IUS ET SCIENTIA
2021 • Vol. 7 • Nº 2• ISSN 2444-8478
https://editorial.us.es/es/revistas/ius-et-scientia
https://dx.doi.org/10.12795/IETSCIENTIA • © Editorial Universidad de Sevilla 2021

Algoritmos en el estrado, ¿realmente los aceptamos? Percepciones del uso de la inteligencia artificial en la toma de decisiones jurídico-penales

ALGORITHMS ON THE STAND, DO WE REALLY ACCEPT THEM? PERCEPTIONS OF THE USE OF ARTIFICIAL INTELLIGENCE IN CRIMINAL-LEGAL DECISION-MAKING

África María Morales Moreno
Contratada predoctoral (FPU). Departamento de Derecho Civil, Penal y Procesal
Universidad de Córdoba
africamorales14@gmail.com 0000-0003-1466-2833

Recibido: 3 de diciembre de 2021 | Aceptado: 19 de diciembre de 2021

RESUMEN

La irrupción que las prácticas basadas en la evidencia, la automatización de decisiones y la inteligencia artificial han tenido en nuestra sociedad también ha alcanzado al sistema de justicia penal. Jueces y operadores jurídicos comienzan a interactuar con este tipo de herramientas aún sin tener la información suficiente sobre su modo de empleo ni sobre el impacto que realmente pueden llegar a tener. Todo ello, unido a la falta de regulación legal y de requisitos éticos para su utilización, parece estar generando entre la ciudadanía controversias, críticas e incluso cierto rechazo hacia la implementación de tales tecnologías. Con una muestra de 359 participantes, este estudio ofrece una primera aproximación al grado de aceptación ciudadana que existe en relación con el uso de la inteligencia artificial para la toma de decisiones jurídico-penales. Los resultados obtenidos apuntan a que tal nivel de aceptación es bajo, lo cual abre camino al debate sobre qué condiciones y límites deben imponerse para que la aplicación de estas tecnologías sea legítima y acorde a los principios de todo Estado social, democrático y de Derecho.

PALABRAS CLAVE

Justicia penal
Inteligencia artificial (IA)
Algoritmo
Riesgo de reincidencia

ABSTRACT

The irruption that evidence-based practices, decision automation and artificial intelligence have had in our society has also reached the criminal justice system. Judges and legal operators are already interacting with these types of tools without having sufficient information about how they are used or the impact they can really have. All of this, together with the lack of legal regulation and ethical requirements for their use, seems to be generating controversy, criticism and even a certain rejection of the implementation of

KEYWORDS

Criminal justice
Artificial intelligence (AI)
Algorithm
Recidivism risk

these technologies among citizens. With a sample of 359 participants, this study offers a first approximation of the degree of public acceptance of the use of artificial intelligence in legal-criminal decision-making. The results obtained suggest that this level of acceptance is low, which opens the way for a debate on what conditions and limits should be imposed for the application of these technologies to be legitimate and in accordance with the principles of any social, democratic and rule of law state.

I. Introducción

En plena era digital, rodeados de cantidades cada vez mayores de datos y motivados por los avances en inteligencia artificial (IA), la toma de decisiones en las sociedades contemporáneas se delega cada vez más en estos sistemas. En este contexto de "datificación" (Van Dijck, 2014), nos dejamos guiar diariamente por las recomendaciones de Siri o Cortana; "chateamos" con nuestro gestor bancario como si verdaderamente hubiera una persona física respondiendo a nuestras dudas al otro lado de la pantalla; e incluso empezamos a desplazarnos en vehículos que no están siendo conducidos por ningún agente humano. Asistimos, de forma inconsciente, a una profunda transformación en todos los ámbitos que nos rodean diariamente, incluido el campo la justicia penal y la organización penitenciaria.

De esta forma, las tradicionales funciones de evaluación judicial humana se complementan cada vez más con una gama de herramientas actuariales, algorítmicas, de aprendizaje automático e IA que pretenden proporcionar capacidades predictivas precisas y evaluaciones de riesgo objetivas y consistentes (Martínez Garay, 2018; Miró Llinares, 2018; Solar Cayón, 2020). La concesión de permisos penitenciarios, la gestión de los recursos e incluso la puesta en libertad condicional de un sujeto son ejemplos de decisiones que comienzan a ser tomadas en base a los resultados arrojados por este tipo de herramientas. Sin embargo, las consecuencias, conocidas y desconocidas, derivadas del uso de este tipo de instrumentos, junto con la falta de regulación legal y de requisitos éticos para su utilización, están generando múltiples controversias y críticas por parte de la población.

Ante este panorama, el objetivo principal que se pretende lograr con este trabajo es conocer el nivel de aceptación que existe entre la ciudadanía sobre el uso de herramientas dotadas de IA en la toma de decisiones jurídico-penales mediante la realización de un estudio empírico. La elección de este objeto de estudio parte de la consideración de que, en un Estado Social y Democrático de Derecho en el que se reconoce y garantiza la participación ciudadana, la automatización de decisiones y el uso de la IA solo triunfará y será legítimo si es aceptado por la población a la que se dirige.

Para alcanzar tal objetivo, se llevará a cabo un estudio exhaustivo de este tipo de herramientas y de su progresivo desarrollo e implementación. A continuación, se realizará una revisión bibliográfica de aquellos trabajos que ya hayan analizado el nivel de aceptación social de la automatización de decisiones, tanto en el ámbito jurídico como en otros campos de la sociedad. Por último, se presentará un caso real acaecido

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

en California y que pone de manifiesto la importancia de tener en cuenta las opiniones de la ciudadanía antes de incorporar el uso de herramientas de IA en el ámbito de la justicia penal. Finalmente, se presentará el estudio realizado y los resultados obtenidos con la finalidad última de proporcionar una primera aproximación a las percepciones que la ciudadanía tiene actualmente sobre el empleo de estas herramientas en el sistema de justicia penal.

II. Las herramientas de inteligencia artificial en el ámbito de la justicia penal

1. Los orígenes de su aplicación: las herramientas de valoración del riesgo

El Derecho Penal no se ocupa solo de tipificar conductas y proteger bienes jurídicos sino que son numerosas las cuestiones que aborda y que, cada vez más, comparte con otras ramas de conocimiento. Entre ellas, la peligrosidad del reo o del condenado ha sido desde siempre un punto de unión con la Criminología, ya que es un criterio fundamental que debe ser valorado a hora de determinar ciertos aspectos en un proceso penal o en el ámbito penitenciario (como conceder o no permisos penitenciarios o la libertad condicional, determinar el régimen penitenciario de un determinado preso, etc.). Para tomar estas decisiones los jueces y el resto de los operadores jurídicos se han apoyado tradicionalmente en informes o valoraciones de psiquiatras, médicos forenses u otros especialistas procedentes de las juntas de tratamiento en el ámbito penitenciario; todos ellos expertos en la materia. Estos informes eran elaborados en base a la libre interpretación de la información del agresor, delincuente o paciente por parte del profesional, sin utilizar ningún tipo de herramienta o reglas fijas y basándose exclusivamente en su propio conocimiento, en su experiencia profesional individual (Martínez Garay y Montes Suay, 2018).

Esta tradicional forma de evaluar la peligrosidad de un sujeto es el denominado método clínico puro o juicio clínico no estructurado y, pese a ser conscientes de la ambigüedad y limitaciones del concepto de "peligrosidad" como predictor de la conducta violenta, tales valoraciones se llevaban a cabo porque la ley así lo exigía en determinados casos y porque no existía ningún otro método para ello. Sin embargo, en las últimas décadas del siglo XX, las ciencias sociales y médicas llevaron a cabo un examen muy crítico sobre las posibilidades reales de detectar "científicamente" la peligrosidad y concluyeron que las dificultades teóricas y prácticas para prever la conducta futura de las personas, unidas a las altas tasas de error que ya se estaban constatando empíricamente, restaban legitimidad a las consecuencias jurídicas restrictivas de derechos que la legislación hacía depender de tales pronósticos (Martínez Garay, 2018).

La principal consecuencia de tales descubrimientos fue la sustitución, por parte de la Criminología, del concepto de "peligrosidad" por el de "riesgo". Para este nuevo enfoque la "peligrosidad" es considerada una cualidad estática, presente en el sujeto de

IUS ET SCIENTIA • 2021
Vol. 7 • N° 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

forma permanente desde el momento en que comete un primer delito. En cambio, el concepto de "riesgo" hace referencia a un conjunto de factores personales y ambientales que se consideran dinámicos, que favorecen en mayor o menor grado la comisión de nuevos delitos y que permiten elaborar pronósticos en términos de probabilidad sobre el riesgo de reincidencia futura que presenta el sujeto. Ya no se trataba de determinar si el individuo posee o no la cualidad subjetiva de peligroso (pues, efectivamente, la posee), sino de evaluar la presencia o ausencia de tales factores en el sujeto y en ese determinado momento (Andrés Pueyo y Redondo Illescas, 2007; Andrés Pueyo y Echeburúa, 2010; Loinaz, 2017). Y, junto a este cambio de conceptualización también se produjo una transformación en la forma de medir dicho riesgo.

