complemented by instruments that help elucidate how professionals perceive these technologies, the levels of trust they inspire, the barriers they encounter, and the conditions under which they are prepared to incorporate them. Digital transformation is not only a technical process but also a human and organizational one: without a foundation of professional acceptance, even the most legally and ethically robust solutions may fail in practice.

CONCLUSION 9. The introduction of new technologies into the criminal justice system takes place in a context where individuals do not perceive all areas of the system in the same way, nor do they attribute equal levels of risk to every type of decision. These differences indicate that uniform solutions cannot be applied across all domains in which artificial intelligence intervenes. Public policies must be adapted to the specific characteristics of each area and to the impact that decisions in those areas may have on the public. This requires combining technical, organizational, and institutional

considerations so that regulation and practice align with the system's actual needs and with societal expectations regarding its functioning. Such an approach facilitates maintaining algorithmic governance in alignment with the principles of the rule of law and with the expectations of those who participate in or are affected by these technologies.

Ultimately, the digital transformation of the criminal justice system - driven by digitalization, algorithmic processes, and artificial Intelligence - reveals a complex scenario in which technological innovation coexists with profound social, ethical, and institutional demands. The research shows that these tools not only reconfigure the ways in which information is produced, risks are assessed, and decisions are justified, but also introduce challenges for democratic legitimacy and the exercise of authority. Both legal practitioners and the public agree that the future of justice does not hinge on automation, but on the enhancement of human capacities through supportive technologies, integrated within collaborative models and meaningful oversight.

The social acceptance of these systems is tied to perceptions of impartiality, transparency, proportionality, and human control, indicating that their legitimacy cannot rest solely on technical performance. The challenges identified underscore that this technological transformation will only be accepted if it is oriented toward a model that strengthens the human dimension, mitigates risks, and consolidates a form of digital justice compatible with the principles of the rule of law. In essence, the digitalization of the criminal justice system should not be conceived as a replacement for human work, but as an opportunity to reinforce it and ensure that technology remains, always, in the service of the public.

REFERENCIAS

Adams, I. T. (2025). Automation and artificial intelligence in police body-worn cameras: Experimental evidence of impact on perceptions of fairness among officers. *Journal of Criminal Justice, 97*, 102373. https://doi.org/10.1016/j.jcrimjus.2025.102373

Agudo Díaz, U. (2021). *La influencia de los algoritmos en las decisiones y juicios humanos: Experimentos en contextos de política, citas y arte* (Tesis doctoral, Universidad de Deusto). https://osf.io/t7dbv/?view_only=21d75e0817984300bea0340bc64d7539

Aguilar, L. J. (2021). *Internet de las cosas: Un futuro hiperconectado: 5G, inteligencia artificial, Big Data, Cloud, Blockchain, Ciberseguridad.* Alpha Editorial

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17*(4), 351–371. https://doi.org/10.1177/1094428114547952

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 5*0(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of applied social psychology, 32*(4), 665-683. https://doi.org/10.1111/j.1559-1816.2002.tb00236.x

Alalwan, A. A., Dwivedi, Y. K., & Rana, N. P. (2017). Factors influencing adoption of mobile banking by Jordanian bank customers: Extending UTAUT2 with trust. *International Journal of Information Management*, 37(3), 99–110. https://doi.org/10.1016/j.ijinfomgt.2017.01.002

Alarie, B., Niblett, A., & Yoon, A. H. (2018). How artificial intelligence will affect the practice of law. *University of Toronto Law Journal*, *68*(Supplement 1), 106–124. https://doi.org/10.3138/utlj.2017-0052

Alemán Aróstegui, L. (2023). El uso de RISCANVI en la toma de decisiones penitenciarias. *Estudios Penales y Criminológicos, 44*(Ext.), 1–43. https://doi.org/10.15304/epc.44.8884

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science, 2*, e93. http://dx.doi.org/10.7717/peerj-cs.93

Allport, G. W. (1935). Attitudes. In *A Handbook of Social Psychology* (pp. 798–844). Clark University Press.

Alnemr, N. (2024). Democratic self-government and the algocratic shortcut: the democratic harms in algorithmic governance of society. *Contemporary political theory, 23*(2), 205-227. https://doi.org/10.1057/s41296-023-00656-y

Alzahrani, A., & Alzahrani, A. (2025). Understanding ChatGPT adoption in universities: the impact of faculty TPACK and UTAUT2. *RIED-Revista Iberoamericana de Educación a Distancia*, *28*(1), 37-58. https://doi.org/10.5944/ried.28.1.41498

Anderson, J. R. (2015). *Cognitive psychology and its implications* (8th ed.). Worth Publishers.

Andrews, D. A., & Bonta, J. (2010). *The Psychology of Criminal Conduct* (5th ed.). New Providence, NJ: LexisNexis Matthew Bender.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2011). The risk-need-responsivity (RNR) model: Does adding the good lives model contribute to effective crime prevention?. *Criminal justice and behavior, 38*(7), 735-755. https://doi.org/10.1177/0093854811406356

Andrews, D.A., & Bonta, J. (1995). *The Level of Service Inventory-Revised*. Toronto, Ontario, Canada: Multi-Health Systems.

Andrews, D.A., Bonta, J., & Wormith, S.J. (2008). *The Level of Service/Risk-Need-Responsivity (LS/RNR)* . Toronto, Ontario, Canada: Multi-Health Systems.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arruda, H., Silva, E. R., Lessa, M., Proença, D., & Bartholo, R. (2022). VOSviewer and Bibliometrix. *Journal of the Medical Library Association, 110*(3), 392. https://doi.org/10.5195/jmla.2022.1434

Ashley, K. D. (2017). *Artificial intelligence and legal analytics: New tools for law practice in the digital age.* Cambridge University Press.

Bagozzi, R. P. (2007). The Legacy of the Technology Acceptance Model and a proposal for a paradigm shift. *Journal of the Association for Information Systems*, *8*(4), 244–254. https://doi.org/10.17705/1jais.00122

Bansak, K. (2019). Can nonexperts really emulate statistical learning methods? A comment on "The accuracy, fairness, and limits of predicting recidivism". *Political Analysis, 27*(3), 370–380. https://doi.org/10.1017/pan.2018.55

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review, 104*(3), 671-732. http://dx.doi.org/10.15779/Z38BG31

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities.* MIT press.

Barona Vilar, S. (2019a). Cuarta revolución industrial (4.0.) o ciberindustria en el proceso penal: revolución digital, inteligencia artificial y el camino hacia la robotización de la justicia. *Revista jurídica digital UANDES, 3*(1), 1-17.

Barona Vilar, S. (2019b). Inteligencia artificial o la algoritmización de la vida y de la justicia: ¿Solución o problema?. *Revista Boliviana de Derecho, 28*, 18–49.

Barry, B. M. (2021). *How judges judge. Empirical insights into judicial decision-*

making. Routledge.

Barysé, D., & Sarel, R. (2023). Algorithms in the court: Does it matter which part of the judicial decision-making is automated? *Artificial Intelligence and Law.* https://doi.org/10.1007/s10506-022-09343-6

Baumgartner, F. R. (2020). *Suspect citizens: What 20 million traffic stops tell us about policing and race.* Cambridge University Press.

Beck, J., & Burri, T. (2024). From "human control" in international law to "human oversight" in the new EU act on artificial intelligence. In *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems* (pp. 104-130). Edward Elgar Publishing. https://doi.org/10.4337/9781802204131.00014

Bergdahl, J., Latikka, R., Celuch, M., Savolainen, I., Mantere, E. S., Savela, N., & Oksanen, A. (2023). Self-determination and attitudes toward artificial intelligence: Cross-national and longitudinal perspectives. *Telematics and Informatics, 82,* 102013. https://doi.org/10.1016/j.tele.2023.102013

Beriain, I. D. M. (2018). Does the use of risk assessments in sentences respect the right to due process? A critical analysis of the *Wisconsin v. Loomis* ruling. *Law, Probability and Risk, 17*(1), 45–53. https://doi.org/10.1093/lpr/mgy001

Berk, R. (2019). Machine learning risk assessments in criminal justice settings. Springer. https://doi.org/10.1007/978-3-030-02272-3

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3–44. https://doi.or10.1177/0049124118782533

Berlanga, V., Rubio Hurtado, M. J., & Vilà Baños, R. (2013). Cómo aplicar árboles de decisión en SPSS. *REIRE. Revista d'Innovació i Recerca en Educació, 2013, vol. 6, num. 1, p. 65-79*.

Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-

corruption tools for societies. Government Information Quarterly, 27(3), 264–271. https://doi.org/10.1016/j.giq.2010.03.001

Bieneck, S. (Ed.). (2009). *How adequate is the vignette technique as a research tool for psycho-legal research?* In M. E. Oswald, S. Bieneck, & J. Hupfeld-Heinemann (Eds.), *Social psychology of punishment of crime* (pp. 255–271). John Wiley & Sons Ltd.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21-34. https://doi.org/10.1016/j.cognition.2018.08.003

Binns, R. (2018). Algorithmic accountability and public reason. Philosophy & Technology, 31, 543–556. https://doi.org/10.1007/s13347-017-0263-5

Binns, R. (2022). Human Judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & governance*, *16*(1), 197-211. https://doi.org/10.1111/rego.12358

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3173574.3173951

Boba, R. (2019). *Predictive policing: Where is the evidence*. In Weisburd, D & Braga, A.A. (Eds.), *Police innovation. Contrasting perspectives.* 2d edition. Cambridge: Cambridge University Press.

Boden, M. A. (2016). *AI: Its nature and future*. Oxford University Press.

Boer, D. P., Hart, S., Kropp, P. R., y Webster, C. D. (1997). The SVR-20 Guide for assessment of sexual risk violence, *Mental Health, Law and Policy Institute, Simon Fraser University, Vancouver.*

Boix Palop, A. (2020). Los algoritmos son reglamentos: La necesidad de extender las

garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones. *Revista de Derecho Público. Teoría y Método, 1*, 223–269. https://doi.org/10.37417/RPD/vol_1_2020_33

Boletín Oficial del Estado. (2025). Ley Orgánica 1/2025, de 2 de enero, de medidas en materia de eficiencia del Servicio Público de Justicia. Madrid: BOE. https://www.boe.es/diario_boe/txt.php?id=BOE-A-2025-76

Bolívar, A. (1994). *La investigación educativa: de la teoría a la práctica*. Madrid: La Muralla.

Bond, C. E., & Jeffries, S. (2011). Indigeneity and the judicial decision to imprison: A study of Western Australia's higher courts. *The British Journal of Criminology, 51*(2), 256-277. https://doi.org/10.1093/bjc/azr001

Borum, R., Bartel, P., y Forth, A. (2005). Structured Assessment of Violence Risk in Youth. En Grisso, T., Vincent, G., y Seagrave, D. (Eds.), *Mental Health Screening and Assessment in Juvenile Justice*. New York: The Guilford Press.

Brandariz García, J. A. (2016). *El modelo gerencial-actuarial de penalidad. Eficiencia, riesgo y sistema penal*. Madrid: Dykinson.

Brantingham, P. J., Valasik, M., & Mohler, G. O. (2018). Does predictive policing lead to biased arrests? Results from a randomized controlled trial. *Statistics and public policy*, *5*(1), 1-6. https://doi.org/10.1080/2330443X.2018.1438940

Brantingham, P., y Brantingham, P. (2013). Crime pattern theory. En Wortley, R., y Townsley, M. (Eds.), *Environmental Criminology and Crime Analysis.* 1st edition. London: Willan.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior, 36*(1), 21–40. https://doi.org/10.1177/0093854808326545

Brennen, J. S., & Kreiss, D. (2016). *Digitalization.* En K. B. Jensen, R. T. Craig, J. Pooley,

& E. W. Rothenbuhler (Eds.), The international encyclopedia of communication theory and philosophy. Wiley-Blackwell. https://doi.org/10.1002/9781118766804.wbiect111

Briñol, P., Falces, C., y Becerra, A. (2006). *Actitudes*. En J. F. Morales, C. Huici, M. Moya, y E. Gaviria (Eds.), Psicología social (3rd ed.). McGraw-Hill..

Brittain, B. J., Georges, L., & Martin, J. (2021). Examining the predictive validity of the Public Safety Assessment (PSA). *Criminal Justice and Behavior, 48*(10), 1369–1389. https://doi.org/10.1177/00938548211005836

Bruun, M. H. (2024). Algorithmic governance, public participation and trust: Citizen–state relations in a smart city project. *Social Anthropology/Anthropologie Sociale, 32*(4), 13–30. https://doi.org/10.3167/saas.2024.320402

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 77–91.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public administration review*, *81*(5), 825-836. https://doi.org/10.1111/puar.13293

Buttarelli, G. (2019). Ethics and Data Protection in Artificial Intelligence: Continuing the Debate. European Data Protection Supervisor.

Campusano Droguett, R. F., & Giesen Espejo, J. B. (2024). La Ley de Inteligencia artificial de la UE: un hito normativo en la escena global. Actualidad Jurídica, 50 (julio), 105–136.

Carmena, M. (1997). Crónica de un desorden: Notas para reinventar la Justicia.

Madrid: Alianza Editorial.

Carrillo, F. J. G. (2007). Claves de la modernización de la justicia en España. Agenda Internacional, 14(25), 249-280.

Casiraghi, S., Burgess, J. P., & Lidén, K. (2021). Social acceptance and border control technologies. *Border Control and New Technologies*, 99-115.

Castellanos Claramunt, J. (2019). La democracia algorítmica: Inteligencia artificial, democracia y participación política. *Revista General de Derecho Administrativo, 50,* 1–32. Iustel.

Castelluccia, C., & Le Métayer, D. (2019). Understanding algorithmic decision-making: Opportunities and challenges. European Parliamentary Research Service (EPRS). Recuperado de https://www.europarl.europa.eu/thinktank

Castro-Toledo, F. J., & Gómez-Bellvís, A. B. (2024). Using Technology Democratically: Development of a Self-Assessment Tool to Support Urban Security Authorities. CrimRxiv. https://doi.org/10.21428/cb6ab371.57080d50

Castro-Toledo, F.J. (ed.) (2022). *La transformación algorítmica del sistema de justicia penal.* Aranzadi, Thomson Reuter.

Caterini, M. (2022). El sistema penal en la encrucijada ante el reto de la inteligencia artificial . *IDP. Revista de Internet, Derecho y Política, 35*, 1–19. https://doi.org/10.7238/idp.v0i35.392754

Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), https://doi.org/10.1098/rsta.2018.0080

CEPEJ (2018). European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment. Council of Europe.

CEPEJ. (2024). Eficacité et qualité de la justice en Europe: Rapport d'évaluation des systèmes judiciaires européens 2024. Conseil de l'Europe.

https://www.coe.int/en/web/portal/-/efficacit%C3%A9-et-qualit%C3%A9-de-la-justice-en-europe-le-conseil-de-l-europe-publie-son-rapport-2024

Cerezo-Martínez, P., Sánchez, A. N., & Castro-Toledo, F. J. (2024). Discriminatory bias (and how to prevent it) in the European Union: A review of the ethical and regulatory framework for artificial intelligence. [Información adicional no proporcionada].