Surgen, entonces, las herramientas de valoración del riesgo (HVR) en sus dos vertientes conocidas: actuariales y de juicio clínico estructurado. El funcionamiento de ambos tipos de HVR se basa en la observación empírica de grupos de sujetos y en la cuantificación y combinación estadística de los factores de riesgo (en ocasiones, también de protección) que concurren en ellos y que han demostrado estar significativamente asociados a la aparición de conducta violenta o delictiva (Martínez Garay y Montes Suay, 2018). Si bien, la principal diferencia entre ellas es que las actuariales proporcionan una estimación del riesgo de forma automática, calculada con un algoritmo a partir de la puntuación que el sujeto haya obtenido en los diferentes factores de riesgo que incluya el instrumento. En cambio, las de juicio clínico estructurado son guías que indican cuáles son los factores que deben ser tenidos en cuenta y cómo deben ser apreciados, pero en las que la evaluación final queda en manos del profesional que pondera el peso que tiene cada factor, e incluso puede añadir otros que la herramienta no contempla pero que él considera decisivos en el caso concreto (Loinaz, 2017).

Con esta nueva forma de proceder, el recelo y escepticismo antes mencionado empiezan a ser sustituidos por una mayor confianza hacia estas herramientas que permitían elaborar pronósticos basados en la evidencia y con mayor rigor y transparencia técnica (Andrés Pueyo, 2017). De esta forma, el éxito de estos instrumentos no se ha limitado al ámbito teórico o académico, sino que numerosos países los han en diversos estadios del proceso penal porque, también en este ámbito, el riesgo de comisión de futuros delitos debe ser tenido en cuenta. A nivel internacional, Canadá y EEUU fueron los primeros en hacerlo y a ellos se han sumado progresivamente otros países vecinos, como España desde hace 10 o 15 años (Martínez Garay y Montes Suay, 2018). A este respecto, especial mención merece en este trabajo la HVR *RisCanvi* que ha sido una importante contribución a la gestión y rehabilitación de los delincuentes en el sistema penitenciario catalán.

RisCanvi se diseñó para evaluar el riesgo de violencia en los centros penitenciarios de Cataluña y desde 2009 es la principal herramienta para prevenir la violencia carcelaria y gestionar la rehabilitación y la reincidencia de los presos. RisCanvi se puso en marcha en un momento en el que las cárceles catalanas se estaban masificando por el alargamiento de las penas de prisión y se tenía que encontrar un equilibrio entre el mantenimiento de los programas de tratamiento, la introducción de nuevas medidas penales alternativas y la respuesta a las constantes demandas de la sociedad. De esta

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

forma, RisCanvi fue diseñada para lograr dos objetivos principales: mejorar las predicciones individualizadas del riesgo de reincidencia; y generalizar el uso de HVR como procedimiento habitual entre los profesionales penitenciarios. A través de una evaluación estructurada de varios factores de riesgo, RisCanvi determina el nivel de riesgo presente para cada recluso en relación con tres posibles conductas violentas en el futuro: violencia autodirigida, violencia en las instalaciones, y probabilidad de cometer nuevos delitos violentos. Esta herramienta también estima el riesgo de quebrantamiento de permisos y otras situaciones similares (Soler, C., 2013; Singh, J. P., et al., 2018).

RisCanvi se ha probado empíricamente y sus parámetros métricos son similares a los de otras HVR. Concretamente, en el último estudio disponible sobre la validez predictiva de RisCanvi su autor concluyó que la sensibilidad de la herramienta era excelente - un 77,15% de los sujetos que había identificado como riesgo alto efectivamente reincidieron - y su efectividad aceptable - un 57,26% de los sujetos que clasificó como de riesgo bajo no reincidieron - (Capdevilla, M., et al, 2015). De esta forma, esta HVR se encuentra plenamente incorporada en varios procedimientos de gestión de los presos y es un recurso a disposición de los profesionales de los centros penitenciarios catalanes para potenciar la toma de decisiones efectiva en todos los ámbitos en los que la evaluación y gestión del riesgo de violencia son importantes (Soler, C., 2013; Singh, J. P., et al., 2018).

En definitiva, pese a que los textos jurídico-penales sigan aludiendo a la "peligrosidad" como fundamento para poder tomar ciertas decisiones (revisión o suspensión de la pena, imposición de medidas de seguridad, concesión de libertad condicional, etc.), lo que realmente interesa es conocer el riesgo de reincidencia que presente el procesado en cada caso concreto. Es por ello que el uso de las HVR se está extendiendo sin precedentes a todo el ámbito de la administración de la justicia penal y, en algunos países, están llegando incluso a ser utilizadas en la fase de determinación de la pena. Esta última aplicación es conocida como *evidence-based sentencing* y, dado que las consecuencias que su uso está generando no son para nada irrelevantes, en las líneas siguientes serán abordados los orígenes de su aplicación e implementación, así como los problemas a los que se enfrenta esta nueva práctica.

2. El evidence based-sentencing

El evidence based-sentencing encuentra su origen en las políticas del "tough on crime" y "la guerra contra las drogas" aplicadas en EEUU durante los años 80-90. El aumento del número de delitos castigados con cadena perpetua, la derogación de la *parole* (lo que equivale a nuestra libertad condicional) en muchos estados, la aprobación de leyes que imponían unas duras condenas mínimas de cumplimiento obligatorio, o la tipificación de penas más duras e internamientos indefinidos para cierto tipo de delincuentes fueron algunas de las reformas que caracterizaron a estas políticas. Como consecuencia de todo ello, se ocasionaron tremendas disparidades raciales (al afectar desproporcionadamente a los habitantes de bajos ingresos y de color), un aumento de los gastos federales y estatales, así como un incremento de la población penitenciaria y del número

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

de sentencias que imponían condenas privativas de libertad (Botnick, 2015; Martínez Garay, 2020). Ante este panorama, a finales de la década de los 90, investigadores y responsables de las políticas criminales se vieron obligados a buscar nuevas soluciones que permitirán paliar estas devastadoras consecuencias. Dicha búsqueda de respuestas coincidió en el tiempo con el auge que los métodos de evaluación del riesgo estaban teniendo en el ámbito policial y penitenciario, lo que llevó a reflexionar sobre la posibilidad de incorporar estas técnicas también en la fase de determinación penal. Surge, entonces, el evidence-based sentencing (Botnick, 2015).

La expresión evidence-based sentencing (EBS) hace referencia a la aplicación, en el momento de la imposición de la pena y durante la ejecución de la misma, de una serie de pautas y criterios que derivan de los resultados de investigaciones científicas rigurosas sobre sobre la capacidad predictiva de las modernas HVR. Se trata de que el juez pueda utilizar los datos de forma actuarial, en lugar de depender de sus propios juicios profesionales o "clínicos" (Star, 2014; Martínez Garay, 2020). Así, ante las elevadas tasas de encarcelamiento y los prejuicios raciales presentes en el sistema de justicia penal estadounidense, el EBS se presentaba como la mejor solución para disminuir el encarcelamiento masivo y, al mismo tiempo, proporcionar una mejor protección a la comunidad, reducir la reincidencia y aumentar la coherencia, la objetividad y la transparencia en la toma de decisiones jurídico penales, al ser estas adoptadas en base a estimaciones científicas sobre el riesgo de reincidencia (Starr, 2014; Botnick, 2015; Garrett y Monahan, 2018; Scurich y Krauss, 2019; Martínez Garay, 2020).

En l994, el estado de Virginia fue el primero en incorporar el uso de HVR en la fase de determinación de la pena (Botnick, 2015) y, desde esta primera experiencia, el estudio y desarrollo de estos instrumentos ha tenido una evolución exponencial. A fecha de 2019, ya eran 28 los estados que se habían sumado a esta nueva práctica (Stevenson y Doleac, 2019). Sin embargo, hoy en día, los objetivos perseguidos con estas reformas no están siendo alcanzados. Casi dos décadas después de la llegada del EBS, las tasas de encarcelamiento estadounidenses no se han visto reducidas. Por citar un ejemplo concreto, en el año 2013, el número de personas encarceladas en el estado de Missouri era de 41.998. Junto a ello, la aparición de una serie de amenazas para los derechos fundamentales y las garantías de los ciudadanos (Freeman, 2016; Martínez Garay, 2019) está dando lugar a un importante debate que cuestiona si realmente las prácticas basadas en la evidencia deben tener o no cabida en el ámbito de la justicia penal.

3. Problemas derivados de la aplicación de estas herramientas

La discusión sobre la aplicación de las HVR en el ámbito judicial y del EBS, parece tener su origen en el año 2013, cuando Eric Loomis se convirtió en la primera persona condenada a seis años de prisión y otros cinco en régimen de libertad vigilada en base al nivel de riesgo que presentaba según la herramienta COMPAS (Northpointe, 1998). La defensa del condenado recurrió la sentencia alegando que se había vulnerado el derecho a un proceso con todas las garantías porque no podía discutir los métodos utilizados

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

por el sistema COMPAS, pues su algoritmo era secreto y solo lo conocía la empresa que lo había desarrollado. Tales argumentos no fueron acogidos por la Corte Suprema del Estado de Wisconsin, que ratificó la condena (Freeman, K., 2016; Martínez Garay, 2018). Pese a ello, en el caso *State vs. Loomis* encuentra su origen el actual debate acerca del uso que debe darse a estas herramientas.

Dudas sobre el nivel de acierto de estas herramientas, la presencia de sesgos en los datos que utiliza, la posible vulneración de derechos fundamentales, o su falta de encaje en las teorías convencionales sobre las finalidades de la pena son algunos de los aspectos que comienzan a ser cuestionados. Todo ello sin olvidar que, en último término, la decisión final queda en manos del juez, cuya tradicional forma de proceder sigue teniendo un importante peso en sus decisiones. Sin restar importancia a ninguna de estas cuestiones y, teniendo en cuenta los objetivos de este trabajo, en las próximas líneas se analizarán dos de ellas.