Cerrillo i Martínez, A. (2019). El impacto de la inteligencia artificial en el derecho administrativo: ¿Nuevos conceptos para nuevas realidades técnicas? *Revista General de Derecho Administrativo, 50.*

Chan, J. (2021). *Policing in the digital age: Challenges and opportunities*. Oxford University Press.

Chatziathanasiou, K. (2022). Beware the lure of narratives: Hungry judges should not motivate the use of artificial intelligence in law. Artificial Intelligence & Law, 30(4), 345–372.

Chen, X.X. & Slade, E. (2024) Theory of Planned Behaviour: A review. In S. Papagiannidis (Ed), TheoryHub Book.

Choi, D. D., Harris, J. A., & Shen-Bayh, F. (2022). Ethnic bias in judicial decision-making: Evidence from criminal appeals in Kenya. Journal of African Law, 66(1), 23–49. https://doi.org/10.1017/S000305542100143X

Choung, H., David, P., & Ross, A. (2022). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human–Computer Interaction.* Advance online publication. https://doi.org/10.1080/10447318.2022.2050543

Christin, A. (2020). The ethnographer and the algorithm: Beyond the black box. *Theory and Society, 49*(5-6), 897–918.

Church, A. (1936). An unsolvable problem of elementary number theory. *American*

journal of mathematics, *58*(2), 345-363. https://doi.org/10.2307/2371045

Clark, J., Demircan, M., & Kettas, K. (2024, 25 de abril). Europe: The EU AI Act's relationship with data protection law – key takeaways. Blog Privacy Matters, DLA Piper. Recuperado de https://privacymatters.dlapiper.com/2024/04/europe-the-eu-ai-acts-relationship-with-data-protection-law-key-takeaways/

Colado, S., Gutiérrez, A., Vives, C. J., & Valencia, E. (2014). *Smart city: Hacia la gestión inteligente.* Marcombo.

Comisión Europea (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Comisión Europea. (2020). White Paper on Artificial Intelligence - A European Approach to Excellence and Trust. Brussels.

Comisión Europea. (2024). Artificial Intelligence Act: Ensuring Human Oversight and Transparency. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Comisión Europea. (2024). Informe sobre el Estado de Derecho 2024 – Capítulo sobre España. Bruselas: Comisión Europea. https://commission.europa.eu/document/download/2bd09a6f-ef56-494a-8303-e0de808ee981_es

Comisión Europea. (2024). Ley de IA (Artificial Intelligence Act) – Marco regulatorio de la IA. En Shaping Europe's Digital Future. Recuperado de https://digital-strategy.ec.europa.eu/es/policies/regulatory-framework-ai

Concannon, C., & Na, C. (2022). Examining racial and ethnic disparity in prosecutor's bail requests and downstream decision-making. Criminal Justice Policy Review, 33(2), 198–220.

Consejo de Europa / Comisión Europea para la Eficacia de la Justicia (CEPEJ). (2024). European judicial systems – CEPEJ Evaluation Report 2022: Data for Spain. Estrasburgo: Consejo de Europa. https://www.poderjudicial.es/cgpj/es/Temas/Estadistica-Judicial/Estadistica-por-temas/Aspectos-internacionales/Informes-Organismos-Extranjeros/CEPEJ--Comision-Europea-para-eficiencia-de-la-justicia-/

Consejo de Europa. (2023). *Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*. Council of Europe.

Consejo de la Unión Europea. (2024, 21 de mayo). Reglamento de Inteligencia artificial : el Consejo da luz verde definitiva a las primeras normas del mundo en materia de inteligencia artificial [Comunicado de prensa]. Recuperado de https://www.consilium.europa.eu/es/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/

Consejo General de la Abogacía Española. (2024). Informe sobre la eficiencia de la justicia 2024 (datos 2022). Consejo General de la Abogacía Española. https://www.abogacia.es/actualidad/noticias/informe-sobre-la-eficiencia-de-la-justicia-de-2024-datos-2022

Consejo General del Poder Judicial, Ministerio de Justicia, Fiscalía General del Estado, Comunidades Autónomas competentes & Instituto Nacional de Estadística. (2021). Plan Nacional de Estadística Judicial 2021-2024. Madrid: Comité Nacional de Estadística Judicial. https://datos.justicia.es/national-judicial-statistics-committee

Consejo General del Poder Judicial. (2018). La carga de trabajo es el principal factor de riesgo psicosocial de jueces y magistrados. Madrid: CGPJ. https://www.poderjudicial.es/cgpj/es/Poder-Judicial/En-Portada/La-carga-de-trabajo--principal-factor-de-riesgo-psicosocial-de-jueces-y-magistrados

Consejo General del Poder Judicial. (2022). *Marco de Competencias Digitales para el ámbito de la Justicia* [Documento técnico]. Centro de Estudios Jurídicos. Gobierno de España.

Consejo General del Poder Judicial. (2025). La sobrecarga de trabajo y las dificultades para conciliar, los obstáculos que más frenan el ascenso de las mujeres en la Carrera Judicial. Madrid: CGPJ. https://www.poderjudicial.es/cgpj/en/Judiciary/Panorama/La-sobrecarga-de-trabajo-y-las-dificultades-para-conciliar--los-obstaculos-que-mas-frenan-el-ascenso-de-las-mujeres-en-la-Carrera-Judicial

Consejo General del Poder Judicial. (2025). Resumen datos estadísticos por partidos judiciales 2024. Madrid: CGPJ. https://www.poderjudicial.es/cgpj/es/Temas/Estadistica-Judicial/Actividad-judicial-por-territorio/Resumen-datos-estadisticos-por-partidos-judiciales-2024

Cordella, A. (2020). Tecnologías digitales para mejorar los sistemas de justicia: Un conjunto de herramientas para la acción. Banco Interamericano de Desarrollo. https://publications.iadb.org/publications/spanish/document/Tecnologias-digitales-para-mejorar-los-sistemas-de-justicia-un-conjunto-de-herramientas-para-la-accion.pdf

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to Algorithms (3rd ed.).* MIT Press. https://dl.acm.org/doi/book/10.5555/1614191

Correia, P. M. A. R., Pereira, S. P. M., & Bilhim, J. A. D. F. (2024). Research of Innovation and Digital Transformation in Justice: A Systematic Review. Journal of Digital Technologies and Law, 2(1), 221-240. https://doi.org/10.21202/jdtl.2024.12

Cortina Orts, A. (2019). Ética de la inteligencia artificial . *Anales de la Real Academia de Ciencias Morales y Políticas, 96*, 379–394.

Cotino Hueso, L. (2022). Nuevo paradigma en las garantías de los derechos

fundamentales y una nueva protección de datos frente al impacto social y colectivo de la inteligencia artificial . In M. Bauzá Reilly (Coord.) & L. Cotino Hueso (Dir.), *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas* (pp. 69–105). Aranzadi.

Cotino Hueso, L. (2024). El uso jurisdiccional de la inteligencia artificial : Habilitación legal, garantías necesarias y la supervisión por el CGPJ. *Actualidad Jurídica Iberoamericana, 21*, 494–527.

Cotino Hueso, L. C., & Reilly, M. B. (2022). Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas. Thomson Reuters Aranzadi.

Cotino Hueso, L., & Simón Castellano, P. (Dirs.). (2024). *Tratado sobre el reglamento de inteligencia artificial de la Unión Europea*. Aranzadi.

Council of Europe. (2024). Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225). https://rm.coe.int/cets-225-framework-convention-on-ai-and-human-rights/1680ac16ed

Council of Europe. (2024). Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law: Explanatory Report (CM(2024)52-addfinal). Committee of Ministers. https://www.coe.int/cm

Crano, W. D., Cooper, J., & Forgas, J. P. (2010). The psychology of attitudes and attitude change. Taylor & Francis.

Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.

Cummings, M. L. (2014). Man versus machine or man + machine? *IEEE Intelligent Systems, 29*(5), 62–69. https://doi.org/10.1109/MIS.2014.87.

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, *108*(17), 6889-

6892.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in policing and security. Harvard Business Review, 97(5), 112-121.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340. https://doi.org/10.2307/249008

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management science*, *35*(8), 982-1003.

de Fine Licht, K., & de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. AI & society, 35(4), 917-926. https://doi.org/10.1007/s00146-020-00960-w

De la Cuétara, J. L. (2023). Derecho Penal y la Inteligencia artificial : Retos y Perspectivas Regulatorias. Revista Electrónica de Ciencia Penal y Criminología, 25, 1-32.

De Sousa, W. G., Fidelis, R. A., de Souza Bermejo, P. H., da Silva Gonçalo, A. G., & de Souza Melo, B. (2022). Artificial intelligence and speedy trial in the judiciary: Myth, reality or need? A case study in the Brazilian Supreme Court (STF). *Government information quarterly*, *39*(1), 101660. https://doi.org/10.1016/j.giq.2021.101660

Deeks, A. S. (2019). The Judicial Demand for Explainable Artificial Intelligence. Columbia Law Review, 119(7), 1829-1850.

Defensor del Pueblo. (2025). Informe anual 2024. Defensor del Pueblo. https://www.defensordelpueblo.es/wp-

content/uploads/2025/03/Defensor-del-Pueblo_Informe-anual-2024.pdf

Dehghanniri, H., & Borrion, H. (2019). Crime scripting: A systematic review. *European Journal of Criminology, 18*(4), 504-525. https://doi.org/10.1177/1477370819850943

Delbecq, A. L., & Van de Ven, A. H. (1971). A group process model for problem identification and program planning. *The journal of applied behavioral science*, *7*(4), 466-492.

DeMichele, M., Baumgartner, P., Barrick, K., Wenger, M., & Comfort, M. (2020). Public Safety Assessment: A re-validation and assessment of predictive utility and differential prediction by race in Kentucky. *Criminology & Public Policy, 19*(3), 1061–1087. https://doi.org/10.1111/1745-9133.12481

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.

Donoghue, J. (2017). The rise of digital justice: Courtroom technology, public participation and access to justice. *The Modern Law Review, 80*(6), 995–1025. https://doi.org/10.1111/1468-2230.12300

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *https://doi.org/10.48550/arXiv.1702.08608*

Dou, L., & Dou, X. (2025). Towards just AI: Challenges and solution framework for algorithmic discrimination in judicial system. *International Journal of Digital Law and Governance, 2*(1), 39-81. https://doi.org/10.1515/ijdlg-2024-0020

Douglas, K. S., Hart, S. D., Webster, C. D., y Belfrage, H. (2013). Assesing risk for violence. *Version 3, Mental Health, Law and Policy Institute, Simon Fraser University, Burnaby, BC*.

Draws, T., Szlávik, Z., Timmermans, B., Tintarev, N., Varshney, K. R., & Hind, M.

(2021). Disparate impact diminishes consumer trust even for advantaged users. In *International Conference on Persuasive Technology* (pp. 135-149). Cham: Springer International Publishing.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1). https://doi.org/10.1126/sciadv.aao5580

Dwivedi, Y. K., Rana, N. P., Chen, H., & Williams, M. D. (2011). A Meta-analysis of the Unified Theory of Acceptance and Use of Technology (UTAUT). In Governance and Sustainability in Information Systems. Managing the Transfer and Diffusion of IT: IFIP WG 8.6 International Working Conference, Hamburg, Germany, September 22-24, 2011. Proceedings (pp. 155-170). Springer Berlin Heidelberg.

Dwivedi, Y. K., Rana, N. P., Jeyaraj, A., Clement, M., & Williams, M. D. (2019). Re-examining the Unified Theory of Acceptance and Use of Technology (UTAUT): Towards a revised theoretical model. *Information Systems Frontiers*, 21, 719–734. https://doi.org/10.1007/s10796-017-9774-y

Ecker, A., Ennser-Jedenastik, L., & Haselmayer, M. (2019). Gender bias: Evidence for leniency toward token women. *European Journal of Political Research, 58*(4), 999–1021.

Eerland, A., & Rassin, E. (2010). Biased evaluation of incriminating and exonerating (non)evidence. Applied Cognitive Psychology, 24(1), 122–135.

Elvin, J. (2010). The continuing use of problematic sexual stereotypes in judicial decision-making. *Feminist Legal Studies, 18*(3), 275–297.

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision-making. *Psychological Science, 17*(3), 189–195.

Enqvist, L. (2023). 'Human oversight' in the EU artificial intelligence act: what, when and by whom? *Law, Innovation and Technology*, *15*(2), 508–535. https://doi.org/10.1080/17579961.2023.2245683

Entcheva, K., & Mazilescu, I. (2024). Artificial intelligence and digitalisation of judicial cooperation: the main provisions in recent EU Legislation. *Eucrim: the European Criminal Law Associations' fórum,* (3), 202-205.

Escajedo, R. (2024). Reconocimiento biométrico remoto y excepciones en el Reglamento de IA. En L. Cotino Hueso & M. Simón Castellano (Eds.), Tratado sobre el Reglamento de Inteligencia Artificial (pp. 218-222). Tirant lo Blanch.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

Eurojust & eu-LISA. (2021). Artificial intelligence and cross-border cooperation in criminal justice: Report. European Commission. https://www.eurojust.europa.eu/sites/default/files/assets/artificial-intelligence-cross-border-cooperation-criminal-justice-report.pdf

European Commission, Directorate-General for Justice and Consumers. (2025). *2025 Rule of Law Report: Communication and country chapters*. Publications Office of the European Union. https://commission.europa.eu/publications/2025-rule-law-report-communication-and-country-chapters_en

European Commission. (2019). Science with and for society in Horizon 2020. Brussels: Directorate-General for Research and Innovation.

European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

European Commission. (2021). Proposal for a regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act). https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

European Commission. (2021). Regulation on a European approach for Artificial Intelligence. https://ec.europa.eu/digital-strategy

European Commission. (2024). *Artificial Intelligence Act* (Regulation (EU) 2024/1689) — human oversight requirements (Art. 14).

European Data Protection Supervisor. (2021). Opinion 20/21 on the Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Brussels.

Eysenck, M. W., & Keane, M. T. (2020). Cognitive psychology: A student's handbook. Psychology press.

Fair Trials. (2023, 16 de junio). EU Parliament approves landmark AI law. Recuperado de https://www.fairtrials.org/articles/news/eu-parliament-approves-landmark-ai-law/

Farfán Intriago, J. L., Farfán Largacha, J. A., Farfán Largacha, B., & Núñez Vera, J. P. (2023). Inteligencia artificial y Derecho: ¿La justicia en manos de la IA? Frónesis. *Frónesis, 30*(2), 173-197. https://produccioncientificaluz.org/index.php/fronesis/article/view/40853

Fariña, F., Arce, R., & Novo, M. (2003). Aislamiento del heurístico de anclaje y su impacto en decisiones judiciales. *Revista Española de Psicología Jurídica, 10*(2), 45–67.

Fazel, S., Burghart, M., Fanshawe, T., Gil, S. D., Monahan, J., & Yu, R. (2022). The predictive performance of criminal risk assessment tools used at sentencing: Systematic review of validation studies. Journal of Criminal Justice, 81, 101902. https://doi.org/10.1016/j.jcrimjus.2022.101902

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297–327. https://doi.org/10.1146/annurev.psych.54.101601.145225

Fedorczyk, F. (2024). Navigating the dichotomy of smart prisons: between surveillance and rehabilitation. *Law, Innovation and Technology*, *16*(1), 243-260. https://doi.org/10.1080/17579961.2024.2313793

Feenberg, A. (2002). Transforming technology: A critical theory revisited. Oxford University Press.