A) La existencia de sesgos en los jueces

Los psicólogos Kahneman y Tversky llevaron a cabo, en la década de los 70, una serie de estudios empíricos que demostraron que las valoraciones de las personas sobre la probabilidad de que se produzcan resultados inciertos incurren en errores cognitivos sistemáticos (Gallo, 2011). En el ámbito del Derecho, estas investigaciones han dado origen a una amplia literatura científica que pone de manifiesto la existencia de signos inequívocos de sesgo en las decisiones de los jueces (Fariña et al., 2002). Un sesgo cognitivo es una desviación sistemática, involuntaria e inconsciente de una norma o de un estándar de racionalidad al emitir un juicio perceptual o conceptual, al recordar un evento o al hacer una predicción. Es decir, no se trata de un simple error, sino de comportamientos que ocurren consistentemente bajo circunstancias similares y que, por lo tanto, son predecibles y replicables. Estos sesgos son causados, principalmente, por el uso de heurísticos, que son atajos que utilizamos en el procesamiento de información. Pero también son producto de las limitaciones de nuestra capacidad cerebral de procesar información, de influencias emocionales, morales y sociales, y de distorsiones en el proceso de almacenamiento y búsqueda de información en la memoria (Kahneman, Slovic, y Tversky, 1982).

Desde el inicio de un proceso hasta el momento en que el juez debe decidir qué pena imponer, las heurísticas que utilizan de forma inconsciente producen múltiples sesgos que interfieren negativamente en sus razonamientos y decisiones. No se trata simplemente se sesgos raciales, sexuales o religiosos, sino de sesgos implícitos propios del procesamiento de la información, tanto sensorial como conceptual (Guthrie et al., 2001). Así, en el ámbito de las decisiones jurídico-penales, los sesgos más discutidos por la literatura científica de los últimos años son: el de anclaje (Kahneman y Tversky, 1974/1986; Fariña et al., 2002; Gallo, 2006, 2011; Múñoz Aranguren, 2011), el de representatividad (Kahneman y Tversky, 1972, 1973; Gallo, 2006, 2011; Múñoz Aranguren, 2011), el de disponibilidad (Kahneman y Tversky, 1973; Gallo, 2006, 2011; Múñoz Aranguren,

IUS ET SCIENTIA • 2021
Vol. 7 • N° 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

2011), el retrospectivo (Fischhoff, 1975; Gallo, 2006, 2011; Múñoz Aranguren, 2011); el de confirmación (Wason, 1960; Myers y Lam, 1976); y el de grupo (Múñoz Aranguren, 2011). Junto a estos sesgos, el denominado efecto marco (Kahneman y Tversky, 1979; Gallo, 2011) también ha sido objeto de estudio en el ámbito del Derecho Penal.

La existencia de tales sesgos explica que en numerosas ocasiones la opinión pública ponga en duda la imparcialidad de los jueces y el resto de los operadores jurídicos al tomar sus decisiones. De hecho, el impacto que esta forma de proceder de la mente humana tiene en las decisiones jurídico-penales se aprecia, aún más, con la llegada del EBS, pues los estudios realizados hasta el momento (y que serán analizados en mayor profundidad a continuación) reflejan cómo los jueces estadounidenses no están teniendo en cuenta los resultados arrojados por las herramientas (Stevenson y Doleac, 2018; Stevenson y Doleac, 2019; Terranova et al., 2020; Garret y Monahan, 2020). Consecuentemente, en lugar de conseguir que las decisiones judiciales sean más objetivas y precisas gracias a la información aportada por las HVR, los jueces siguen decidiendo en base a su particular experiencia y a sus propios criterios que, en ningún caso, están libres de sesgos.

B) La "desconocida" validez predictiva de las HVR

La segunda problemática que interesa destacar en este trabajo viene referida a la forma de conocer la validez predictiva de estas herramientas; cuestión para nada baladí si se tiene en cuenta que la capacidad predictiva de una HVR puede ser expresada con muchos indicadores diferentes cada uno de los cuales mide una dimensión distinta de esta (Loinaz, 2017). De esta forma, algunos expresan cómo de bien detecta el instrumento la reincidencia (sensibilidad) y la no reincidencia (eficacia), y otros cómo de bien la predice (valor predictivo positivo) o no la predice (valor predictivo negativo); algunos son medidas de riesgo relativo (esto es, informan sobre el mayor riesgo de reincidir de los clasificados como riesgo alto en relación con clasificados como de riesgo bajo), y otros de riesgo absoluto (es decir, informan sobre la probabilidad de que la reincidencia ocurra en un grupo concreto al transcurrir un periodo de tiempo dado), etc.

Dado que cada indicador mide una dimensión diferente de la capacidad predictiva, cada uno puede adoptar valores muy distintos para una misma HVR y podría darse el caso de que una misma herramienta sea calificada de "buena" (en el sentido de fiable, precisa) y "mala" a la vez (Martínez Garay y Montes Suay, 2018). No obstante, en la mayoría de los casos, los jueces no poseen esta información, lo cual origina dos posibles consecuencias. En primer lugar, que los jueces interpreten erróneamente los resultados arrojados por la HVR y apliquen medidas restrictivas de derechos fundamentales a sujetos que, en realidad, no las merecen (Martínez Garay y Montes Suay, 2018). En segundo lugar, que esta falta de información sobre cómo interpretar los resultados se convierta en una motivación más para no tener en cuenta la información que proporcionan y seguir decidiendo en base a sus propios criterios, con la consiguiente influencia de los sesgos y heurísticos presentes en ellos.

IUS ET SCIENTIA • 2021
Vol. 7 • N° 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

En definitiva, pese a que con un uso adecuado de estas herramientas el sistema de justicia penal podría poner fin a muchos de sus actuales problemas, la falta de conocimiento sobre su forma de proceder y sobre cómo interpretar sus resultados dificultan en gran medida su efectiva utilización. Por otro lado, gran parte de la sociedad también comienza a mostrar su rechazo hacia estas prácticas a causa de las posibles discriminaciones raciales y vulneraciones de derechos antes mencionadas. Se nos presenta, entonces, la siguiente pregunta: ¿confían realmente los seres humanos en las predicciones y decisiones algorítmicas?

III. Percepciones ciudadanas en torno a las decisiones algorítmicas, ¿qué sabemos?

En 1954, en su libro *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence*, Paul Meehl revisó los resultados de 20 estudios sobre predicciones en diversos dominios y fue el primero en mostrar que las predicciones realizadas por algoritmos superaban a las realizadas por humanos. En efecto, los algoritmos pueden permitir una toma de decisiones eficiente, optimizada y basada en datos (Kyung Lee, 2018). Esta visión optimista ha impulsado la adopción de algoritmos para la toma de decisiones en multitud de ámbitos de nuestra sociedad, como en el campo de los negocios (Siegel, 2016), en la salud (Jee et al., 2013), en la educación (Selwyn, 2015; Baker, 2016), en la política (Kim et al., 2014) y, poco a poco, también en la justicia y la organización penitenciaria (Botnick, 2015; Kehl et al., 2017; Stevenson y Doleac, 2018, 2019; Scurich y Krauss, 2019; McKay, 2019; De Michele et al., 2019; Garret y Monahan, 2020; Terranova et al., 2020).

La irrupción de este fenómeno está trasformando radicalmente el funcionamiento tradicional de nuestra sociedad y, aunque las ventajas son innegables, también los son las consecuencias negativas derivadas del uso de esta tecnología. Todo ello ha despertado el interés de sociólogos, psicólogos e investigadores de otros campos del conocimiento y está dando lugar a una importante literatura científica centrada en conocer cuáles son las percepciones que las personas tenemos sobre la utilización de estos algoritmos y la consiguiente automatización de decisiones.

1. Aversión algorítmica

Al ser humano siempre le ha costado admitir que una máquina pueda incorporar capacidades mentales en el más amplio sentido de la expresión (Amador, 1996). Por ello, varios estudios se han propuesto analizar si realmente existe o no esta desconfianza.

Uno de los primeros trabajos realizados sobre el tema tuvo lugar en el campo de los medios de comunicación y tenía por objeto conocer si las percepciones de los lectores de un periódico digital variaban en función del conocimiento que tuvieran sobre quién o qué había decidido qué tipo de noticias mostrarle. A los participantes se le ofrecieron seis noticias distintas que tuvieron que leer y se establecieron cuatro condiciones experimentales

IUS ET SCIENTIA • 2021
Vol. 7 • N° 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

en torno a quién había seleccionado las noticias: (a) las noticias habían sido seleccionadas por los redactores del periódico; (b) las noticias habían sido seleccionadas por un algoritmo de IA; (c) las noticias habían sido seleccionadas por otros usuarios lectores del periódico; (d) al último grupo se le asignó una tarea de pseudo selección que les llevó a creer que las noticias fueron elegidas por ellos mismos. Las opiniones de los participantes sobre el periódico fueron más negativas en el grupo que había sido informado de que las noticias eran seleccionadas por un sistema artificialmente inteligente. Así, una de las conclusiones a la que llegaron los autores fue que las percepciones que las personas tienen de una decisión se ven afectadas por el hecho de que tal decisión haya sido tomada por un algoritmo más que por una persona, con independencia de la calidad de sus resultados (Sundar y Nass, 2001).

Más recientes en el tiempo son los estudios que se han encargado de evaluar cómo las personas "castigan" más a un algoritmo después de verlo errar, mientras que con sus propios errores o los de otro ser humano se muestran mucho más tolerantes. Es decir, después de conocer que un algoritmo se ha equivocado o que presenta cierto margen de error, automáticamente, las personas confían más en el juicio humano, incluso cuando el rendimiento general de aquel es mejor que el de este. En Dietvorst et al. (2015), a través de cinco estudios con diferentes manipulaciones, los autores aportaron evidencia sobre cómo ver a un algoritmo errar hace que las personas sean menos propensas a confiar en sus predicciones. Esto ocurría incluso en los supuestos en los que los participantes también veían errar al ser humano o en los que, pese a que algoritmo no hacía la predicción exacta, sí que superaba a la del ser humano. Además, esta preferencia por la predicción humana frente a la algorítmica se daba sin importar si el pronosticador humano era el propio participante u otro participante anónimo.