Feldman, S. M. (2006). Empiricism, religion, and judicial decision-making. University of Pennsylvania Law Review, 154(3), 537–583.

Fenoll, J. N. (2025). Los sesgos cognitivos y la prueba: huyendo de la intuición del juez. *InDret*, (1), 382-405.

Férez-Mangas, D. (2017a). *Eficàcia del RisCanvi Complet en la predicció del trencament de permís de sortida*. Barcelona: Centro de Estudis Jurídicos y Formación Especializada.

Férez-Mangas, D. (2017b). *Factores de riesgo y predicción del quebrantamiento de los permisos penitenciarios de salida (Tesis doctoral,* Universidad de Barcelona).

Ferguson, A. G. (2017). *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement.* NY: NYU Press.

Fernández-Güell, J. M. (2015). Ciudades inteligentes: la mitificación de las nuevas tecnologías como respuesta a los retos de las ciudades contemporáneas. *Economía industrial*, (395), 17-28.

Ferrara, E. (2024). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci, 6(1), 3. https://doi.org/10.3390/sci6010003

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Fine, A., Berthelot, E. R., & Marsh, S. (2025). Public Perceptions of Judges' Use of AI Tools in Courtroom Decision-Making: An Examination of Legitimacy, Fairness, Trust, and Procedural Justice. *Behavioral Sciences, 15*(4), 476. https://doi.org/10.3390/bs15040476

Fine, A., Miller, M. K., & Le, S. (2024). Content Analysis of Judges' Sentiments Toward Artificial Intelligence Risk Assessment Tools. *Criminology, Criminal Justice, Law & Society 24* (2), 31–46. https://doi.org/10.54555/CCJLS.8169.84869

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human perception and performance, 1*(3), 288.

Fishbein, M., & Ajzen, I. (1975). Belief, attitude, intention and behavior: An introduction to theory and research. Addison-Wesley.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior : the reasoned action approach.* Psychology Press.

Floridi, L. and Cowls, J. (2022). A Unified Framework of Five Principles for AI in Society . In Machine Learning and the City, S. Carta (Ed.). https://doi.org/10.1002/9781119815075.ch45

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People,An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689-707.

Forgas, J. P. (2008). The role of affect in attitudes and attitude change. In W. D. Crano & R.Prislin (Eds.), Attitudes and attitude change (pp. 131–158). New York: Psychology Press.

Forza, A., Menegon, G., & Rumiati. (2024). *El juez emotivo: La decisión entre razón y emoción.* Marcial Pons.

Fosch-Villaronga, E., & Malgieri, G. (2024). Queering the ethics of AI. In *Handbook on the Ethics of Artificial Intelligence*. Edward Elgar Publishing. (En prensa).

Fundación Aranzadi La Ley. (2024). *Informe 2024 del Observatorio de la Actividad de la Justicia.*

Gabinete de Estudios del CGPJ & Consejo General de Procuradores de España. (2024). La Justicia en España 2024: El atasco, un problema estructural. Madrid: CGPE. https://www.icpb.es/wp-content/uploads/2024/10/Informe-La-Justicia-en-Espana-2024-Procura_compressed.pdf

Gaede, J., & Rowlands, I. H. (2018). Visualizing social acceptance research: A bibliometric review of the social acceptance literature for energy technology and fuels. *Energy research & social science*, *40*, 142-158.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349-360. https://doi.org/10.1093/poq/nfp031

Galli, F., & Sartor, G. (2023). AI approaches to predictive justice: A critical assessment. *Humanities and Rights Global Network Journal*, *5*(2).

Gansser, O. A., & Reich, C. S. (2021). A new acceptance model for artificial intelligence with extensions to UTAUT2: An empirical study in three segments of application. *Technology in Society, 65*, 101535.

García-Sánchez, I. M., Rodríguez-Ariza, L., & Frías-Aceituno, J. V. (2013). The cultural system and integrated reporting. *International Business Review, 22*(5), 828-838. https://doi.org/10.1016/j.ibusrev.2013.01.007

Gendreau, P., Little, T., y Goggin, C (1996). A meta-analysis of the predictors of adult offender recidivism: What works!. *Criminology*, 34(4).

Gibson, J. L., & Caldeira, G. A. (1995). The legitimacy of transnational legal institutions: Compliance, support, and the European Court of Justice. *American Journal of Political Science, 39*(2), 459–489. https://doi.org/10.2307/2111621

Gigerenzer, G., Todd, P. M., & ABC Research Group, T. (2000). Simple heuristics that make us smart. Oxford University Press.

Gil-Monte, P. R., López-Vílchez, J., Llorca-Rubio, J. L., & Sánchez Piernas, J. (2016). Prevalencia de riesgos psicosociales en personal de la administración de justicia de la Comunidad Valenciana (España). Liberabit, 22(1), 7-19.

Gill, R. D., Kagan, M., & Marouf, F. (2019). The impact of maleness on judicial decision making: Masculinity, chivalry, and immigration appeals. Politics, Groups, and

Identities, 7(3), 509–528.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121-127. ttps://doi.org/10.1136/amiajnl-2011-000089

Gómez Bellvís, A. B., & Bañón, Z. E. (2022). El quién y el qué de la política criminal: una aproximación general a sus actores e instituciones. En Manual de política criminal (pp. 75-104). Atelier.

González-Anleo, J. M., Delbello, L., Martínez-Gonzálo, J. M., & Gómez, A. (2024). Sociodemographic impact on the adoption of emerging technologies. Journal of Small Business Strategy, 34(2), 42-50.

González-Bravo, L., & Valdivia-Peralta, M. (2015). Posibilidades para el uso del modelo de aceptación de la tecnología (TAM) y de la teoría de los marcos tecnológicos para evaluar la aceptación de nuevas tecnologías para el aseguramiento de la calidad en la educación superior chilena. *Revista electrónica educare*, *19*(2), 181-196.

González-Fuster, G., & Van Brakel, R. (2020). The Fundamental Rights Implications of AI and Predictive Analytics in Law Enforcement. European Parliament, LIBE Committee.

Goodman-Delahunty, J., & Sporer, S. L. (2010). Unconscious influences in sentencing decisions: A research review of psychological sources of disparity. *Australian Journal of Forensic Sciences, 42*(1), 19–36. https://doi.org/10.1080/00450610903391440

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine, 38(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gravett, W. H. (2017). The myth of objectivity: Implicit racial bias and the law. South African Journal on Human Rights, 33(4), 409–432.

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency* , 90-99. https://doi.org/10.1145/3287560.3287563

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association *Test. Journal of Personality and Social Psychology, 74*(6), 1464–1480.

Grimmelikhuijsen, S., & Meijer, A. (2022). Legitimacy of algorithmic decision-making: Six threats and the need for a calibrated institutional response. *Perspectives on Public Management and Governance, 5(*3), 232-242. https://doi.org/10.1093/ppmgov/gvac008

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Grupo de Expertos de Alto Nivel sobre Inteligencia artificial de la Comisión Europea. (2019). Directrices éticas para una Inteligencia artificial fiable. Comisión Europea.

Güerri Ferrández, C., Martí Barrachina, M., & Pedrosa, A. (2021). Abriendo ventanas virtuales en los muros de la prisión. *IDP: revista de Internet, derecho y política= revista d'Internet, dret i política*, (32), 5. https://doi.org/10.7238/idp.v0i32.375209

Guillén Burguillos, M., & Serrano Robles, E. (2024). Los seres humanos en la toma de decisiones automatizada en el marco del RGPD y la Ley de IA. *Revista CIDOB d'Afers Internacionals*, (138), 145–170.

Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2001). Inside the judicial mind. *Cornell Law Review, 86*(4), 777–830.

Gutterman, A. (2023). Sustainability Reporting and Communications. SSRN. https://doi.org/10.2139/ssrn.4433818

Habermas, J. (1973). *Problemas de legitimación en el capitalismo tardío.* Madrid: Cátedra.

Hannah-Moffat, K. (2013). Actuarial Sentencing: An "Unsettled" Proposition. *Justice Quarterly, 30*(2), 270–296. https://doi.org/10.1080/07418825.2012.682603

Hanson, R. K., y Bussiere, M. T. (1998). Predicting relapse: a meta-analysis of sexual offender recidivism studies. *Journal of consulting and clinical psychology*, 66(2).

Harbers, M., Peeters, M. M. M., & Neerincx, M. A. (2017). Perceived Autonomy of Robots: Effects of Appearance and Context. En Aldinhas Ferreira, M., Silva Sequeira, J., Tokhi, M., Kadar, E., & Virk, G. (Eds.), A World with Robots. Intelligent Systems, Control and Automation: Science and Engineering (pp. 19-33). Springer.

Harcourt, B. E. (2019). *Against prediction: Profiling, policing, and punishing in an actuarial age.* University of Chicago Press.

Hardyns, W., y Rummens, A. (2018). Predictive policing as a new tool for law enforcement? Recent developments and challenges. *European journal on criminal policy and research, 24*(3), 201-218. https://doi.org/10.1007/s10610-017-9361-2

Harris, A. P., & Sen, M. (2019). Bias and judging: A literature review. *Harvard Law Review, 132(*6), 1285–1310.

Hayward, K. J., & Maas, M. M. (2020). Artificial intelligence and crime: A primer for criminologists. *Crime, Media, Culture, 17*(2), 209-233. https://doi.org/10.1177/1741659020917434

He, J., & Zhang, Z. (2025). Algorithm Power and Legal Boundaries: Rights Conflicts and Governance Responses in the Era of Artificial Intelligence. Laws, 14(4), 54.

Hellman, D. (2020). Measuring algorithmic fairness. *Virginia Law Review, 106(*4), 811–866.

High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*. European Commission.

High-Level Expert Group on Artifcial Intelligence. (2019). A defnition of AI: Main capabilities and scientifc disciplines. European Commission. https://ec.europa.eu/newsroom/ dae/document.cfm?doc_id=56341

High-Level Expert Group on Justice & CEPS. (2021). Criminal justice, fundamental rights and the rule of law in the digital age. CEPS. https://cdn.ceps.eu/wp-content/uploads/2021/05/Criminal-Justice-Fundamental-Rights-and-the-Rule-of-law-in-the-Digital-Age.pdf

Hildebrandt, M. (2015). *Smart technologies and the End(s) of Law*. Edwar Elgar Publishing.

Hildebrandt, M. (2020). *Law for Computer Scientists and Other Folk*. Oxford University Press.

Ho, A., Shlosberg, A., & Lesneskie, E. (2018). Sensitivity of error: An examination of the impact of human mistakes on offender risk classification validity. *Justice System Journal*, 39(2), 171–188. https://doi.org/10.1080/0098261X.2018.1430632

Hochstedler Webb, K., Riley, S., & Wells, M. T. (2024). An assessment of racial disparities in pretrial decision-making using misclassification models. SSRN. [Preimpresión].

Holford, W. D. (2022). 'Design-for-responsible' algorithmic decision-making systems: A question of ethical judgement and human meaningful control. *AI and Ethics, 2*(4), 827–836. https://doi.org/10.1007/s43681-022-00144-w

Horowitz, M. C., Kahn, L., Macdonald, J., & Schneider, J. (2024). Adopting AI: How familiarity breeds both trust and contempt. *AI & Society, 39*(5), 1721–1735.

https://doi.org/10.1007/s00146-023-01666-5

Hossain, L., & de Silva, A. (2009). Exploring user acceptance of technology using social networks. *The Journal of High Technology Management Research*, *20*(1), 1-18.

Hueso, A. M., García, A. V. M., & Fincias, P. T. (2022). Revisión sistemática de aceptación de la tecnología digital en personas mayores. Perspectiva de los modelos TAM. *Revista Española de Geriatría y Gerontología, 57*(2), 105-117.

Imandeka, E., Hidayanto, A. N., & Mahmud, M. (2024). *Smart prison technology and challenges: A systematic literature reviews*. *IAES International Journal of Artificial Intelligence (IJ-AI), 13*(2), 1214–1226. https://doi.org/10.11591/ijai.v13.i2.pp1214-1226

Innerarity, D., (2024). Defensa y crítica de la gobernanza algorítmica. Revista CIDOB d'Afers Internacionals, (138), 11-26. https://doi.org/10.24241/rcai.2024.138.3.11

Insko, C. A. (1965). Verbal reinforcement of attitude. *Journal of Personality and Social Psychology*, 2(4), 621–623. https://doi.org/10.1037/h0022489

Isaac, W. (2017). Hope, hype, and fear: The promise and potential pitfalls of AI in criminal justice. In Proceedings of the 1st ACM Conference on Fairness, Accountability and Transparency.

Jain, S., & Murugesan, S. (2021). *Smart connected world: A broader perspective*. En S. Jain & S. Murugesan (Eds.), *Smart connected world: Technologies and applications shaping the future* (pp. 3–23). Springer. https://doi.org/10.1007/978-3-030-76387-9_1

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw inhuman judgment*. Little, Brown Spark.

Kastellec, J. P. (2020). Race, context, and judging on the courts of appeals: Race-based panel effects in death penalty cases. *American Journal of Political Science, 64*(2), 289–305.

Katsh, E., & Rabinovich-Einy, O. (2017). *Digital justice: Technology and the internet of disputes.* Oxford University Press.

Katz, D., & Scotland, C. G. (1959). Function of attitudes: A review of theory and research. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 1, pp. 150–193). Academic Press.

Kehl, D., Guo, P., & Kessler, S. (2017). Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. *Harvard Law School, Berkman Klein Center for Internet & Society.*

Kelly, S., Kaye, S.-A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics, 77,* 101925. https://doi.org/10.1016/j.tele.2022.101925

Kim, S. W., & Lee, Y. (2024). Investigation into the influence of socio-cultural factors on attitudes toward artificial intelligence. *Education and Information Technologies, 29*(8), 9907-9935.

Kim, T., & Peng, W. (2024). Do we want AI judges? The acceptance of AI judges' judicial decision-making on moral foundations. *AI & Society,40. 3683–3696.* https://doi.org/10.1007/s00146-024-02121-9

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics, 133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Knuth, D. E. (1968). *The art of computer programming, vol 1: Fundamental. Algorithms.* Reading, MA: Addison-Wesley.

Kozak, J., & Fel, S. (2024). How sociodemographic factors relate to trust in artificial intelligence among students in Poland and the United Kingdom. *Scientific Reports, 14,* 28776. https://doi.org/10.1038/s41598-024-80305-5

Krištofík, A. (2025). Bias in AI (supported) decision making: Old problems, new technologies. *International Journal for Court Administration, 16*(1), 3–27. https://doi.org/10.36745/ijca.598

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633–705.

Kropp, P. R., Hart, S. D., Webster, C. D., & Eaves, D. (1995). Manual for the Spousal Assault Risk Assessment Guide (SARA). British Columbia, Canada: The British Columbia Institute on Family Violence.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213-236. https://doi.org/10.1002/acp.2350050305

Kuen, L., Westmattelmann, D., Bruckes, M., & Schewe, G. (2023). Who earns trust in online environments? A meta-analysis of trust in technology and trust in provider for technology acceptance. *Electronic Markets, 33*, 61. https://doi.org/10.1007/s12525-023-00672-1

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, 207(4430), 557–558. https://doi.org/10.1126/science.7352271

Lafont, C. (2019). *Democracy without shortcuts: A participatory conception of deliberative democracy.* Oxford Scholarship Online. https://doi.org/10.1093/oso/9780198848189.001.0001

Lahdili, N., Önder, M., & Nyadera, I. N. (2024). Artificial intelligence and citizen participation in governance: Opportunities and threats. *Amme İdaresi Dergisi, 57*(3), 202–229.

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In Proceedings of the conference on fairness, accountability, and transparency (pp. 29-38).

Lazcoz Moratinos, G. (2024). La vigilancia o supervisión humana en el artículo 14 del reglamento de inteligencia artificial : ¿Un mero requisito obligatorio para los sistemas de alto riesgo? In L. Cotino Hueso & P. Simón Castellano (Dirs.), Tratado sobre el reglamento europeo de inteligencia artificial . Aranzadi.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society, 5(1), 1–16. https://doi.org/10.1177/2053951718756684

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction, 3*(182). 1 -26. https://doi.org/10.1145/3359284

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology, 31*(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x

Levy, K., Chasalow, K. E., & Riley, S. (2021). Algorithms and decision-making in the public sector. *Annual Review of Law and Social Science, 17*(1), 309-334.

Ley 18/2011, de 5 de julio, reguladora del uso de las tecnologías de la información y la comunicación en la Administración de Justicia. Boletín Oficial del Estado, 160, 71320-71348. https://www.boe.es/eli/es/l/2011/07/05/18

Ley 3/2020, de 18 de septiembre, de medidas procesales y organizativas para hacer frente al COVID-19 en el ámbito de la Administración de Justicia. Boletín Oficial del Estado, 250. https://www.boe.es/eli/es/l/2020/09/18/3

Ley 42/2015, de 5 de octubre, de reforma de la Ley 1/2000, de 7 de enero, de Enjuiciamiento Civil. Boletín Oficial del Estado, 239, 90240-90288. https://www.boe.es/eli/es/l/2015/10/05/42

Ley Orgánica 1/2025, de 2 de enero, de eficiencia digital del servicio público de Justicia. Boletín Oficial del Estado, 3. https://www.boe.es/eli/es/lo/2025/01/02/1/con

Ley Orgánica 16/1994, de 8 de noviembre, por la que se modifica la Ley Orgánica 6/1985, de 1 de julio, del Poder Judicial. Boletín Oficial del Estado, 268, 34605-34628. https://www.boe.es/eli/es/lo/1994/11/08/16

Liebe, U., Preisendörfer, P., & Enzler, H. B. (2020). The social acceptance of airport expansion scenarios: A factorial survey experiment. *Transportation Research Part D: Transport and Environment*, *84*, 102363.

Liljander, V., Gillberg, F., Gummerus, J., & van Riel, A. (2006). Technology readiness and the evaluation and adoption of self-service technologies. *Journal of Retailing and Consumer Services, 13*(3), 177–191. https://doi.org/10.1016/j.jretconser.2005.08.004

Lin, J. S. C., & Hsieh, P. L. (2007). The influence of technology readiness on satisfaction and behavioral intentions toward self-service technologies. *Computers in Human Behavior, 23*(3), 1597–1615. https://doi.org/10.1016/j.chb.2005.07.006

Lin, Z. J., Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. Science Advances, 6(7).

Lind, E. A. (2001). Fairness heuristic theory: Justice judgments as pivotal cognitions in organizational relations. *Advances in Organizational Justice, 56*(1), 56-88.

Lind, E. A., & Tyler, T. R. (1988). *The social psychology of procedural justice*. Springer.

Long, H. A., French, D. P., & Brooks, J. M. (2020). Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative

evidence synthesis. *Research Methods in Medicine & Health Sciences*, *1*(1), 31-42.

López Lorca, B. (2023). La digitalización de las prisiones y el uso de la inteligencia artificial: Marcadores de última generación para la normalización del entorno penitenciario y la redefinición del proceso de resocialización*. En F. Miró (Coord.),* Digitalización y algoritmización de la justicia*. IDP. Revista de Internet, Derecho y Política*, (39). https://doi.org/10.7238/idp.v0i39.416671

Lowder, E. M., Morrison, M. M., Kroner, D. G., & Desmarais, S. L. (2019). Racial bias and LSI-R assessments in probation sentencing and outcomes. *Criminal Justice and Behavior, 46*(2), 210-233.

Lum, C., Koper, C. S., & Wu, X. (2022). The impact of technology on policing strategy and effectiveness: A systematic review. *Criminology & Public Policy,* 21(3), 456-479.

Lupo, G., & Velicogna, M. (2018). *Making EU justice smart? Looking into the implementation of new technologies to improve the efficiency of cross border justice services delivery. In Smart Technologies for Smart Governments: Transparency, Efficiency and Organizational Issues* (pp. 95-121). Cham: Springer International Publishing.

MacVaugh, J., & Schiavone, F. (2010). Limits to the diffusion of innovation: A literature review and integrative model. *European journal of innovation management, 13*(2), 197-221. https://doi.org/10.1108/14601061011040258

Madan, R., & Ashok, M. (2023). AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly, 40*(1), 101774. https://doi.org/10.1016/j.giq.2022.101774

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm

aversion. *Technological Forecasting and Social Change, 175*, 121390. https://doi.org/10.1016/j.techfore.2021.121390

Malek, M. A. (2022). Criminal courts' artificial intelligence: The way it reinforces bias and discrimination. AI and Ethics, 2, 233–245.

Manning, K. L., Carroll, B. A., & Carp, R. A. (2004). Does age matter? Judicial decision-making in age discrimination cases. *Law & Society Review, 38*(1), 85–102.

Mantelero, A. (2018). AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review, 34*(4), 754-772.

Maravall, J. M., & Przeworski, A. (2003). Democracy and the Rule of Law. Cambridge: Cambridge University Press.

Marchena Gómez, M. (2022). *Inteligencia artificial y jurisdicción penal.* Discurso pronunciado en su ingreso como Académico de Número en la Real Academia de Doctores de España, 26 de octubre de 2022.

Marikyan, D.& Papagiannidis, S. (2024) Technology Acceptance Model: A review. In S. Papagiannidis (Ed), TheoryHub Book. Available at https://open.ncl.ac.uk / ISBN: 9781739604400

Marín, R. (1976). *Psicología social*. Madrid: Alhambra.

Maroney, T. A. (2011). The persistent cultural script of judicial dispassion. California Law Review, 99, 629. https://scholarship.law.vanderbilt.edu/faculty-publications/765

Martín Diz, F. (2024). Justicia híbrida: la tecnología disruptiva al servicio del proceso. Ius et Veritas, 68, 113–129. https://doi.org/10.18800/iusetveritas.202401.008

Martin, N. B. (2019). Algoritmos predictivos al servicio de la justicia: ¿Una nueva forma de minimizar el riesgo y la incertidumbre? Revista da Faculdade Mineira de Direito, 22(43).

Martínez Garay, L., Boix Palop, A., Briz Redón, Á., Flores Giménez, F., García Ortiz, A., Molina Sánchez, M., Montes Suay, F., Palma Ortigosa, A., Peris Manguillot, A., & Soriano Arnanz, A. (2024). *Three predictive policing approaches in Spain: VioGén, RisCanvi and VeriPol. Assessment from a human rights perspective*. Universitat de València. https://doi.org/10.7203/PUV-OA-451-9

Martínez i Coma, F., & Sanz-Labrador, I. (2009). ¿Qué determinan las opiniones sobre la justicia? Un estudio cuantitativo. Revista Española de Ciencia Política, (21), 69–90. https://dialnet.unirioja.es/servlet/articulo?codigo=3069821

Mayoral Díaz-Asensio, J. A., & Martínez i Coma, F. (2013). La calidad de la justicia en España: ¿Cómo evalúan los españoles el funcionamiento de las instituciones judiciales y qué se puede hacer para mejorarlas?. Fundación Alternativas.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Dartmouth College.

McKay, C. (2019). Predicting risk in criminal procedure: Actuarial tools, algorithms, AI and judicial decision-making. Legal Studies Research Paper Series, The University of Sydney Law School, 19/67, 1–31.

Meijer, A., & Wessels, M. (2019). Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration*, *42*(12), 1031-1039. https://doi.org/10.1080/01900692.2019.1575664

Miller, A. L. (2018). Expertise fails to attenuate gendered biases in judicial decision-making. Law and Human Behavior, 42(5), 472–487.

Minaggia, M. (2023). La inteligencia artificial en el derecho penal. Ministerio Público de la Defensa. https://repositorio.mpd.gov.ar/handle/123456789/4810

Ministerio de Justicia. (2020). Plan Justicia 2030: Una justicia accesible, eficiente y sostenible. Gobierno de España. Recuperado de: https://www.justicia2030.es/

Miró Llinares, F. (2018). Inteligencia artificial y justicia penal: más allá de los resultados lesivos causados por robots. Revista de Derecho Penal y Criminología, (20), 87-130. https://revistas.uned.es/index.php/RDPC/article/view/26446

Miró Llinares, F. (2020). El sistema penal ante la inteligencia artificial: actitudes, usos, retos. En Kiefer, M. y Dupuy, D. (Dirs.)., *Cibercrimen III. Inteligencia artificial . Automatización, algoritmos y predicciones en el Derecho penal y procesal penal*. Argentina: BdeF.

Miró Llinares, F. (2022). Policía predictiva: Realismo frente a utopías y distopías. In F. J. Castro Toledo (Dir.), *La transformación algorítmica del sistema de justicia penal* (pp. 177–198). Aranzadi.

Miró Llinares, F. (2025). Derecho penal y desafíos de la inteligencia artificial (en el contexto del nuevo marco regulatorio europeo). En M. E. Casas Baamonde (Dir.) & D. Pérez del Prado (Coord.), *Derecho y tecnologías* (pp. 177–212). Editorial Centro de Estudios Ramón Areces. ISBN: 978-84-9961-466-3

Miró Llinares, F., & Santisteban Galarza, A. (2024). Policía predictiva y evaluación del riesgo criminal en el marco del IA Act. En L. Cotino Hueso & M. Simón Castellano (Eds.), Tratado sobre el Reglamento de Inteligencia Artificial (pp. 319, 333). Tirant lo Blanch.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Mitchell, T. L., Haw, R. M., Pfeifer, J. E., & Meissner, C. A. (2005). Racial bias in mock juror decision-making: A meta-analytic review of defendant treatment. *Law and Human Behavior, 29*(6), 621–637.

Mitchell, T. M. (1997). *Machine learning.* McGraw-Hill.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature machine*

*intelligence, 1(*11), 501-507. https://doi.org/10.1038/s42256-019-0114-4

Mittelstadt, B. D., Russell, C., & Wachter, S. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLOS Medicine, 6(7).

Momani, A. M. (2020). The unified theory of acceptance and use of technology: A new approach in technology acceptance. International Journal of Sociotechnology and Knowledge Development (IJSKD), 12(3), 79-98.

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology, 12*, 489–513. https://doi.org/10.1146/annurev-clinpsy-021815-092945

Morales, F. J., Moya, M., Gaviria, E., & Cuadrado, I. (2007). *Psicología Social* (3ª ed.). Mc Graw-Hill.

Morales, J. F. (1999). *Psicología social*. Madrid: McGraw-Hill.

Morris, C. G. (1996). *Psicología: Una introducción*. México: Pearson Educación.

Moses, L. B., & Chan, J. (2018). Algorithmic prediction in policing: Implications for fairness and efficiency. AI & Society, 33(1), 1-10.

Mosier, K. L., & Skitka, L. J. (1999, September). Automation use and automation bias. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 43, No. 3, pp. 344-348). Sage CA: Los Angeles, CA: SAGE Publications.

Muñoz Aranguren, A. (2011). La influencia de los sesgos cognitivos en las decisiones jurisdiccionales: el factor humano. Una aproximación. *InDret, 2*.

Nadler, J., & McDonnell, M.-H. (2012). Moral character, motive, and the psychology of blame. Cornell Law Review, 97, 255.

Natarajan, S., Mathur, S., Sidheekh, S., Stammer, W., & Kersting, K. (2025). Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?. Proceedings of the AAAI Conference on Artificial Intelligence, 39(27), 28594-28600. https://doi.org/10.1609/aaai.v39i27.35083

Nazareno, D. O. D. L. (2023). Digitization, digitalización y transformación digital: conceptos clave para la práctica empresarial. Serie Científica De La *Universidad De Las Ciencias Informáticas, 16*(10), 44-68.

Nicolás-Sánchez, A., & Castro-Toledo, F. J. (2024). Uncovering the social impact of digital steganalysis tools applied to cybercrime investigations: A European Union perspective. *Crime Science, 1*3(11).

Nieto, A. (2004). El desgobierno judicial. Madrid: Editorial Trotta.

Nilsson, N. J. (2009). *The quest for artificial intelligence: A history of ideas and achievements*. Cambridge University Press.

Ning, H., Liu, H., Ma, J., Yang, L. T., Wan, Y., Ye, X., & Huang, R. (2015). From Internet to smart world. *IEEE Access*, *3*, 1994-1999.

Noble, S. U. (2018). *Algorithms of oppression. In Algorithms of Oppression*. New York University Press.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown Publishing.

OECD - Organisation for Economic Co-operation and Development. (2024). *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions.* OECD Publishing.

OECD (2024). Recomendación del Consejo sobre la inteligencia artificial (OECD/LEGAL/0449, versión modificada 03.05.2024).

OECD. (2019). *OECD Principles on Artificial Intelligence*. OECD Publishing.

OECD. (2019). *Recommendation of the Council on Artificial Intelligence*. OECD Legal

Instruments. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

OECD. (2020). *Framework for the Classification of AI Systems*. OECD Publishing.

Oren-Kolbinger, O. (2019). Measuring the effect of social background on judicial decision-making. *Journal of Law & Society, 46*(3), 478–502.

Ortega Giménez, J. (2024). Implicaciones procesales del incumplimiento del Reglamento de IA. En L. Cotino Hueso & M. Simón Castellano (Eds.), Tratado sobre el Reglamento de Inteligencia Artificial (p. 133). Tirant lo Blanch.

Oswald, M. (2020). Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretion. *Philosophical Transactions of the Royal Society A, 378*(2166), 20190359.

Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2020). Algorithmic risk assessment policing models: Lessons from the Durham HART model and 'experimental' proportionality. *Information & Communications Technology Law, 29*(2), 1–30. https://doi.org/10.1080/13600834.2018.1458455

Owen, R., Macnaghten, P., & Stilgoe, J. (2020). Responsible research and innovation: From science in society to science for society, with society. In Emerging technologies (pp. 117-126). Routledge.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Bouton, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). Declaración PRISMA 2020: Una guía actualizada para la publicación de revisiones sistemáticas. *Revista Española de Cardiología, 74*, 790–799.