De las aportaciones de estos estudios nace el concepto de aversión algorítmica (*algorithm aversion*) que hace referencia a la desconfianza humana en el uso de algoritmos para la toma de decisiones. No obstante, este fenómeno no es ni exclusivo ni excluyente.

2. Apreciación algorítmica

Dentro de la literatura sobre automatización de decisiones, también encontramos otro grupo de estudios cuyos resultados reflejan que, para cierto tipo de decisiones y en determinados ámbitos, las personas se adhieren más a recomendaciones algorítmicas frente a las procedentes de seres humanos. Así, en Araujo et al. (2019), se trató de analizar si las percepciones de utilidad, equidad y riesgo que las personas tienen sobre una decisión tomada por un sistema de IA en comparación con un ser humano experto en la materia variaban en función del contexto y de las consecuencias derivadas de la misma. En el estudio, los tres ámbitos analizados fueron mass media, salud y justicia; y diferenciaron entre nivel de impacto bajo o alto. Los resultados mostraron que: (a) las decisiones del algoritmo se percibieron como más justas en el ámbito de la salud y de la justicia, y sin diferencias en relación con los mass media; (b) los participantes percibieron en todos los contextos que en las decisiones del algoritmo el riesgo era menor,

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

especialmente, cuando las consecuencias eran de mayor impacto; y (c) las decisiones algorítmicas se percibieron como mucho más útiles en el ámbito de la salud, mientras que en los demás contextos no se vieron diferencias.

En esta misma línea, en Logg et al. (2019), los resultados de seis experimentos mostraron que las personas legas en conocimientos informáticos se adhieren más a las recomendaciones procedentes de un algoritmo frente a las procedentes de una persona. En el estudio, se pedía a los sujetos que hicieran predicciones numéricas sobre un estímulo visual y sobre la popularidad de una serie de canciones. Para ello, se les ofreció recomendaciones procedentes de un algoritmo y otras realizadas por un ser humano, y los resultados mostraron que los participantes confiaron más en las estimaciones algorítmicas. Sólo disminuyó la apreciación algorítmica cuando los participantes tenían que elegir entre la estimación de un algoritmo y la suya propia, prefiriendo en estos casos sus propias predicciones.

De conformidad con lo anterior, surge el concepto de aceptación algorítmica (*algorithm appreciation*) para referirse a aquellos supuestos en los que las personas perciben de forma favorable la automatización de decisiones.

3. Otros factores influyentes

A la evidencia sobre la aversión y aceptación algorítmica le han acompañado otras investigaciones centradas en analizar de forma más concreta qué elementos o factores hacen que los individuos se muestren favorables o adversos a la automatización de decisiones. Así, en un contexto de gestión de tareas, el estudio realizado por Kyung Lee (2018) sugiere que las características de las actividades afectadas por la decisión son importantes para comprender las experiencias de las personas con las tecnologías algorítmicas. En él, las personas percibieron que determinadas tareas requieren habilidades "humanas" (como sería el caso de realizar juicios subjetivos o de decisiones que requieren una mínima capacidad emocional), mientras que otras exigen habilidades más "mecánicas" (por ejemplo, el procesamiento de datos cuantitativos para medidas objetivas o la asignación de tareas en una empresa). Para las primeras, los sujetos participantes eligieron que la decisión fuera tomada por un ser humano, mientras que para las segundas les era indiferente que la decisión la tomara un ser humano o un algoritmo.

Otro grupo de estudios se ha centrado en conocer cómo debe configurarse un algoritmo para que las personas consideren sus decisiones "justas". En todos ellos, los participantes se mostraron desfavorables al uso de decisiones algorítmicas si los datos en los que se basa el algoritmo pueden originar discriminaciones por razón de raza o sexo. Por ejemplo, en Pierson et al. (2018), los participantes se mostraron desfavorables a que el género sea un factor a tener en cuenta por un algoritmo que recomienda asignaturas optativas a estudiantes si su inclusión hace que las estudiantes de género femenino sean menos propensas a participar en cursos de ciencias naturales. Y, en Saxena et al. (2020), los sujetos encuestados se mostraron en contra de que la raza fuera tenida en cuenta por un algoritmo que en una entidad financiera se encarga de decir a qué sujetos conceder un préstamo y a cuáles no.

IUS ET SCIENTIA • 2021
Vol. 7 • N° 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

Como una primera conclusión, se advierte que nos encontramos ante un fenómeno muy poliédrico, ya que no existen factores que determinen de forma absoluta la existencia de una mayor o menor aceptación sobre el uso de estos algoritmos. El ámbito en el que se toma la decisión, las consecuencias que se puedan derivar de dicha decisión, ver errar al algoritmo, o los datos en los que éste se basa para decidir, son sólo ejemplos de algunos aspectos influyentes, pero no exclusivos ni generalizables. Además, la literatura existente es escasa en general y lo es, mucho más, en el ámbito de la justicia penal, pese a ser éste un campo que empieza a verse afectado por la automatización de decisiones con la llegada de las HVR.

Ante este panorama, el estudio de las percepciones ciudadanas sobre el uso de predicciones y decisiones algorítmicas en la practica jurídico-penal se reviste necesario por dos razones. En primer lugar, porque es el juez quién, siendo también un ciudadano, va a decidir si tener en cuenta o no los resultados arrojados por la HVR y, a su vez, esta decisión se verá afectada por la percepción que dicho juez tenga de este instrumento. En segundo lugar, porque los destinarios de estas decisiones van a ser los miembros de la sociedad y, dado que pueden ver afectados algunos de sus derechos, es necesario que se respete mínimamente la opinión que tengan al respecto; de lo contrario, incluso podría comenzar a verse a afectada la legitimidad del sistema. En definitiva y como se anunciaba al principio de este trabajo, la importancia de conocer tales percepciones reside en el hecho de que la automatización de las decisiones judiciales sólo triunfará y será legítima si es aceptada por la ciudadanía a la que se dirige.

IV. Importancia de las percepciones sociales sobre las decisiones algorítmicas en el ámbito de la justicia penal

1. El juez

El sistema de justicia penal estadounidense arrastra, desde el siglo pasado, una serie de problemas tanto a nivel social como económico. Con su aparición, el EBS se presentó como la solución perfecta a todos ellos, incorporándose progresivamente en más de la mitad de sus estados. Se pensó, entonces, que las tasas de encarcelamiento se verían reducidas y que la coherencia y transparencia en la toma de decisiones jurídico-penales aumentaría (Botnick, 2015; Garrett y Monahan, 2018; Scurich y Krauss, 2019). No obstante, se anunciaba unas líneas más arriba que estas reformas no están teniendo el impacto que se esperaba.

En los tribunales estadounidenses la incorporación del EBS realmente no ha privado a los jueces de su margen de discreción y no lo ha hecho porque la decisión final queda siempre en manos del juez. Por tanto, independientemente del nivel de riesgo que haya arrojado el instrumento, tales predicciones no son vinculantes para ellos. A este hecho debemos sumarle otros dos condicionantes: por un lado, la presencia de los sesgos y heurísticos antes mencionada; y, por otro, la existencia de prioridades políticas propias que tienen los jueces (como aspiraciones profesionales de reelección o nuevos

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

nombramientos, opinión pública, etc.) y que les hacen ignorar la evaluación de riesgos por completo o responder a ella estratégicamente, usándola para avanzar en su propia agenda (Cowgill y Stevenson, 2019; Lim et al., 2015; Berdej y Yuchtman, 2013; en Stevenson y Doleac, 2019). De esta forma, los trabajos que han tratado de conocer cómo se está articulando la relación entre los jueces y las predicciones algorítmicas arrojan interesantes resultados.

Estudios llevados a cabo en Kentucky y Virginia (Stevenson y Doleac, 2019; Garret y Monahan, 2020) ofrecen evidencias sobre cómo en estos estados, tras adoptar evaluaciones de riesgo algorítmicas con el objetivo declarado de reducir el encarcelamiento de los infractores de bajo riesgo, las tasas de internamiento penitenciario no disminuyeron sustancialmente. En el primer estado, Kentucky, los resultados mostraron que dos de cada tres veces los jueces no tienen en cuenta el nivel de riesgo que predice el algoritmo y, en lugar de conceder la libertad sin fianza a aquellos sujetos que presentan un riesgo bajo o moderado (que es el objetivo perseguido con la reforma), siguen decretando su entrada en prisión preventiva. En lo que respecta al estado de Virginia, la reforma persigue reducir la condena de aquellos procesados que presenten bajo riesgo, pero de nuevo, tal objetivo no se está viendo cumplido, pues en todos los casos analizados la mayor parte de los jueces no tiene en cuenta las recomendaciones de la herramienta y continúan condenando con penas elevadas. La misma desconfianza hacia las predicciones algorítmicas realizadas por las HVR se ha observado en el estado de Missouri (Botnick, 2015) y en Nueva Jersey (Garrett y Monahan, 2020).