Parasuraman, A. (2000). Technology readiness index (TRI): A multiple-item scale to measure readiness to embrace new technologies. Journal of Service Research, 2(4), 307–320. https://doi.org/10.1177/109467050024001

Parasuraman, A., & Colby, C. L. (2001). *Techno-ready marketing: How and why your customers adopt technology.* Free Press.

Parasuraman, A., & Colby, C. L. (2015). An updated and streamlined Technology Readiness Index: TRI 2.0. Journal of Service Research, 18(1), 59–74. https://doi.org/10.1177/1094670514539730

Parlamento Europeo & 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 (Reglamento de inteligencia artificial)*. Diario Oficial de la Unión Europea, 168.

Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

Pasquale, F. (2020). New laws of robotics: Defending human expertise in the age of AI. Harvard University Press.

Peeters, R., & Schuilenburg, M. (2018). Machine justice: Governing security through the bureaucracy of algorithms. Information Polity, 23(3), 267-280.

Pereira Iubel de Oliveira, D., Pinheiro de Lima, E., Gonçalves Machado, C., Gouvêa da Costa, S.E. (2020). A Review Content Analysis Between Industry 4.0 and Sustainable Manufacturing. In: Anisic, Z., Lalic, B., Gracanin, D. (eds) Proceedings on 25th International Joint Conference on Industrial Engineering and Operations Management. Lecture Notes on Multidisciplinary Industrial Engineering. Springer, Cham. https://doi.org/10.1007/978-3-030-43616-2_2

Peresie, J. L. (2005). Female judges' matter: Gender and collegial decision-making in the federal appellate courts. Yale Law Journal, 114(7), 1759–1790.

Pérez Domínguez, S., & Simón Castellano, P. (2023). Attitudes and perceptions regarding algorithmic judicial judgement: Barriers to innovation in the judicial system? In F. Miró (Coord.), Digitization and algorithmization of justice. *IDP. Revista de Internet, Derecho y Política,* 39. http://dx.doi.org/10.7238/idp.v0i39.417206

Pérez Domínguez, S., Castro-Toledo, F. J., & Miró-Llinares, F. (2019). Prevalencia, factores asociados y diferencias de género en el cumplimiento de la propiedad intelectual: Una revisión sistemática. *Revista Electrónica de Criminología*, 02–04.

Poder Judicial de España. (2024). Evaluación de la justicia en España por la CEPEJ. Consejo General del Poder Judicial. https://www.poderjudicial.es/cgpj/es/Temas/Estadistica-Judicial/Estadistica-por-temas/Aspectos-internacionales/Informes-Organismos-Extranjeros/CEPEJ--Comision-Europea-para-eficiencia-de-la-justicia-

Portela, M., Castillo, C., Tolan, S., Karimi-Haghighi, M., & Andres Pueyo, A. (2025). A comparative user study of human predictions in algorithm-supported recidivism risk assessment. *Artificial Intelligence and Law, 33*(3), 471–517. https://doi.org/10.1007/s10506-024-09393-y

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*(1), 109-130. https://doi.org/10.1093/poq/nfh008

Rachlinski, J. J., & Wistrich, A. J. (2017). Judging the judiciary by the numbers: Empirical research on judges. *Annual Review of Law and Social Science, 13*(1), 203-229. https://doi.org/10.1146/annurev-lawsocsci-110615-085032

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. *Nature, 568*(7753), 477-486. https://doi.org/10.1038/s41586-019-1138-y

Ratcliffe, J. (2019). Predictive policing. En Weisburd, D & Braga, A.A. (Eds.), P*olice innovation. Contrasting perspectives. 2d edition.* Cambridge: Cambridge University Press.

Real Castrillo, C. (2021). *La gobernanza de la ciberseguridad en España: un estudio empírico de los actores, redes de colaboración y prospectiva desde las teorías de la seguridad plural* (Tesis doctoral, Universidad de Cádiz).

Real Decreto 1065/2015, de 27 de noviembre, sobre comunicaciones electrónicas en la Administración de Justicia en el ámbito territorial del Ministerio de Justicia y por el que se regula el sistema LexNET. Boletín Oficial del Estado, 312, 113314-113331. https://www.boe.es/eli/es/rd/2015/11/27/1065

Recupero, P. R., Christopher, P. P., Stong, D. R., Price, M., & Harms, S. E. (2015). Gender bias in judicial decisions of undue influence in testamentary challenges. *Journal of the American Academy of Psychiatry and the Law, 43*(2), 60-68.

Reglamento (UE) 2019/2144 del Parlamento Europeo y del Consejo, de 27 de noviembre de 2019, relativo a los requisitos de homologación de tipo de los vehículos de motor.

Reiling, A. D. (2020). Courts and Artificial Intelligence. *International Journal for Court Administration*, *11*(2). https://doi.org/10.36745/ijca.343

Reis, J., Amorim, M., Melão, N., Cohen, Y., & Rodrigues, M. (2020). Digitalization: A literature review and research agenda. In *International Joint conference on industrial engineering and operations management* (pp. 443-456). Springer, Cham. https://doi.org/10.1007/978-3-030-43616-2_47

Rekunenko, I., Koldovskyi, A., Hordiienko, V., Yurynets, O., Abu Khalaf, B., & Ktit, M. (2025). Technology adoption in government management: Public sector transformation analysis. *Journal of Governance & Regulation, 14*(1), 150–160. https://doi.org/10.22495/jgrv14i1art14

Reyes, R., & Reyes, J. W. (2019). Religion in judicial decision-making: An empirical analysis. BYU Law Review, 2019(1), 293–336.

Rich, E. (1983). *Artificial Intelligence*. McGraw-Hill.

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review, 94*(1), 192–233.

Rizer, A., y Watney, C. (2018). Artificial Intelligence Can Make Our Jail System More Efficient, Equitable and Just, 2018. Disponible en Internet en: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3129576

Rogers, E. M. (2003). Diffusion of innovations (5th ed.). Free Press.

Rogers, E. M., Singhal, A., & Quinlan, M. M. (2014). *Diffusion of innovations*. En D. K. Kim & J. W. Dearing (Eds.), An integrated approach to communication theory and research (pp. 432–448). Routledge.

Romain Dagenhardt, D. M. (2021). "You know baseball? 3 strikes": Understanding racial disparity with mixed methods for probation review hearings. *Social Sciences, 10*(6). https://doi.org/10.3390/socsci10060235

Rosenberg, M. J. (1965). When dissonance fails: On eliminating evaluation apprehension from attitude measurement. Journal of Personality and Social Psychology, 1, 28–42.

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Prentice Hall.

Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Saavedra-Vera, C. O., Jáuregui Bustamante, K. del R., & Arista Bustamante, L. L. (2023). La incidencia del sesgo algorítmico en la justicia predictiva del sistema judicial. *TZHOECOEN, 15*(2), 79–97. https://doi.org/10.26495/tzh.v15i2.2592

Sahin, I. (2006). Detailed review of Rogers' diffusion of innovations theory and educational technology-related studies based on Rogers' theory. Turkish Online Journal of Educational Technology (TOJET), 5(2), 14–23.

Salanova, M., Llorens, S., & Cifre, E. (2013). The dark side of technologies: technostress among users of information and communication technologies.

International journal of psychology: *Journal international de psychologie, 48*(3), 422–436. https://doi.org/10.1080/00207594.2012.680460

Sartor, G. (2020). Artificial intelligence and human rights: Between law and ethics. Maastricht Journal of European and Comparative Law, 27(6), 705-719. https://doi.org/10.1177/1023263X20981566

Scaria, A. G., Subramanian, V., George, N. K., & Sengupta, N. (2024). Algorithms and recidivism: A multi-disciplinary systematic review. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7 (pp. 1292–1305).

Schiavo, G., Businaro, S., & Zancanaro, M. (2024). Comprehension, apprehension, and acceptance: Understanding the influence of literacy and anxiety on acceptance of artificial intelligence. *Technology in Society, 77,* 102537. https://doi.org/10.1016/j.techsoc.2024.102537

Schwab, K. (2016). *La cuarta revolución industrial.* Debate

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–457. https://doi.org/10.1017/S0140525X00005756

Segura, R. E. (2023). Inteligencia artificial y administración de justicia: Desafíos derivados del contexto latinoamericano. *Revista de Bioética y Derecho*, 58, 45–72. https://doi.org/10.1344/rbd2023.58.40601

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59-68). https://doi.org/10.1145/3287560.3287598

Sempere, A. (2024). El régimen sancionador del Reglamento de IA y sus implicaciones penales. En L. Cotino Hueso & M. Simón Castellano (Eds.), *Tratado sobre el Reglamento de Inteligencia Artificia*l (pp. 875-877). Tirant lo Blanch.

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: definition and

background. In Mission AI: The new system technology. *Research for Policy. Springer, Cham.* https://doi.org/10.1007/978-3-031-21448-6_2

Shen, H., Cabrera, Á. A., Perer, A., & Hong, J. (2022). "Public(s)-in-the-Loop": Facilitating deliberation of algorithmic decisions in contentious public policy domains. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–7). https://doi.org/10.48550/arXiv.2204.10814

Simón Castellano, P. (2021). *Justicia cautelar e inteligencia artificial : La alternativa a los atávicos heurísticos judiciales.* J. M. Bosch. https://doi.org/10.2307/j.ctv1tqcxbh

Simón Castellano, P. (2023). *La evaluación de impacto algorítmico en los derechos fundamentales.* Aranzadi.

Sisk, G. C., & Heise, M. (2005). Judges and ideology: Public and academic debates about statistical measures. Northwestern University Law Review, 99(2), 743–791.

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, & recidivism: Predictive bias and disparate impact. *Criminology, 54*(4), 680–712.

Skitka, L. J., Mosier, K. L., & Burdick, M. (2012). Automation bias and decision-making: Psychological mechanisms and policy implications. Journal of Applied Psychology, 87(5), 999.

Slobogin, C. (2021). *Just algorithms: Using science to reduce incarceration and inform a jurisprudence of risk. Cambridge* University Press.

Slobogin, C. (2021). Just algorithms: Using science to reduce incarceration and inform a jurisprudence of risk. Cambridge University Press.

Socol de la Osa, D. U., & Remolina, N. (2024). Artificial intelligence at the bench: Legal and ethical challenges of informing -or misinforming- judicial decision-making through generative AI. *Data & Policy*, *6*, e59. . https://doi.org/10.1017/dap.2024.53

Soriano Arnanz, A. (2023). Creando sistemas de inteligencia artificial no discriminatorios: Buscando el equilibrio entre la granularidad del código y la generalidad de las normas jurídicas. *IDP. Revista de Internet, Derecho y Política, 38*, 1–12. https://doi.org/10.7238/idp.v0i38.403794

Sourdin, T. (2018). Judge v Robot?: Artificial intelligence and judicial decision-making. *University of New South Wales Law Journal, The*, *41*(4), 1114-1133.

Spohn, C., & Sample, L. L. (2013). The dangerous drug offender in federal court: Intersections of race, ethnicity, and culpability. Crime & Delinquency, 59(1), 3–31.

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology, 55*(1), 167-194. https://doi.org/10.1111/j.1744-6570.2002.tb00108.x

Starke, C., & Lünich, M. (2020). Artificial intelligence for political decision-making in the European Union: Effects on citizens' perceptions of input, throughput, and output legitimacy. *Data & Policy, 2*, e16. https://doi.org/10.1017/dap.2020.19

Starke, C., Baleis, J., Keller, B., & De Fine Licht, J. (2022). Artificial intelligence in government: A systematic literature review on applications, impacts, and trust. Government Information Quarterly, 39(3), 101704. https://doi.org/10.1016/j.giq.2022.101704

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review, 66*(4), 803–872.

Steffensmeier, D., & Demuth, S. (2000). Ethnicity and sentencing outcomes in U.S. federal courts: Who is punished more harshly? *American Sociological Review, 65*(5).

Steffensmeier, D., Ulmer, J., & Kramer, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology, 36*(4), 763–798. https://doi.org/10.1111/j.1745-

9125.1998.tb01265.x

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2495-2507. https://doi.org/10.1145/3630106.3659051

Sunstein, C. R. (2018). *#Republic: Divided democracy in the age of social media.* Princeton: Princeton University Press.

Susskind, R. (2019). Online courts and the future of justice. Oxford University Press. https://doi.org/10.1093/oso/9780198838364.001.0001

Tam, C., & Oliveira, T. (2017). Literature review of mobile banking and individual performance. *International Journal of Bank Marketing*, 35(7), 1042–1065. https://doi.org/10.1108/IJBM-09-2015-0143

Tarafdar, M., Cooper, C. L., & Stich, J. F. (2019). The technostress trifecta-techno eustress, techno distress and design: Theoretical directions and an agenda for research. *Information systems journal, 29*(1), 6-42. https://doi.org/10.1111/isj.12169

Tashakkori, A., & Creswell, J. W. (2007). The new era of mixed methods. *Journal of mixed methods research, 1*(1), 3-7. https://doi.org/10.1177/2345678906293042

Ter-Minassian, L. (2025). *Democratizing AI Governance: Balancing Expertise and Public Participation.* arXiv preprint arXiv:2502.08651.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*(4), 529–554. https://doi.org/10.1086/214483

Toharia, J. J. (1975). El juez español: un análisis sociológico. Madrid: Tecnos.

Torres Albero, C., Robles, J. M., De Marco, S., & Antino, M. (2017). Revisión analítica del modelo de aceptación de la tecnología: el cambio tecnológico. *Papers*

(Universitat Autònoma de Barcelona), *102*(1), 0005-27. https://doi.org/10.5565/rev/papers.2233

Treasury Board of Canada Secretariat. (2019). *Directive on Automated Decision-Making*. Government of Canada.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*(1), 230–265. https://doi.org/10.1112/plms/s2-42.1.230

Tversky, A., & Kahneman, D. (1974). *Judgment under uncertainty: Heuristics and biases*. Science, 185.

Tyler, T. R. (2006). *Why people obey the law* (2nd ed.). Princeton University Press.

Tyler, T. R., & Huo, Y. J. (2002). Trust in the law: Encouraging public cooperation with the police and courts. Russell Sage Foundation.

Ucín, M. C. (2022). ¿Jueces sensibles? Una introducción al análisis del rol de las emociones en la decisión judicial. Doxa. Cuadernos de Filosofía del Derecho, 45, 191–219. https://doi.org/10.14198/DOXA2022.45.07

UNESCO (Ed.) (2021). Recomendación sobre la ética de la Inteligencia Artificial. UNESCO.

Urquidi Martin, A. C., Calabor Prieto, M. S., & Tamarit Aznar, C. (2019). Entornos virtuales de aprendizaje: modelo ampliado de aceptación de la tecnología. *Revista electrónica de investigación educativa*, *21*. https://doi.org/10.24320/redie.2019.21.e22.1866

Van De Poel, I. (2020). Three philosophical perspectives on the relation between technology and society, and how they affect the current debate about artificial intelligence. *Human Affairs,* *30*(4), 499-511. https://doi.org/10.1515/humaff-2020-0042

Van den Bos, K. (2001). Uncertainty management: The influence of uncertainty salience on reactions to perceived procedural fairness. *Journal of Personality*

and Social Psychology, 80(6), 931-941. https://doi.org/10.1037/0022-3514.80.6.931

van den Bos, K., Wilke, H. A. M., Lind, E. A., & Vermunt, R. (1998). Evaluating outcomes by means of the fair process effect: Evidence for different processes in fairness and satisfaction judgments. *Journal of Personality and Social Psychology, 74*(6), 1493–1503. https://doi.org/10.1037/0022-3514.74.6.1493

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–538.

Varela, L. A. Y. (2004). Modelo de aceptación tecnológica (TAM) para determinar los efectos de las dimensiones de cultura nacional en la aceptación de las TIC. *Revista Internacional de Ciencias Sociales y Humanidades, SOCIOTAM*, *14*(1), 131-171.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4(*2), 1-17. https://doi.org/10.1177/2053951717743530

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425–478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly*, 36(1), 157–178. https://doi.org/10.2307/41410412

Venkatesh, V., Thong, J. Y., & Xu, X. (2016). Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the association for Information Systems, 17*(5), 328-376.

Verplanck, W. S. (1955). The operant conditioning of human motor behavior. *Psychological Bulletin, 53*(1), 70–83. https://doi.org/10.1037/h0041472

Viana, R. A., & Arranz, J. M. (2021). Efficiency of university education: a partial frontier analysis. *Centro de Investigación y Docencia Económicas (CIDE).* https://repositorio-digital.cide.edu/handle/11651/4680

Vicente, L. y Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports* , *13* (1), 15737.

Wachter-Boettcher, S. (2017). *Technically wrong: Sexist apps, biased algorithms, and other threats of toxic tech*. W. W. Norton & Company.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76-99.

Walczuch, R., Lemmink, J., & Streukens, S. (2007). The effect of service employees' technology readiness on technology acceptance. Information & Management, 44(2), 206–215. https://doi.org/10.1016/j.im.2006.12.005

Waldron, R. J., Quarles, C. L., McElreath, D., Waldron, M. E., & Milstein, D. (2009). *The criminal justice system: An introduction*. CRC Press.

Wang, W., & Siau, K. (2019). Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. Journal of Database Management, 30(1). https://doi.org/10.4018/JDM.2019010104

Wang, X., Zhang, Y., & Zhu, R. (2022). A brief review on algorithmic fairness. *Management System Engineering, 1(*7), 1–13. https://doi.org/10.1007/s44176-022-00006-z

Weinstein, S. (2022). *Lawyers' perceptions on the use of AI.* In: Custers, B., Fosch-Villaronga, E. (eds) Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice. Information Technology and Law Series, 35. T.M.C. Asser Press, The Hague. https://doi.org/10.1007/978-94-6265-523-2_21

Weisburd, D., y Eck, J. E. (2004). What can police do to reduce crime, disorder, and fear?. *The Annals of the American Academy of Political and Social Science*, 593(1), 42-65. https://doi.org/10.1177/0002716203262548

Westerman, G., Bonnet, D., & McAfee, A. (2014). *Leading digital: Turning technology into business transformation.* Harvard Business Press.

Westerman, G., Bonnet, D., & McAfee, A. (2014). The nine elements of digital transformation. *MIT Sloan Management Review, 55*(3), 1-6.

Williams, M. D., Rana, N. P., & Dwivedi, Y. K. (2015). The unified theory of acceptance and use of technology (UTAUT): A literature review. *Journal of Enterprise Information Management*, 28(3), 443–488. https://doi.org/10.1108/JEIM-09-2014-0088

Williams, M. S., & Law, A. O. (2012). Understanding judicial decision making in immigration cases at the US Courts of Appeals. *Justice System Journal, 33(*1), 97-120.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2018). Artificial Intelligence and the Public Sector—Applications and Challenges. International Journal of Public Administration, 42(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Wischmeyer, T., & Rademacher, T. (Eds.). (2020). Regulating artificial intelligence 1, (1). Cham: Springer.

Wisser, L. (2019). Pandora's algorithmic black box: The challenges of using algorithmic risk assessments in sentencing. *American Criminal Law Review*, 56, 1811- 1832.

Wofford, C. B. (2017). Avoiding adversariness? The effects of gender on litigant decision-making. *Politics & Gender, 13*(4), 656–682. https://doi.org/10.1017/S1743923X17000071

Wright, R. W., Brand, R. A., Dunn, W., & Spindler, K. P. (2007). How to write a

systematic review. Clinical Orthopaedics and Related Research, 455, 23-29. https://doi.org/10.1097/BLO.0b013e31802c9098

Xu, N., & Wang, K. J. (2021). Adopting robot lawyer? The extending artificial intelligence robot lawyer technology acceptance model for legal industry by an exploratory study. *Journal of Management & Organization, 27*(5), 867–885.

Yalcin, G., Themeli, E., Stamhuis, E., Philipsen, S., & Puntoni, S. (2023). Perceptions of justice by algorithms. *Artificial intelligence and Law, 31*(2), 269-292. https://doi.org/10.1007/s10506-022-09312-z

Yeung, K. (2019). Regulation by blockchain: The emerging battle for supremacy between the code of law and code as law. *Modern Law Review*, 82(2), 207-239. https://doi.org/10.1111/1468-2230.12399

Yigitcanlar, T., Mehmood, R. & Corchado, J.M. (2021). Green Artificial Intelligence: Towards an Efficient, Sustainable and Equitable Technology for Smart Cities and Futures. *Sustainability*, 13(6), 8952. https://doi.org/10.3390/su13168952

Yong Varela, L. A., Rivas Tovar, L. A., & Chaparro, J. (2010). Modelo de aceptación tecnológica (TAM): un estudio de la influencia de la cultura nacional y del perfil del usuario en el uso de las TIC. *Innovar*, *20*(36), 187-203.

Zalnieriute, M., Moses, L. B., & Williams, G. (2019). The rule of law and automation of government decision-making. *The Modern Law Review*, *82*(3), 425-455. https://doi.org/10.1111/1468-2230.12412

Zanón, A. G. (1990). La Técnica del Grupo Nominal. *Documentación administrativa*, (223), 51-98.

Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology, & Human Values, 41*(1), 118-132. https://doi.org/10.1177/0162243915605575

Završnik, A. (2020). Criminal justice, artificial intelligence systems and human rights. *ERA Forum, 20*, 567–583. https://doi.org/10.1007/s12027-020-00602-0

Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology, 18*(5), 623–642. https://doi.org/10.1177/1477370819876762

Zedner, L. (2007). Pre-crime and post-criminology?. *Theoretical Criminology, 11*(2), 261-281.https://doi.org/10.1177/1362480607075851

Zeng, J., Ustun, B., & Rudin, C. (2016). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A, 180*(3), 689–722. https://doi.org/10.1111/rssa.12227

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? Philosophy & Technology, 32(4), 661-683.

Zouridis, S., Van Eck, M., & Bovens, M. (2020). Automated discretion. In: Evans, T., Hupe, P. (eds) Discretion and the Quest for Controlled Freedom. Palgrave Macmillan, Cham, 313-329. https://doi.org/10.1007/978-3-030-19566-3_20

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

ANEXOS

Anexo 1. Evaluación de la calidad de los estudios incluidos en la revisión sistemática del cap 4.

Tabla 32.

Resultados de la evaluación de la calidad de los estudios incluidos en la revisión.

	ESTUDIOS																															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**
Definición clara de los objetivos	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Congruencia metodológica.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Adecuación método-objetivos.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Adecuación de la técnica de recogida de datos con la pregunta de investigación y el método.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Aspectos éticos	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Análisis de datos riguroso.	✓	✓	✓	✓	✓	✓	±	✓	✓	✓	±	✓	✓	✓	✓	✓	±	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Aplicabilidad de resultados	✓	✓	✓	✓	✓	±	±	±	±	±	±	✓	✓	±	✓	✓	±	✓	✓	✓	±	✓	✓	±	✓	✓	✓	±	✓	✓	±	±
Total (máx. 7):	7	7	7	7	7	6.5	6	6.5	6.5	6.5	6.5	7	7	6.5	7	7	6	7	7	7	6.5	7	7	6.5	7	7	7	6.5	7	7	6.5	6.5

Anexo 2. Instrumento de los estudios empíricos incluidos en el capítulo 6.

Impacto del uso de algoritmos predictivos para la investigación y tratamiento jurídico de la delincuencia sobre la legitimidad percibida por la ciudadanía y la aceptación social.

El uso de algoritmos en el sistema judicial ha generado un intenso debate. Mientras algunos defienden que la automatización de ciertos procesos judiciales puede ser beneficiosa, otros temen que los algoritmos puedan generar resultados sesgados y agravar las desigualdades. Siendo así, nos interesa conocer su opinión respecto a la inclusión de estas herramientas en el ámbito judicial.

Solicito su participación en el proyecto de investigación titulado IusMachina: sobre las bases normativas y el impacto real de la utilización de algoritmos predictivos en los ámbitos judicial y penitenciario, cuyo/a investigador/a principal es Fernando Miró Llinares. Consiste en el análisis de la aceptación del uso de herramientas automatizadas en el ámbito judicial y penitenciario. Se podrá cumplimentar desde 1/02/2024 hasta 31/04/2024. La participación es totalmente voluntaria (si no desea participar o si se retira anticipadamente no habrá ninguna consecuencia) y anónima (no se dispondrá de ningún dato que le identifique). Si tiene alguna pregunta puede consultar en este correo: s.perezd@crimina.es. Si usted responde se entiende de forma tácita que ha comprendido en que consiste este estudio, que ha podido preguntar y aclarar las dudas que se le hubieran planteado y que acepta participar. El equipo investigador le agradece su valioso tiempo

No olvides que se trata de una encuesta TOTALMENTE ANÓNIMA que no te ocupará más de 15 MINUTOS y que nos aportará datos de gran interés.

- He leído y acepto el tratamiento de mis datos personales para la investigación académica que se ha descrito

BLOQUE 1.- Datos sociodemográficos

P1 - ¿Cuál es su sexo?

- Hombre
- Mujer

P2 - ¿Cuál es su edad?

_____

P3 - ¿Cuál es su nivel de estudios?

- Primaria
- Secundaria
- Bachillerato/FP
- Universitarios

P4 - ¿Tiene formación en alguno de los siguientes ámbitos?:

- Licenciatura o Grado en Derecho
- Ejercicio práctico de la abogacía
- Magistrado/a o juez/a
- Consultor/a o asesor/a sector legal
- He trabajado o trabajo en empresas legal tech
- Técnico/a en análisis de datos
- He participado o participo en el desarrollo de soluciones tecnológicas
- Conocimiento práctico de sistemas que apliquen IA
- Ninguno de los anteriores

P5 - ¿Estás familiarizado/a con el concepto de algoritmos predictivos utilizados en el ámbito jurídico?

- Sí
- No

BLOQUE 2.- Casos para el estudio experimental

GRUPO 1. TOTAL + DISEÑO HUMANO + PROPORCIOANDO

Se ha instaurado en el sistema penitenciario español un algoritmo predictivo diseñado por expertos en psicología, criminología y derecho para evaluar el riesgo de reincidencia de un recluso que solicita la libertad condicional. El algoritmo considera factores como el historial delictivo, comportamiento en la prisión y participación en programas de rehabilitación para determinar la probabilidad de reincidencia.

Iván Pérez fue condenado a cumplir una pena de 7 años de prisión por delitos relacionados con robo agravado. Durante su tiempo en prisión, ha mostrado un comportamiento relativamente bueno, participando activamente en programas de rehabilitación, como cursos de educación, talleres de habilidades laborales y terapia de control de impulsos pero ha cometido 3 infracciones disciplinarias durante su tiempo en prisión. El algoritmo predictivo concluye que tiene una baja probabilidad de reincidencia; decide conceder la libertad condicional.

¿Estás de acuerdo con la decisión tomada?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

Se ha instaurado en el sistema de justicia español un algoritmo predictivo diseñado por expertos en psicología, criminología y derecho para determinar sentencias penales en casos de delitos no violentos. El sistema utiliza algoritmos que consideran la gravedad del delito, antecedentes penales del acusado y factores socioeconómicos para determinar la sentencia.

Carlos García, es acusado de un robo a mano armada en una joyería. Según la acusación, Carlos, junto con otro cómplice, ingresó a la joyería durante el horario laboral y amenazó al personal y a los clientes con un arma de fuego, llevándose joyas

valuadas en 50.000€. Este delito se clasifica como robo con violencia y se considera un delito grave en la ley penal. El algoritmo predictivo decide imponer una sanción de 4 años de prisión.

¿Estás de acuerdo con la decisión tomada?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

¿En qué medida le parece aceptable que se usen las herramientas anteriores en el sistema judicial y penitenciario?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

GRUPO 2. TOTAL + DISEÑO HUMANO + DESPROPORCIONADO

Se ha instaurado en el sistema penitenciario español un algoritmo predictivo diseñado por expertos en psicología, criminología y derecho para evaluar el riesgo de reincidencia de un recluso que solicita la libertad condicional. El algoritmo considera factores como el historial delictivo, comportamiento en la prisión y participación en programas de rehabilitación para determinar la probabilidad de reincidencia.

Iván Pérez fue condenado a cumplir una pena de 7 años de prisión por delitos relacionados con robo agravado. Durante su tiempo en prisión, ha mostrado un comportamiento relativamente bueno, participando activamente en programas de rehabilitación, como cursos de educación, talleres de habilidades laborales y terapia de control de impulsos pero ha cometido 3 infracciones disciplinarias durante su tiempo en prisión. El algoritmo predictivo concluye que tiene una baja probabilidad de reincidencia; decide rechazar la libertad condicional.

¿Está de acuerdo con la decisión tomada?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

Se ha instaurado en el sistema de justicia español un algoritmo predictivo diseñado por expertos en psicología, criminología y derecho para determinar sentencias penales en casos de delitos no violentos. El sistema utiliza algoritmos que consideran la gravedad del delito, antecedentes penales del acusado y factores socioeconómicos para determinar la sentencia.

Carlos García, es acusado de un robo a mano armada en una joyería. Según la acusación, Carlos, junto con otro cómplice, ingresó a la joyería durante el horario laboral y amenazó al personal y a los clientes con un arma de fuego, llevándose joyas valuadas en una gran cantidad de dinero. Este delito se clasifica como robo con violencia y se considera un delito grave en la ley penal. el algoritmo predictivo decide imponer una multa de 10.000€ sin prisión.

¿Está de acuerdo con la decisión tomada?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

¿En qué medida le parece aceptable que se usen las herramientas anteriores en el sistema judicial y penitenciario?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

GRUPO 3. TOTAL + DISEÑO MAQUINA + PROPORCIONADO

Se ha instaurado en el sistema penitenciario español un algoritmo predictivo diseñado mediante machine learning para evaluar el riesgo de reincidencia de un recluso que solicita la libertad condicional. El algoritmo considera factores como el historial delictivo, comportamiento en la prisión y participación en programas de rehabilitación para determinar la probabilidad de reincidencia.