Por otro lado, en Stevenson y Doleac (2019) se observó cómo los jueces otorgan sistemáticamente indulgencia a los jóvenes acusados a pesar de su alto riesgo de reincidencia. Este alto riesgo de reincidencia se debe a que, la evidencia criminológica demuestra que la corta edad es uno de los principales factores de riesgo asociados a la reincidencia delictiva y, por ello, las HVR otorgan mayor importancia a la variable "edad" cuanto más joven sea el sujeto. Sin embargo, este hecho se opone a una práctica judicial estadounidense según la cual la menor edad del delincuente debe considerarse atenuante en la sentencia debido a la menor percepción de culpabilidad y a las consecuencias negativas que para estos jóvenes puede ocasionar el paso por prisión. Esta práctica está muy arraigada entre los jueces y magistrados norteamericanos, lo cual explica que decidan ignorar las predicciones algorítmicas en estos casos y absuelvan a los delincuentes más jóvenes sin tener en cuenta el riesgo que este hecho puede suponer para la sociedad.

En tercer lugar, en Skeem et al. (2020) se observa cómo las decisiones judiciales se ven influenciadas por las características sociodemográficas de los acusados. En él, jueces del este, medio oeste y suroeste de Estados Unidos con experiencia en sentencias penales participaron en un experimento controlado para conocer cómo interactúan las decisiones de los jueces con el porcentaje de reincidencia arrojado por la herramienta y con las características de la persona procesada. Los resultados revelaron que la información proporcionada por la herramienta redujo la probabilidad de encarcelamiento de los acusados relativamente ricos, pero la misma información aumentó la probabilidad

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

de encarcelamiento de los acusados relativamente pobres. Es decir, que incluso cuando los jueces están teniendo en cuenta los resultados arrojados por las HVR, la decisión final sigue estando influenciada por sus propios sesgos.

En definitiva, los estudios realizados hasta el momento ponen de manifiesto la existencia de una percepción negativa hacia el uso de estas herramientas por parte de los jueces quienes, haciendo uso de su margen de discrecionalidad, prefieren continuar tomando sus decisiones en base a sus propios criterios. Esto explica que las tasas de encarcelamiento sigan sin verse reducidas, que el gasto estatal destinado al mantenimiento de la población reclusa continúe creciendo y que el racismo, la xenofobia y la aporofobia sigan reflejados en muchas sentencias. La ciudadanía en su conjunto acaba siendo, una vez más, la que resulta mayormente perjudicada al ser sus miembros los principales destinatarios de estas decisiones. Por ello, para quien suscribe estas páginas, igual atención debe prestarse a las percepciones que tales ciudadanos tienen sobre este tipo de cambios en el sistema de justicia, porque si queremos que esta clase de reformas sean socialmente aceptadas es necesario que respeten de alguna forma las intuiciones de justicia de la ciudadanía a la que van dirigidas.

2. La ciudadanía: el caso de la California Proposition 25

En la actualidad, no existen estudios que se hayan centrado en conocer de forma específica las percepciones que la ciudadanía tiene sobre la incorporación de herramientas de IA en el ámbito de la justicia penal. Pese a ello, se advierte que la llegada de tales predicciones algorítmicas al ámbito de la justicia penal ha originado entre algunos miembros de la población una desconfianza generalizada debido a las implicaciones éticas y a la posible vulneración de derechos que parece derivarse de su uso. Por un lado, algunos de sus críticos alegan que muchas de las variables que tiene en cuenta el algoritmo a la hora de evaluar a un sujeto (como el lugar de residencia, el nivel socioeconómico, el número de arrestos previos, etc.), no son sino una forma más de perpetuar prejuicios y sesgos ya existentes. Por otro lado, también destacan el hecho de que algunas HVR son desarrolladas por empresas privadas que contratan las jurisdicciones locales para proporcionar sus algoritmos patentados, impidiendo a los acusados impugnar las evaluaciones que brindan (tal y como ocurrió en el caso Loomis). Finalmente, algunos profesionales del Derecho y la Criminología también muestran cierto recelo apoyándose en los problemas relativos a su validez predictiva (a los que se ha hecho referencia unas líneas más arriba) y a la falta de información sobre su utilización.

Esta desconfianza no debe ser obviada si tenemos en cuenta que son varios los autores que desde hace años defienden que la percepción pública de la legitimidad y equidad de las instituciones y de la toma de decisiones legales afecta al cumplimiento y respeto de estas, así como al grado de compromiso que los ciudadanos tienen con el sistema de justicia penal (Robinson, 1995, 2000, 2007, 2013; Tyler & Jackson, 2014). Es

IUS ET SCIENTIA • 2021
Vol. 7 • N° 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

decir, un sistema de justicia penal que es percibido como justo por la comunidad a la que rige gana en su capacidad de obtener respeto y conformidad; mientras que si es visto como dispuesto a cometer injusticias y a tolerar cómodamente errores de la justicia pierde su credibilidad moral con la comunidad y es más proclive a provocar resistencia y subversión. Precisamente, lo ocurrido en el estado de California con la denominada *Proposal 25* es un claro ejemplo de la importancia que tiene conocer y atender a las actitudes de los miembros de la sociedad hacia el uso de estas herramientas (Gasek, J., 2019; Scurich y Krauss, 2019).

La legislación tradicional californiana establecía que, cuando una persona es detenida como sospechosa de un delito, su puesta en libertad hasta el momento de celebración del juicio queda subordinada al pago de una fianza. Sin embargo, durante muchos años se ha criticado que esta medida perjudicaba a los ciudadanos más pobres que no podían permitirse el pago de la fianza. Esto llevó a que el gobernador Jerry Brown promulgara en agosto de 2018 una nueva ley (*the Proposition 25*) en virtud de la cual el sistema de fianza era sustituido por una evaluación del sujeto con una HVR. Es decir, que en la Audiencia inicial (que se celebra dentro de las 48h siguientes a la detención) será el Juez quién decida, en función del riesgo que arroje la HRV, si el sujeto queda en libertad o si debe quedar bajo prisión provisional hasta la celebración del Juicio penal.

La nueva regulación tenía previsto entrar en vigor el 1 de octubre de 2019. Sin embargo, muchos miembros de la sociedad comenzaron a oponerse a esta reforma alegando que permitiría a los jueces encarcelar a más personas y que no incluía suficiente supervisión sobre las HVR, lo cual suponía una amenaza para las garantías y los derechos fundamentales de los ciudadanos. La preocupación que originó esta reforma legal dio lugar a la creación de *The American Bail Coalition*, una asociación que comenzó a recabar firmas para que, tal y como recoge la Constitución de California en el artículo II de su Sección 9, se pudiera llevar a cabo la celebración de un referéndum que permitiera a los ciudadanos elegir si esta ley debía o no entrar en vigor. En enero de 2019 ya se habían alcanzado las firmas necesarias para su celebración y, por tanto, la entrada en vigor de la ley quedó suspendida hasta que el referéndum tuviera lugar a finales del año siguiente. El 3 de noviembre de 2020 se procedió a la celebración del referéndum en todo el estado y un 56.41% de los californianos votaron en contra de la entrada en vigor de la ley. De esta forma, los ciudadanos dijeron "no" a la utilización de predicciones algorítmicas para decidir si un sujeto debe quedar o no en libertad condicional.

En definitiva, dado que el sistema de justicia penal debe o debería ser acorde a las intuiciones de justicia de la sociedad, considero relevante que las actitudes de la ciudadanía hacia el uso de estas herramientas deben ser conocidas y, en base a ello, determinar si se debe imponer su uso o no en los Juzgados y Tribunales. Es este, pues, el principal objetivo de este trabajo: analizar y evaluar el grado de aceptación ciudadana del uso de herramientas basadas en algoritmos de IA para la toma de decisiones jurídico-penales. Para lograr este objetivo se llevará a cabo el estudio que se presenta a continuación.

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

V. Estudio empírico

1. Objetivos e hipótesis

El objetivo general de este estudio es evaluar el grado de aceptación ciudadana del uso de herramientas de IA para la toma de decisiones jurídico-penales, concretado en este caso en el uso de la herramienta RisCanvi. Para alcanzar este objetivo general, se proponen los siguientes objetivos específicos:

1. Analizar la influencia del conocimiento sobre la validez predictiva de RisCanvi en la aceptación ciudadana de su uso.
2. Evaluar la influencia del nivel de riesgo de reincidencia arrojado por RisCanvi en la aceptación ciudadana de su uso.
3. Conocer si el grado de aceptación ciudadana del uso de RisCanvi se relaciona con las percepciones que los participantes tienen sobre la existencia de imparcialidad en los jueces.

Para la consecución de estos objetivos se plantean las siguientes hipótesis:

a) Conocimiento sobre la validez predictiva. De conformidad con la literatura previa, los sujetos tienden a rechazar las decisiones tomadas por un algoritmo después de verlo errar o de tener información sobre la validez predictiva de dicho algoritmo (Dietvorst et al., 2015). En este sentido, se hipotetiza que:

(H1) Los participantes que han recibido información sobre la validez predictiva de la herramienta presentarán un grado de aceptación menor que los participantes que no han recibido tal información.

b) Nivel de riesgo arrojado. Varios estudios empíricos muestran que en algunos supuestos en los que los jueces disponen de la información sobre el riesgo de reincidencia arrojado por una HVR, esta información no es tenida en cuenta para la toma de su decisión judicial y acaban aplicando su propio criterio (Stevenson y Doleac, 2019; Garret y Monahan, 2020; Botnick, 2015). De conformidad con ello, se hipotetiza que:

(H2) Los participantes tendrán diferente grado de aceptación en función del nivel de riesgo que arroje la herramienta.

c) Percepción de imparcialidad. Tal y como han mostrado algunos estudios, una variable que puede influir en la aceptación ciudadana del uso de algoritmos para la toma de decisiones es la percepción de imparcialidad y racionalidad que las personas tienen de estas herramientas (Dijkstra et al., 1998). Teniendo en cuenta esto, se hipotetiza que:

(H3) Los participantes que evalúan que las decisiones de los jueces no son imparciales, mostrarán un mayor grado de aceptación de la herramienta RisCanvi.

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

2. Muestra

La muestra estuvo compuesta por 359 participantes, de los cuales el 56% son mujeres y 44% hombres con una edad media de 36 años. En relación con el nivel de estudios de los individuos de la muestra, este se distribuye de la siguiente manera: 0,8% educación primaria; 0,8% educación secundaria obligatoria, 7,8% bachillerato; 7,2% formación profesional; 54,3% están graduados o licenciados; 20,6% tienen estudios de máster; y 8,4% han obtenido el Doctorado. Además, un 47,9% poseen estudios en Derecho. La última característica que se quiso conocer de los participantes fue su ideología política, para lo cual se utilizó una escala de 1 a 7, donde 1 = extrema izquierda y 7 = extrema derecha. El 47,4% de la muestra se situó en el espectro de la izquierda, un 28,4% en el centro y el 24,2% restante en el espectro de la derecha.

3. Variables e instrumento

Además de las variables sociodemográficas anteriormente mencionadas, en este estudio se han tenido en cuenta como variable dependiente el grado de aceptación del uso de la IA para la toma de decisiones en el ámbito penal. Esta aceptación se ha medido tanto a nivel general como sobre un caso específico. También se consideraron como variables dependientes la confiabilidad percibida de los resultados de la herramienta en un caso en concreto y la percepción de imparcialidad de los jueces. Por otro lado, como variables independientes se han tenido en cuenta el nivel de riesgo de reincidencia arrojado por la herramienta y la información sobre la validez predictiva de la herramienta RisCanvi. Para medir estas variables independientes se llevó a cabo una manipulación. Por último, previendo que pudiera afectar al objeto de estudio, también se controló la variable conocimiento sobre RisCanvi. En el ANEXO I se encuentran detalladas todas las variables, su operativización y sus correspondientes escalas de medición.

Como instrumento de medición se empleó un cuestionario construido ad hoc resultado de diversas reuniones con expertos en Derecho penal y metodología.

4. Diseño y procedimiento

Para la consecución de los objetivos propuestos y la comprobación de las hipótesis formuladas, se empleó un diseño experimental en el que los participantes fueron asignados aleatoriamente a cada una de las condiciones experimentales. Se empleó un diseño factorial 2x2 en el que se tuvieron en cuenta dos grados de riesgo de reincidencia y dos niveles de información sobre la validez predictiva de la herramienta. Por último, se estableció un quinto grupo que no recibió ningún tipo de manipulación, configurándose como el grupo control.

Para llevar a cabo este diseño factorial se empleó la técnica del caso escenario (Hartley, 2004; Cousin, 2005; Swanborn, 2010). Dado que el objeto de estudio es el nivel de aceptación ciudadana del uso de decisiones algorítmicas en el ámbito de la justicia

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

penal, en primer lugar, se ofreció una introducción sobre la herramienta de valoración del riesgo RisCanvi a todos los participantes y que, en el caso de los grupos 2 y 4, también incluyó información sobre la validez predictiva de la herramienta. A continuación, a los grupos 1, 2, 3 y 4 se les puso en la situación de que un sujeto que se encontraba en prisión por un delito de agresión sexual del artículo 178 del Código Penal ya había cumplido la mayor parte de su condena y, por ello, el juez debía decidir si podía quedar en libertad condicional o no. Junto con esta información, en el caso escenario se indicaba el riesgo de reincidencia arrojado por la herramienta (bajo para los grupos 1 y 2, alto para los grupos 3 y 4). De esta forma, los 359 participantes fueron distribuidos aleatoriamente entre los cuatro grupos. El grupo control contó con 77 sujetos; el grupo 1 con 96; el 2 lo compusieron 63 participantes; el grupo 3, 62; y, finalmente, el cuarto grupo contó con 61 participantes.

Para la distribución del cuestionario se utilizó el sistema de encuestas gratuitas de Google Forms. Asimismo, el cuestionario fue distribuido a través de diversas redes sociales (Twitter, Facebook, Whatsapp, entre otras), por lo que se utilizó un muestreo no probabilístico por conveniencia. Sin embargo, para asegurar la validez interna del diseño, los participantes fueron signados aleatoriamente a cada una de las condiciones. Para garantizar tal asignación aleatoria se utilizó el programa Sublime Text que permitió unificar los cuatro enlaces de los cuestionarios generados en único enlace que asignaba aleatoriamente a cada sujeto. En este sentido, cabe señalar que cada participante solo tuvo acceso a el cuestionario que aleatoriamente se le había asignado. El cuestionario estuvo disponible de 29/06/2021 a 08/07/2021.

VI. Resultados

1. Conocimiento sobre la validez predictiva

La primera hipótesis planteada en este estudio era que *los participantes que han recibido información sobre la validez predictiva de la herramienta presentarán un grado de aceptación menor que los participantes que no han recibido tal información.* En el presente estudio, la variable aceptación se ha medido de cuatro formas distintas y, dado que nos encontramos ante una comparación de más de tres grupos en la que las variables no son cuantitativas, para conocer si existen diferencias en el nivel de aceptación entre cada uno de los grupos se aplicó el estadístico de contraste *Kruskal-Wallis* (tabla 1).

La prueba Kruskall-Wallis refleja que el nivel de aceptación difiere significativamente entre los cuatro grupos sólo cuando se mide de forma específica a través del ítem *¿En qué medida le parece aceptable que el juez base su decisión de si poner en libertad o no al sujeto en el riesgo arrojado por RisCanvi?* (χ^2 =12,698; p = ,005). No obstante, estos resultados sólo informan de que al menos dos grupos de entre los comparados son diferentes, pero no indica cuales. Para saberlo es necesario comparar los grupos entre ellos a través de la prueba *U de Mann-Whitney* en relación con la variable aceptación específica (I) (figura 1).

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

Tabla 1. Resultados de la prueba Kruskall-Wallis

	Aceptación general (I)	Aceptación general (II)	Aceptación específica (I)	Aceptación específica (II)
H de Kruskal-Wallis	4,118	2,442	12,698	0,646
gl	3	3	3	3
Sig. asin.	0,249	0,486	0,005	0,886

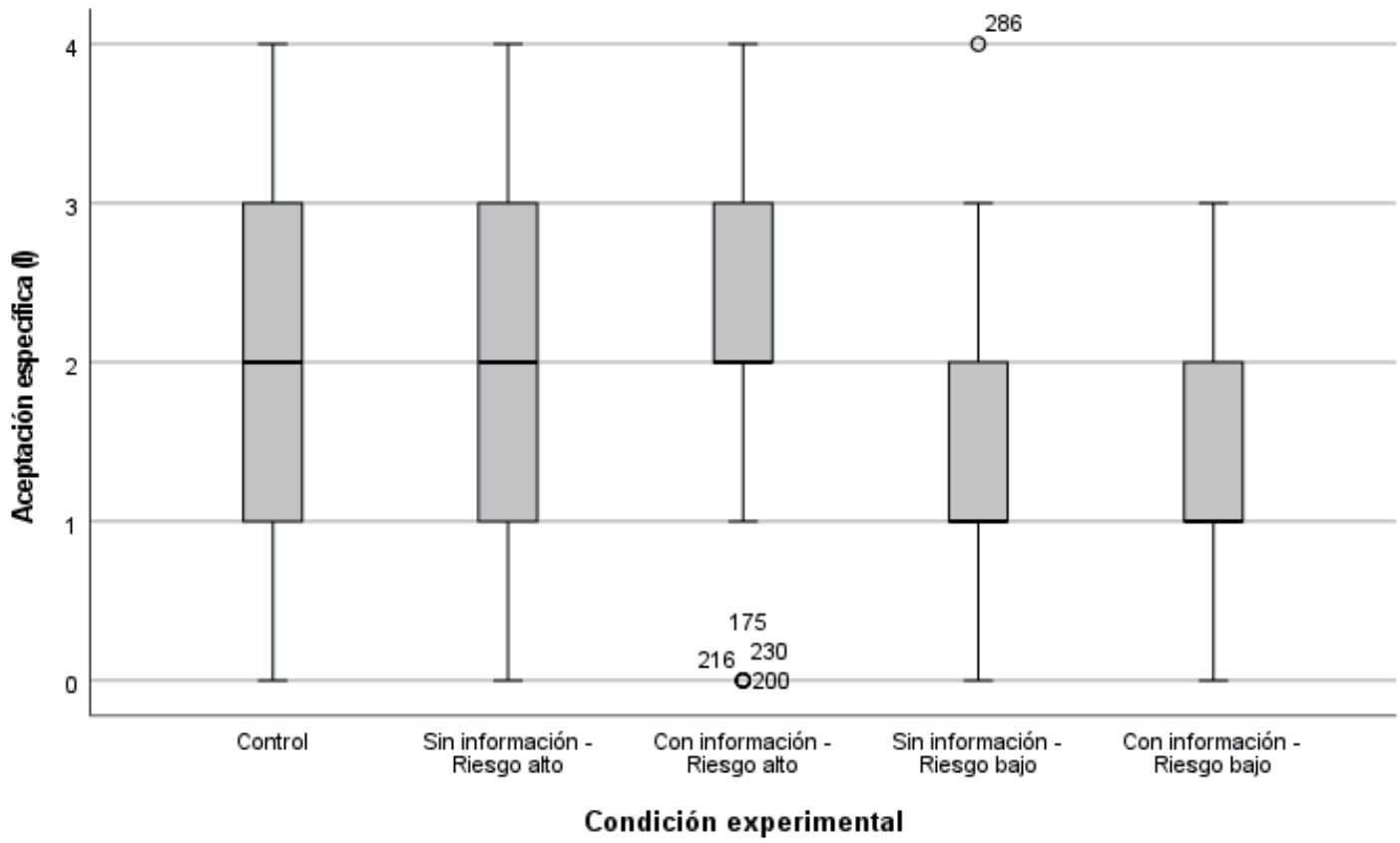

Figura 1. Diagrama de caja Aceptación específica (I) – Condición experimental. *Nada aceptable (0); poco aceptable (1); neutral (2); bastante aceptable (3); totalmente aceptable (4).