Iván Pérez fue condenado a cumplir una pena de 7 años de prisión por delitos relacionados con robo agravado. Durante su tiempo en prisión, ha mostrado un comportamiento relativamente bueno, participando activamente en programas de rehabilitación, como cursos de educación, talleres de habilidades laborales y terapia de control de impulsos pero ha cometido 3 infracciones disciplinarias durante su tiempo en prisión. El algoritmo predictivo concluye que tiene una baja probabilidad de reincidencia; decide conceder la libertad condicional.

¿Está de acuerdo con la decisión tomada?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

Se ha instaurado en el sistema de justicia español un algoritmo predictivo diseñado mediante machine learning para determinar sentencias penales en casos de delitos no violentos. El sistema utiliza algoritmos que consideran la gravedad del delito, antecedentes penales del acusado y factores socioeconómicos para determinar la sentencia.

Carlos García, es acusado de un robo a mano armada en una joyería. Según la acusación, Carlos, junto con otro cómplice, ingresó a la joyería durante el horario laboral y amenazó al personal y a los clientes con un arma de fuego, llevándose joyas valuadas en 50.000€. Este delito se clasifica como robo con violencia y se considera un delito grave en la ley penal. El algoritmo predictivo decide imponer una sanción

de 4 años de prisión.

¿Está de acuerdo con la decisión tomada?

Totalmente en desacuerdo				Totalmente de acuerdo
1	2	3	4	5

¿En qué medida le parece aceptable que se usen las herramientas anteriores en el sistema judicial y penitenciario?

Totalmente en desacuerdo				Totalmente de acuerdo
1	2	3	4	5

GRUPO 4. TOTAL + DISEÑO MAQUINA + DESPROPORCIONADO.

Se ha instaurado en el sistema penitenciario español un algoritmo predictivo diseñado mediante machine learning para evaluar el riesgo de reincidencia de un recluso que solicita la libertad condicional. El algoritmo considera factores como el historial delictivo, comportamiento en la prisión y participación en programas de rehabilitación para determinar la probabilidad de reincidencia.

Iván Pérez fue condenado a cumplir una pena de 7 años de prisión por delitos relacionados con robo agravado. Durante su tiempo en prisión, ha mostrado un comportamiento relativamente bueno, participando activamente en programas de rehabilitación, como cursos de educación, talleres de habilidades laborales y terapia de control de impulsos pero ha cometido 3 infracciones disciplinarias durante su tiempo en prisión. El algoritmo predictivo concluye que tiene una baja probabilidad de reincidencia; decide rechazar la libertad condicional.

¿Está de acuerdo con la decisión tomada por el algoritmo?

¿Está de acuerdo con la decisión tomada por el juez?

Totalmente en desacuerdo				Totalmente de acuerdo
1	2	3	4	5

Se ha instaurado en el sistema de justicia español un algoritmo predictivo diseñado mediante machine learning para determinar sentencias penales en casos de delitos no violentos. El sistema utiliza algoritmos que consideran la gravedad del delito, antecedentes penales del acusado y factores socioeconómicos para determinar la sentencia.

Carlos García, es acusado de un robo a mano armada en una joyería. Según la acusación, Carlos, junto con otro cómplice, ingresó a la joyería durante el horario laboral y amenazó al personal y a los clientes con un arma de fuego, llevándose joyas valuadas en una gran cantidad de dinero. Este delito se clasifica como robo con violencia y se considera un delito grave en la ley penal. El algoritmo predictivo decide imponer una multa de 10.000€ sin prisión.

¿Está de acuerdo con la decisión tomada por el algoritmo?

¿Está de acuerdo con la decisión tomada por el juez?

Totalmente en desacuerdo				Totalmente de acuerdo
1	2	3	4	5

¿En qué medida le parece aceptable que se usen las herramientas anteriores en el sistema judicial y penitenciario?

Totalmente en desacuerdo				Totalmente de acuerdo
1	2	3	4	5

GRUPO 5. SEMI + DISEÑO HUMANO + PROPORCIONADO

Se ha instaurado en el sistema penitenciario español un algoritmo predictivo diseñado por expertos en psicología, criminología y derecho para informar a los jueces sobre el riesgo de reincidencia de un recluso que solicita la libertad condicional. El algoritmo considera factores como el historial delictivo, comportamiento en la prisión y participación en programas de rehabilitación para determinar la probabilidad de reincidencia que permita informar al comité en la toma de decisión.

Iván Pérez fue condenado a cumplir una pena de 7 años de prisión por delitos relacionados con robo agravado. Durante su tiempo en prisión, ha mostrado un comportamiento relativamente bueno, participando activamente en programas de rehabilitación, como cursos de educación, talleres de habilidades laborales y terapia de control de impulsos pero ha cometido 3 infracciones disciplinarias durante su tiempo en prisión.

El algoritmo predictivo, tras evaluar la solicitud de libertad condicional, informó al comité encargado de evaluar dichas solicitudes sobre la baja probabilidad de reincidencia; concede la libertad condicional. Estos revisan el informe generado por el algoritmo predictivo junto con otros informes y evaluaciones pertinentes. Basándose en la evaluación del algoritmo y los informes, concluyen que hay una baja probabilidad de reincidencia; deciden conceder la libertad condicional.

¿Está de acuerdo con la decisión tomada por el algoritmo?

¿Está de acuerdo con la decisión tomada por el juez?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

Se ha instaurado en el sistema de justicia español un algoritmo predictivo diseñado por expertos en psicología, criminología y derecho para informar a los jueces y determinar las sentencias penales en casos de delitos no violentos. El sistema utiliza

algoritmos que consideran la gravedad del delito, antecedentes penales del acusado y factores socioeconómicos para informar a los jueces en la toma de decisión

Carlos García, es acusado de un robo a mano armada en una joyería. Según la acusación, Carlos, junto con otro cómplice, ingresó a la joyería durante el horario laboral y amenazó al personal y a los clientes con un arma de fuego, llevándose joyas valuadas en una gran cantidad de dinero. Este delito se clasifica como robo con violencia y se considera un delito grave en la ley penal.

El algoritmo predictivo establece una sanción de 4 años de prisión. El juez encargado del caso revisa la decisión algorítmica y los hechos, decide imponer una sanción de 4 años de prisión.

¿Está de acuerdo con la decisión tomada por el algoritmo?

¿Está de acuerdo con la decisión tomada por el juez?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

¿En qué medida le parece aceptable que se usen las herramientas anteriores en el sistema judicial y penitenciario?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

GRUPO 6. SEMI + DISEÑO HUMANO + DESPROPORCIONADO

Se ha instaurado en el sistema penitenciario español un algoritmo predictivo diseñado por expertos en psicología, criminología y derecho para informar a los jueces sobre el riesgo de reincidencia de un recluso que solicita la libertad condicional. El algoritmo considera factores como el historial delictivo, comportamiento en la prisión y participación en programas de rehabilitación para determinar la probabilidad de reincidencia que permita informar al comité en la toma de decisión.

Iván Pérez fue condenado a cumplir una pena de 7 años de prisión por delitos relacionados con robo agravado. Durante su tiempo en prisión, ha mostrado un comportamiento relativamente bueno, participando activamente en programas de rehabilitación, como cursos de educación, talleres de habilidades laborales y terapia de control de impulsos pero ha cometido 3 infracciones disciplinarias durante su tiempo en prisión.

El algoritmo predictivo, tras evaluar la solicitud de libertad condicional, informó al comité encargado de evaluar dichas solicitudes sobre la baja probabilidad de reincidencia; concede la libertad condicional. Estos revisan el informe generado por el algoritmo predictivo junto con otros informes y evaluaciones pertinentes. Basándose en la evaluación del algoritmo y los informes, concluyen que hay una alta probabilidad de reincidencia; deciden rechazar la libertad condicional.

¿Está de acuerdo con la decisión tomada por el algoritmo?

¿Está de acuerdo con la decisión tomada por el juez?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

Se ha instaurado en el sistema de justicia español un algoritmo predictivo diseñado por expertos en psicología, criminología y derecho para informar a los jueces y determinar las sentencias penales en casos de delitos no violentos. El sistema utiliza algoritmos que consideran la gravedad del delito, antecedentes penales del acusado

y factores socioeconómicos para informar a los jueces en la toma de decisión

Carlos García, es acusado de un robo a mano armada en una joyería. Según la acusación, Carlos, junto con otro cómplice, ingresó a la joyería durante el horario laboral y amenazó al personal y a los clientes con un arma de fuego, llevándose joyas valuadas en una gran cantidad de dinero. Este delito se clasifica como robo con violencia y se considera un delito grave en la ley penal.

El algoritmo predictivo establece una sanción de 4 años de prisión. El juez encargado del caso revisa la decisión algorítmica y los hechos; decide imponer una multa de 10.000€ sin prisión.

¿Está de acuerdo con la decisión tomada por el algoritmo?

¿Está de acuerdo con la decisión tomada por el juez?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

¿En qué medida le parece aceptable que se usen las herramientas anteriores en el sistema judicial y penitenciario?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

GRUPO 7. SEMI + DISEÑO MAQUINA + PROPORCIONADO.

Se ha instaurado en el sistema penitenciario español un algoritmo predictivo diseñado por machine learning para informar a los jueces sobre el riesgo de reincidencia de un recluso que solicita la libertad condicional. El algoritmo considera factores como el historial delictivo, comportamiento en la prisión y

participación en programas de rehabilitación para determinar la probabilidad de reincidencia que permita informar al comité en la toma de decisión.

Iván Pérez fue condenado a cumplir una pena de 7 años de prisión por delitos relacionados con robo agravado. Durante su tiempo en prisión, ha mostrado un comportamiento relativamente bueno, participando activamente en programas de rehabilitación, como cursos de educación, talleres de habilidades laborales y terapia de control de impulsos pero ha cometido 3 infracciones disciplinarias durante su tiempo en prisión.

El algoritmo predictivo, tras evaluar la solicitud de libertad condicional, informó al comité encargado de evaluar dichas solicitudes sobre la baja probabilidad de reincidencia; concede la libertad condicional. Estos revisan el informe generado por el algoritmo predictivo junto con otros informes y evaluaciones pertinentes. Basándose en la evaluación del algoritmo y los informes, concluyen que hay una baja probabilidad de reincidencia; deciden conceder la libertad condicional.

¿Está de acuerdo con la decisión tomada por el algoritmo?

¿Está de acuerdo con la decisión tomada por el juez?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

Se ha instaurado en el sistema de justicia español un algoritmo predictivo diseñado por machine learning para informar a los jueces y determinar las sentencias penales en casos de delitos no violentos. El sistema utiliza algoritmos que consideran la gravedad del delito, antecedentes penales del acusado y factores socioeconómicos para informar a los jueces en la toma de decisión

Carlos García, es acusado de un robo a mano armada en una joyería. Según la acusación, Carlos, junto con otro cómplice, ingresó a la joyería durante el horario laboral y amenazó al personal y a los clientes con un arma de fuego, llevándose joyas

valuadas en una gran cantidad de dinero. Este delito se clasifica como robo con violencia y se considera un delito grave en la ley penal.

El algoritmo predictivo establece una sanción de 4 años de prisión. El juez encargado del caso revisa la decisión algorítmica y los hechos, decide imponer una sanción de 4 años de prisión.

¿Está de acuerdo con la decisión tomada por el algoritmo?

¿Está de acuerdo con la decisión tomada por el juez?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

¿En qué medida le parece aceptable que se usen las herramientas anteriores en el sistema judicial y penitenciario?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

GRUPO 8. SEMI + DISEÑO MAQUINA + DESPROPORCIONADO

Se ha instaurado en el sistema penitenciario español un algoritmo predictivo diseñado por machine learning para informar a los jueces sobre el riesgo de reincidencia de un recluso que solicita la libertad condicional. El algoritmo considera factores como el historial delictivo, comportamiento en la prisión y participación en programas de rehabilitación para determinar la probabilidad de reincidencia que permita informar al comité en la toma de decisión.

Iván Pérez fue condenado a cumplir una pena de 7 años de prisión por delitos relacionados con robo agravado. Durante su tiempo en prisión, ha mostrado un comportamiento relativamente bueno, participando activamente en programas de rehabilitación, como cursos de educación, talleres de habilidades laborales y terapia de control de impulsos pero ha cometido 3 infracciones disciplinarias durante su tiempo en prisión.

El algoritmo predictivo, tras evaluar la solicitud de libertad condicional, informó al comité encargado de evaluar dichas solicitudes sobre la baja probabilidad de reincidencia; concede la libertad condicional. Estos revisan el informe generado por el algoritmo predictivo junto con otros informes y evaluaciones pertinentes. Basándose en la evaluación del algoritmo y los informes, concluyen que hay una alta probabilidad de reincidencia; deciden rechazar la libertad condicional.

¿Está de acuerdo con la decisión tomada por el algoritmo?

¿Está de acuerdo con la decisión tomada por el juez?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

Se ha instaurado en el sistema de justicia español un algoritmo predictivo diseñado por machine learning para informar a los jueces y determinar las sentencias penales en casos de delitos no violentos. El sistema utiliza algoritmos que consideran la

gravedad del delito, antecedentes penales del acusado y factores socioeconómicos para informar a los jueces en la toma de decisión

Carlos García, es acusado de un robo a mano armada en una joyería. Según la acusación, Carlos, junto con otro cómplice, ingresó a la joyería durante el horario laboral y amenazó al personal y a los clientes con un arma de fuego, llevándose joyas valuadas en una gran cantidad de dinero. Este delito se clasifica como robo con violencia y se considera un delito grave en la ley penal.

El algoritmo predictivo establece una sanción de 4 años de prisión. El juez encargado del caso revisa la decisión algorítmica y los hechos; decide imponer una multa de 10.000€ sin prisión.

¿Está de acuerdo con la decisión tomada por el algoritmo?

¿Está de acuerdo con la decisión tomada por el juez?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

¿En qué medida le parece aceptable que se usen las herramientas anteriores en el sistema judicial y penitenciario?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

GRUPO 9. HUMANO + PROPORCIONADO

En el sistema de justicia español los jueces evalúan el riesgo de reincidencia de un recluso que solicita la libertad condicional. Estos se basan en factores como el historial delictivo, comportamiento en la prisión y participación en programas de rehabilitación para determinar la probabilidad de reincidencia.

Iván Pérez fue condenado a cumplir una pena de 7 años de prisión por delitos relacionados con robo agravado. Durante su tiempo en prisión, ha mostrado un comportamiento relativamente bueno, participando activamente en programas de rehabilitación, como cursos de educación, talleres de habilidades laborales y terapia de control de impulsos pero ha cometido 3 infracciones disciplinarias durante su tiempo en prisión. El comité encargado de evaluar las solicitudes de libertad condicional revisa los informes y evaluaciones pertinentes, concluyen que tiene una baja probabilidad de reincidencia; concede la libertad condicional.

¿Está de acuerdo con la decisión tomada?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

En el sistema de justicia español los jueces determinan las sentencias penales en casos de delitos no violentos. Estos se basan en la gravedad del delito, antecedentes penales del acusado y factores socioeconómicos para determinar la sentencia.

Carlos García, es acusado de un robo a mano armada en una joyería. Según la acusación, Carlos, junto con otro cómplice, ingresó a la joyería durante el horario laboral y amenazó al personal y a los clientes con un arma de fuego, llevándose joyas valuadas en una gran cantidad de dinero. Este delito se clasifica como robo con violencia y se considera un delito grave en la ley penal. El juez encargado del caso decide imponer una sanción de 4 años de prisión.