Tanto a nivel descriptivo como estadísticamente se aprecia que solamente existen diferencias significativas cuando se compara el grupo Con información – Riesgo alto con los grupos Sin información – Riesgo bajo (U = 1428,500; Z = -2,710; p = 0,007) y Con información – Riesgo bajo (U = 1261,000; Z = -3,448; p = <0,001). Siguiendo la clasificación elaborada por Cohen (1992), en la primera pareja el tamaño del efecto es pequeño (r = 0,24) y en la segunda, mediano (r = 0,03).

Por otro lado, si se realiza el contraste considerando como variable de agrupación *información sobre la validez predictiva de RisCanvi* (y, por tanto, no teniendo en cuenta el efecto de interacción entre las dos variables independientes), no se aprecian diferencias significativas entre ninguno de los grupos, sino que todos los grupos presentan un nivel de aceptación similar (U = 13973,500; Z = -0,666; p = 0,505).

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

Tabla 2. Resultados de la prueba U de Mann-Whitney (variable agrupación = información)

Condición	Rango	*U*	Z	*p*
Sin información	177,46	13973,500	-0,666	0,505
Con información	184,81			

De acuerdo con estos resultados, la H1 debe ser rechazada porque los participantes que han recibido información sobre la validez predictiva presentan un nivel de aceptación similar que aquellos participantes que no la han recibido o, incluso, mayor; pero en ningún caso, menor (que es lo que sostenía la H1).

2. Nivel de riesgo arrojado

En segundo lugar, se hipotetizó que *los participantes tendrán diferente grado de aceptación en función del nivel de riesgo que arroje la herramienta*. Para contrastar esta hipótesis basta con atender, de nuevo, a los resultados de la prueba *Kruskal-Wallis* antes realizada (véase Tabla 6) y, al igual que ha ocurrido para la H1, sólo cuando se mide la aceptación específica (I) las diferencias son significativas (χ^2 =12,698; p = ,005). Asimismo, al realizar el contraste por parejas de grupo con el estadístico *U de Mann-Whitney* (véase Tabla 8), los resultados son significativos cuando se compara el grupo Con información – Riesgo alto con los grupos Sin información – Riesgo bajo (U = 1428,500; Z = -2,710; p = 0,007) y Con información – Riesgo bajo (U = 1261,000; Z = -3,448; p = <0,001). De esta forma, en la primera pareja el tamaño del efecto es pequeño (r = 0,24) y en la segunda, mediano (r = 0,03).

Estos resultados indican que los participantes a los que se les informó de que el riesgo era alto presentan un mayor nivel de aceptación que aquellos a los que se les informó de que el riesgo era bajo (véase de nuevo figura 1). Además, si se realiza el contraste considerando como variable de agrupación el nivel de riesgo (lo que conlleva no tener en cuenta la interacción entre ambas variables independientes), las diferencias son claramente significativas (U =8010,500; Z = -2,719; p = 0,007). El tamaño del efecto es, no obstante, pequeño (r = 0,16).

Tabla 3. Resultados de la prueba U de Mann-Whitney (variable de agrupación = nivel de riesgo)

Condición	Rango	*U*	Z	*p*
Riesgo alto	152,62	8010,500	-2,719	0,007
Riesgo bajo	127,13			

Conforme a estos resultados, la predicción realizada en base a la H2 puede ser aceptada ya que los participantes presentan diferente grado de aceptación en función del nivel de riesgo arrojado por la herramienta.

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

3. Percepción de imparcialidad

La última hipótesis planteada en este estudio era que *los participantes que evalúan que las decisiones de los jueces no son imparciales, mostrarán un mayor grado de aceptación de la herramienta RisCanvi*. Una vez más, para conocer si existen diferencias entre los grupos el estadístico de contraste aplicado fue la prueba *Kruskal-Wallis*.

Tabla 4. Resultados de la prueba Kruskal-Wallis para la variable percepción imparcialidad

	Percepción imparcialidad jueces
H de Kruskal-Wallis	4,946
gl	4
Sig. asin.	0,293

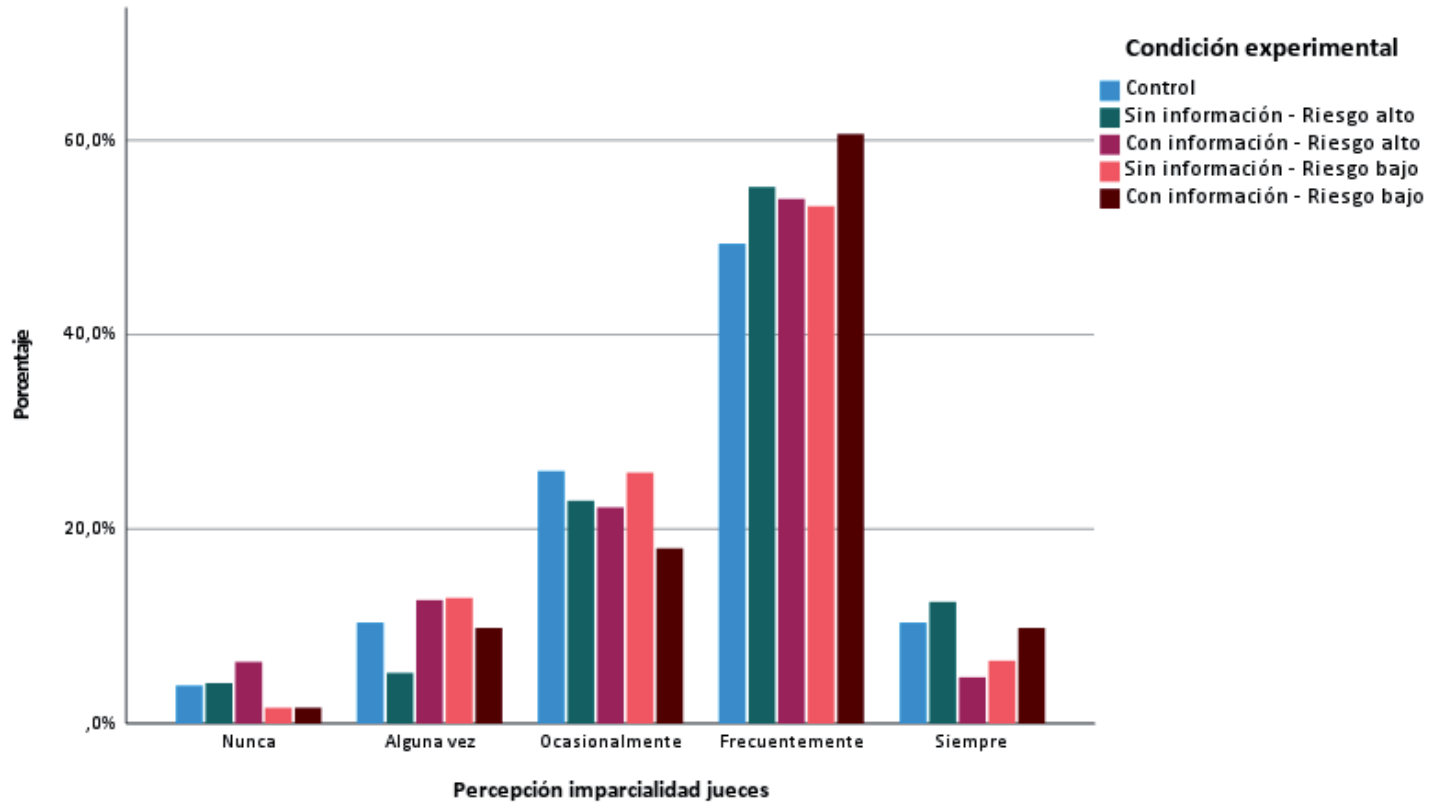

Figura 2. Frecuencias de los valores de la variable percepción imparcialidad

A nivel descriptivo se observa que, en todos los grupos, más de la mitad de los participantes consideran que los jueces toman sus decisiones de forma imparcial con frecuencia ($f_{i\,N}$ = 54,3%). Esto coincide con los resultados de la prueba Kruskal-Wallis en los que se aprecia que no existen diferencias significativas entre ninguno de los grupos (χ^2 = 4,96; p = 0,293). Al respecto, la H3 deber ser rechazada por completo porque la mayoría de los participantes considera que los jueces son frecuentemente imparciales cuando toman sus decisiones jurídico-penales.

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

VII. Discusión

El objetivo general de este estudio ha sido evaluar el grado de aceptación ciudadana del uso de herramientas de IA para la toma de decisiones jurídico-penales. Al respecto, los resultados reflejan que el grado de aceptación es relativamente bajo pues, con independencia de las diferencias que se hayan podido dar entre grupos, cuando se ha evaluado (a través de cuatro ítems distintos) tal aceptación, la gran mayoría de los participantes se han situado en el primer tramo de la escala o en el tramo central.