¿Está de acuerdo con la decisión tomada?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

GRUPO 10. HUMANO + DESPROPORCIONADO.

En el sistema de justicia español los jueces evalúan el riesgo de reincidencia de un recluso que solicita la libertad condicional. Estos se basan en factores como el historial delictivo, comportamiento en la prisión y participación en programas de rehabilitación para determinar la probabilidad de reincidencia.

Iván Pérez fue condenado a cumplir una pena de 7 años de prisión por delitos relacionados con robo agravado. Durante su tiempo en prisión, ha mostrado un comportamiento relativamente bueno, participando activamente en programas de rehabilitación, como cursos de educación, talleres de habilidades laborales y terapia de control de impulsos pero ha cometido 3 infracciones disciplinarias durante su tiempo en prisión. El comité encargado de evaluar las solicitudes de libertad condicional revisa los informes y evaluaciones pertinentes, concluye que tiene una baja probabilidad de reincidencia; rechaza la libertad condicional.

¿Está de acuerdo con la decisión tomada?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

En el sistema de justicia español los jueces determinan las sentencias penales en casos de delitos no violentos. Estos se basan en la gravedad del delito, antecedentes penales del acusado y factores socioeconómicos para determinar la sentencia.

Carlos García, es acusado de un robo a mano armada en una joyería. Según la

acusación, Carlos, junto con otro cómplice, ingresó a la joyería durante el horario laboral y amenazó al personal y a los clientes con un arma de fuego, llevándose joyas valuadas en una gran cantidad de dinero. Este delito se clasifica como robo con violencia y se considera un delito grave en la ley penal. El juez encargado del caso decide imponer una multa de 10.000€ sin prisión.

¿Está de acuerdo con la decisión tomada?

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	**2**	**3**	**4**	**5**

BLOQUE 3.- Aceptación de las herramientas algorítmicas.[15]

Las herramientas algorítmicas son programas de software que utilizan algoritmos para realizar tareas específicas con o sin la necesidad de intervención humana. Estas herramientas están estrechamente conectadas con la inteligencia artificial (IA), ya que la IA hace referencia al desarrollo de sistemas y algoritmos que pueden realizar tareas que normalmente requerirían la inteligencia humana, como el aprendizaje, la toma de decisiones y el reconocimiento de patrones. Las herramientas algorítmicas se utilizan en muchos campos de la IA, como el aprendizaje automático, la minería de datos y la automatización de procesos. A continuación, se le realizará una serie de preguntas generales acerca del conocimiento y de las características de las herramientas algorítmicas en el ámbito judicial.

	Totalmente en desacuerdo				**Totalmente de acuerdo**
Estoy familiarizado con el uso de las herramientas algorítmicas en el sistema de justicia.	1	2	3	4	5
Los jueces y los abogados deben recibir capacitación sobre el posible uso de herramientas algorítmicas en el sistema de justicia.	1	2	3	4	5
El uso de las herramientas algorítmicas en el sistema de justicia puede perpetuar sesgos y discriminación existentes.	1	2	3	4	5
El uso de las herramientas algorítmicas en el sistema de justicia podría ser susceptible a manipulación o sabotaje.	1	2	3	4	5
Las herramientas algorítmicas deberían ser reguladas para evitar posibles discriminaciones o sesgos en las decisiones judiciales.	1	2	3	4	5
Las herramientas algorítmicas utilizadas en el sistema de justicia son objetivas en sus evaluaciones.	1	2	3	4	5

[15] El cuestionario se encuentra publicado en los anexos del artículo Pérez Domínguez, S., & Simón Castellano, P. (2023). *Attitudes and perceptions regarding algorithmic judicial judgement: barriers to innovation in the judicial system*. IDP: Revista de Internet, Derecho y Política, (39), 6. https://doi.org/10.7238/idp.v0i39.417206"

Las herramientas algorítmicas son más justas en comparación con la intuición humana.	1	2	3	4	5
Las herramientas algorítmicas son más imparciales que los jueces humanos en la toma de decisiones judiciales.	1	2	3	4	5

BLOQUE 4. Aceptación de uso de herramientas algorítmicas automatizadas

En la actualidad, el desarrollo de la inteligencia artificial y de las herramientas algorítmicas automatizadas en el sistema de justicia se encuentra en constante actualización y desarrollo. A continuación, nos gustaría conocer su opinión acerca de la inclusión dichas herramientas en diferentes ámbitos del sistema judicial.

En cuál de los siguientes casos que se le expone estaría de acuerdo en que se utilicen las herramientas algorítmicas automatizadas como único método del proceso o con una función relevante y decisiva *(sin intervención humana).*

	Totalmente desacuerdo				**Totalmente de acuerdo**
Predicción del nivel de riesgo de reincidencia	1	2	3	4	5
Decisión sobre el fondo o determinación del fallo (conclusión de un caso legal que resuelve la cuestión principal en disputa)	1	2	3	4	5
Decisión de concesión del tercer grado (acceso a la libertad por parte una vez cumplida una parte de la condena con buen comportamiento)	1	2	3	4	5
Chatbots para responder preguntas comunes y programar citas	1	2	3	4	5
Averiguación del patrimonio o del domicilio	1	2	3	4	5
Adopción de la decisión judicial en el ámbito laboral.	1	2	3	4	5

Adopción de la decisión judicial en la jurisdicción civil (disputas en temas como propiedad, contratos, familia y herencias)	1	2	3	4	5
Monitorear el cumplimiento de las obligaciones contractuales (realizar un seguimiento de los términos y condiciones establecidos en un contrato para asegurarse de que se estén cumpliendo satisfactoriamente).	1	2	3	4	5
Análisis de tratados y acuerdos internacionales (revisión del texto del acuerdo y la identificación de los compromisos específicos)	1	2	3	4	5
Determinación de la cuantía de la indemnización en accidentes.	1	2	3	4	5
Determinar la veracidad de una denuncia.	1	2	3	4	5

En cuál de los siguientes casos que se le expone estaría de acuerdo en que se utilicen las herramientas algorítmicas como ayuda al operador (por ejemplo: abogados, jueces, fiscales, notarios, registradores, mediadores, etc) en el proceso (función accesoria).

	Totalmente en desacuerdo				**Totalmente de acuerdo**
Predicción del nivel de riesgo de reincidencia	1	2	3	4	5
Decisión sobre el fondo o determinación del fallo (conclusión de un caso legal que resuelve la cuestión principal en disputa)	1	2	3	4	5
Decisión de concesión del tercer grado (acceso a la libertad por parte una vez cumplida una parte de la condena con buen comportamiento)	1	2	3	4	5

Chatbots para responder preguntas comunes y programar citas	1	2	3	4	5
Averiguación del patrimonio o del domicilio	1	2	3	4	5
Adopción de la decisión judicial en el ámbito laboral.	1	2	3	4	5
Adopción de la decisión judicial en la jurisdicción civil (disputas en temas como propiedad, contratos, familia y herencias)	1	2	3	4	5
Monitorear el cumplimiento de las obligaciones contractuales (realizar un seguimiento de los términos y condiciones establecidos en un contrato para asegurarse de que se estén cumpliendo satisfactoriamente)	1	2	3	4	5
Análisis de tratados y acuerdos internacionales (revisión del texto del acuerdo y la identificación de los compromisos específicos)	1	2	3	4	5
Determinación de la cuantía de la indemnización en accidentes.	1	2	3	4	5
Determinar la veracidad de una denuncia.	1	2	3	4	5

Anexo 3. Código abierto de los análisis de datos del estudio 2 del capítulo 6.

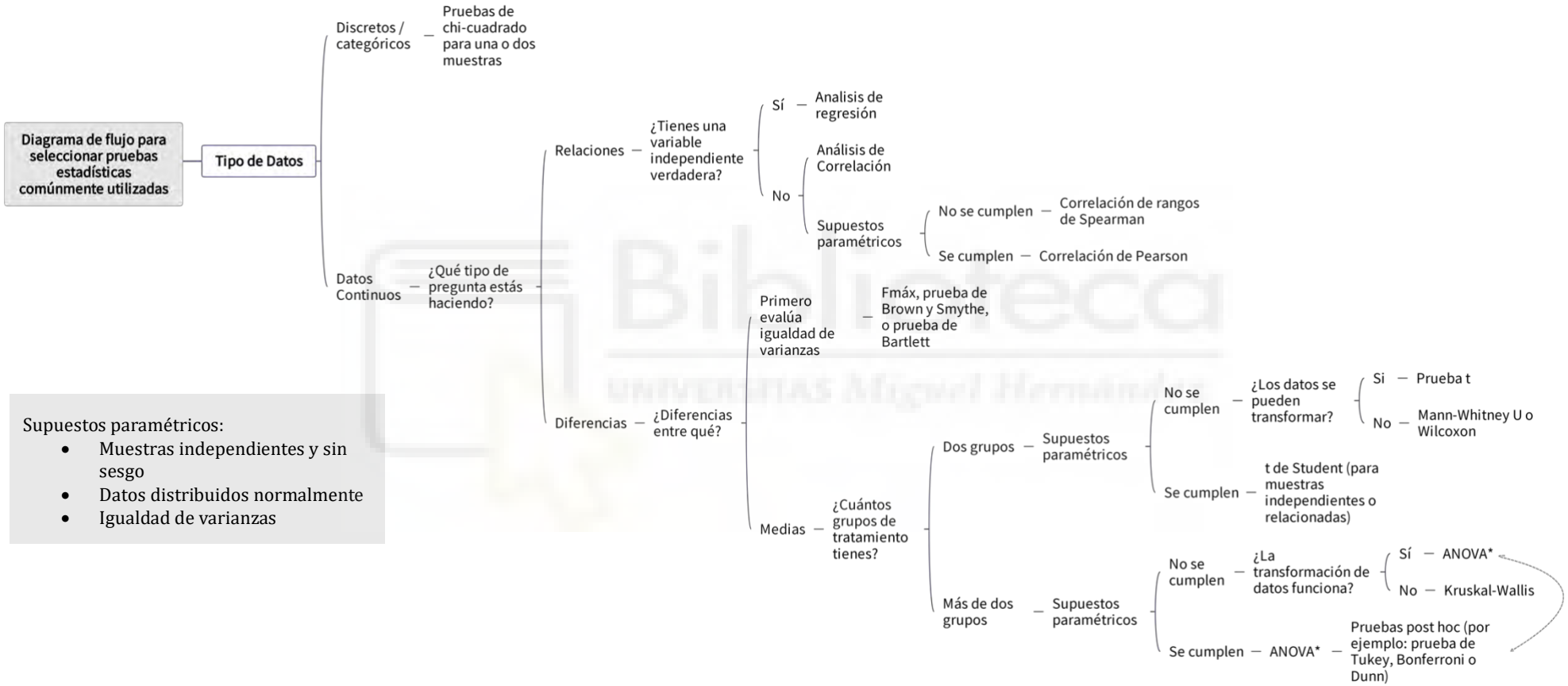
```
install.packages("haven")

library(haven)

install.packages("MASS")

library(MASS)

install.packages(c("broom", "effectsize", "dplyr"))

library(broom)

library(effectsize)

library(dplyr)

data<-datos_ius_machina_NO_exper

summary(data)

head(data)

names(data)[names(data) == "P1"] <- "SEXO"

names(data)[names(data) == "P2"] <- "EDAD"

names(data)[names(data) == "P3"] <- "ESTUDIOS"

names(data)[names(data) == "P4"] <- "FORMACION"

names(data)[names(data) == "P5"] <- "FAMILIARIDAD_ALGORITMICA"

names(data)[names(data) == "P7_1"] <- "FAMILIARIDAD_ALGORITMOS_SJP"

names(data)[names(data) == "P7_2"] <- "CAPACITACION"

names(data)[names(data) == "P7_3"] <- "PERPETUACION_SESGOS"

names(data)[names(data) == "P7_4"] <- "MANIPULACION"
```

```
names(data)[names(data) == "P7_5"] <- "REGULACIÓN"

names(data)[names(data) == "P7_6"] <- "OBJETIVIDAD"

names(data)[names(data) == "P7_7"] <- "JUSTAS"

names(data)[names(data) == "P7_8"] <- "IMPARCIALES"

head(data)

data$EDAD <- as.numeric(as.character(data$EDAD))

data$grupo_edad <- cut(data$EDAD,

                       breaks = c(18, 30, 45, 55, 65, Inf),   # límites de los grupos

                       labels = c("18-29", "30-44", "45-54", "55-64", "65+"),

                       right = TRUE)

table(data$grupo_edad)

vars_likert <- c("FORMACION", "FAMILIARIDAD_ALGORITMICA",
"FAMILIARIDAD_ALGORITMOS_SJP",

          "CAPACITACION", "PERPETUACION_SESGOS", "MANIPULACION",

          "REGULACIÓN", "OBJETIVIDAD", "JUSTAS", "IMPARCIALES")

data[vars_likert] <- lapply(data[vars_likert], factor)

is.numeric(data$ACEP_GENERAL_IA)

modelo <- lm(ACEP_GENERAL_HOTL ~ SEXO + grupo_edad +
FAMILIARIDAD_ALGORITMOS_SJP + CAPACITACION + PERPETUACION_SESGOS +
MANIPULACION + REGULACIÓN + OBJETIVIDAD + JUSTAS + IMPARCIALES, data = data)

tab_base <- tidy(modelo, conf.int = TRUE, conf.level = 0.95)

tab_beta <- standardize_parameters(modelo) |>

  select(term = Parameter, beta = Std_Coefficient)
```

```
tabla_final <- tab_base %>%
 left_join(tab_beta, by = "term") %>%
 mutate(
  IC95 = paste0("[", round(conf.low, 3), "; ", round(conf.high, 3), "]")
 ) %>%
 select(
  Predictor = term,
  B = estimate,
  SE_B = std.error,
  Beta = beta,
  p = p.value,
  IC95
 )
print(tabla_final, n = Inf)
summary(modelo)
modelo <- lm(ACEP_GENERAL_HITL ~ SEXO + grupo_edad + FAMILIARIDAD_ALGORITMOS_SJP + CAPACITACION + PERPETUACION_SESGOS + MANIPULACION + REGULACIÓN + OBJETIVIDAD + JUSTAS + IMPARCIALES, data = data)
tab_base <- tidy(modelo, conf.int = TRUE, conf.level = 0.95)
tab_beta <- standardize_parameters(modelo) |>
 select(term = Parameter, beta = Std_Coefficient)
tabla_final <- tab_base %>%
 left_join(tab_beta, by = "term") %>%
```

```
  mutate(
    IC95 = paste0("[", round(conf.low, 3), "; ", round(conf.high, 3), "]")
  ) %>%
  select(
    Predictor = term,
    B = estimate,
    SE_B = std.error,
    Beta = beta,
    p = p.value,
    IC95
  )
print(tabla_final, n = Inf)
summary(modelo)
```

Anexo 4. Árbol de decisión para los análisis estadísticos.



Supuestos paramétricos:

- Muestras independientes y sin sesgo
- Datos distribuidos normalmente
- Igualdad de varianzas

Adaptado y traducido de: Gerwien (2014).