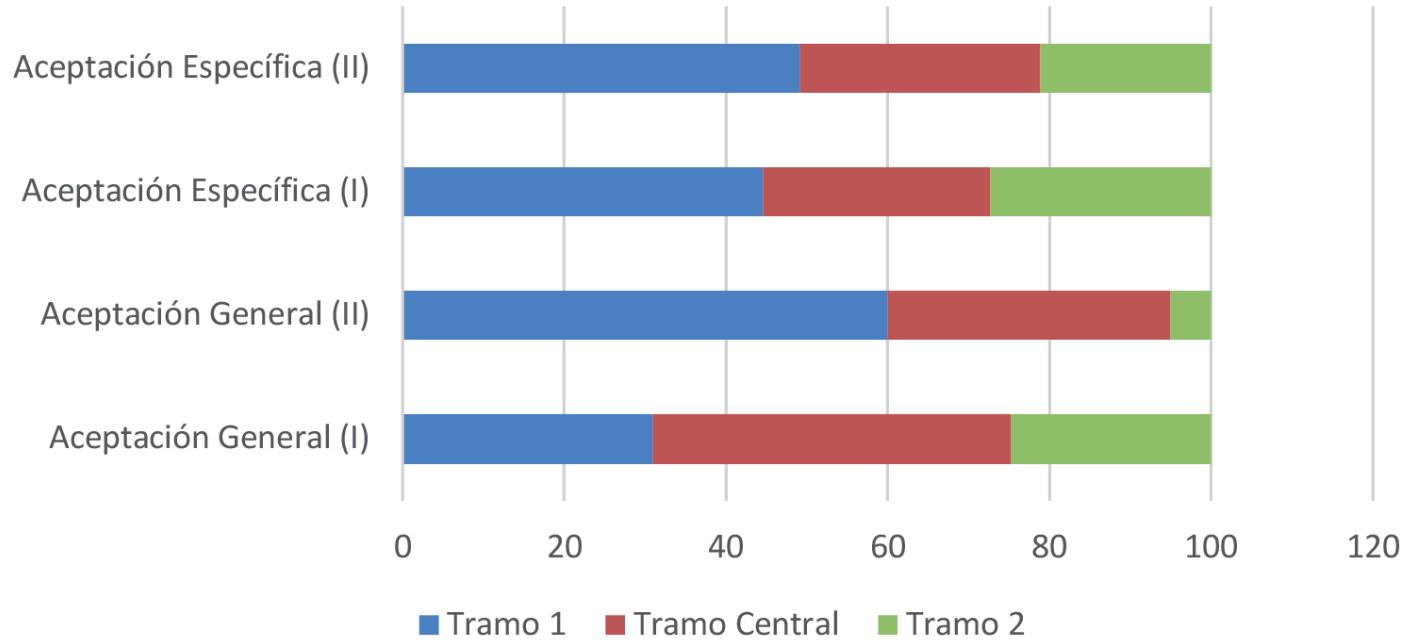

Figura 3. Escala de medida de las variables de aceptación divida en tramos

Estos resultados se encuentran en la misma dirección que la opinión pública que tienen los ciudadanos de California sobre el uso de estas herramientas en el sistema de justicia penal (Gasek, J., 2019; Scurich y Krauss, 2019), quienes han rechazado su incorporación en la toma de decisiones jurídico-penales. Coinciden, también, con los estudios llevados a cabo en otros estados norteamericanos en los que se midió la frecuencia con la que los jueces siguen las recomendaciones hechas por estos sistemas y cuyos resultados han puesto de manifiesto que no atienden a tales predicciones (Botnick, 2015; Stevenson y Doleac, 2019; Garret y Monahan, 2020). Atendiendo a los estudios realizados fuera del ámbito jurídico, los resultados de este trabajo son igualmente acordes con las conclusiones alcanzadas en Sundar y Nass (2001), donde las percepciones que las personas tenían de una decisión se vieron afectadas por el hecho de que tal decisión fuese tomada por un algoritmo más que por una persona, siendo el nivel de aceptación menor cuando la decisión procedía de un algoritmo. Asimismo, en Kyung Lee (2018) se observó cómo los participantes distinguen entre decisiones que requieren habilidades "humanas" (y, por tanto, prefirieron que las tomaran las personas), mientras que otras exigen habilidades más "mecánicas" (aceptando que sean tomadas por algoritmos). Esto

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

puede ser extrapolable al presente estudio, donde sus resultados indican que decidir sobre la libertad de un sujeto se considera una decisión más "humana" que "mecánica".

En relación con la influencia que pueda tener el hecho de poseer información sobre la validez predictiva de la herramienta en el nivel de aceptación ciudadana de su uso, se observa que los participantes que han recibido tal información presentan un nivel de aceptación similar que aquellos participantes que no la han recibido o, incluso, mayor. La primera de las hipótesis planteadas sostenía que el nivel de aceptación sería menor en aquellos sujetos que obtuvieran información sobre la validez predictiva por lo que, teniendo en cuenta los resultados arrojados, esta hipótesis no puede ser aceptada. Al respecto, este estudio difiere de las evidencias aportadas por Dietvorst et al. (2015) donde tener información sobre el margen de error del algoritmo hizo que los participantes fueran menos propensos a acepar sus predicciones.

Por otro lado, las diferencias han sido bastante llamativas en relación con el nivel de riesgo de reincidencia que arroja la herramienta. Así, los participantes a los que se les informó de que el riesgo era alto han presentado un mayor nivel de aceptación que aquellos a los que se les informó de que el riesgo era bajo. Estos resultados han permitido aceptar la predicción realizada en base a la segunda hipótesis ya que, efectivamente, los participantes han presentado diferente grado de aceptación en función del nivel de riesgo arrojado por la herramienta. Precisamente, en Araujo et al. (2019) se observó que las percepciones que las personas tienen sobre una decisión tomada por un sistema de IA en comparación con un ser humano experto en la materia variaban en función del contexto y de las consecuencias derivadas de la misma. En este sentido, podría explicarse que en el presente estudio el nivel de aceptación de la herramienta es mayor cuando esta arroja un nivel de riesgo de reincidencia alto, ya que las consecuencias derivadas de dejar en libertad a un sujeto con riesgo alto son potencialmente más graves que si el sujeto presenta un nivel de riesgo bajo.

En tercer lugar, no se ha observado diferencias en la percepción de imparcialidad sobre los jueces, pues la mayoría de los participantes considera que son frecuentemente imparciales cuando toman sus decisiones en el ámbito penal. Estos resultados nos obligan a rechazar por completo nuestra última hipótesis que predecía que los participantes que evaluasen que las decisiones de los jueces no son imparciales, mostrarían un mayor grado de aceptación del empleo de la herramienta.

Por último, para los cuatro grupos a lo que se les aplicó la manipulación, también se ha medido la confiabilidad que tienen de los resultados. Pese a que no se ha hipotetizado nada al respecto, la aplicación del estadístico *Kruskal-Wallis* sobre esta variable ha arrojado resultados significativos ($\chi^2 = 19{,}391$; $p = 0{,}001$), lo cual indica que existen diferencias entre los grupos. El contraste por parejas realizado mediante el estadístico *U de Mann-Whitney* ha permitido conocer cuáles son estos grupos y, al respecto, se observa que los participantes a los que se le informó de que el riesgo arrojado por la herramienta era bajo, presentan un menor nivel de confiabilidad en los resultados. Con todo, el valor de moda en esta variable ha sido "moderadamente confiable", que ocupa la posición central de la escala ($f_i = 50\%$). Este aspecto merece ser destacado porque indica que los participantes confían de forma notable en los resultados arrojados por la

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

herramienta, pero no se muestran favorables a su uso en la toma de decisiones jurídico-penales, como puede ser poner o no en libertad condicional a un sujeto.

Tabla 5. Contrastes de confiabilidad por parejas de grupos

Grupos*	*U*	Z	*p*
1 – 2	2651,500	-1,535	0,125
1 – 3	2476,000	-1,919	0,055
1 – 4	2242,500	-2,652	0,008
2 – 3	1340,500	-3,321	<0,001
2 – 4	1187,500	-3,996	<0,001
3 – 4	1754,000	-0,760	0,447

*Sin información – Riesgo alto (1); Con información – Riesgo alto (2); Sin información - Riesgo bajo (3); Con información – Riesgo bajo (4).

En relación con las limitaciones que presenta este trabajo, la principal de ellas viene referida a su novedad, pues se trata del primer estudio que se ha hecho en España sobre el nivel de aceptación del uso de herramientas dotadas de IA en el ámbito de la justicia penal y, a nivel internacional, tampoco son numerosos los trabajos que han abordado esta cuestión. Asimismo, sería conveniente poder disponer de más estudios que evalúen la validez predictiva de RisCanvi y del resto de herramientas de valoración del riesgo que ya están siendo utilizadas, pues la escasa información de la que se dispone al respecto también supone una importante limitación. Finalmente, otra importante limitación se encuentra en la valoración de la precepción de imparcialidad de los jueces debido a que la pregunta que evaluaba estaba variable se ha realizado al final del cuestionario y habría sido más adecuado hacerla antes de presentar el caso escenario para garantizar que las respuestas no se vieran influenciadas por la información aportada.

VII. Conclusiones

A finales del siglo XX, Pamela N. Gray en su libro *Artificial Legal Intelligence* (1997) presentaba una reflexión acerca de la implementación tecnológica del razonamiento jurídico a través de modelos computacionales de razonamiento legal y del uso del pensamiento evolutivo sobre el derecho. Si bien, con este trabajo queda demostrado que, pese a ser consideras utópicas e imaginativas en su época, las ideas presentadas por Gray no resultan nada extrañas en la actualidad. La IA ha llegado para quedarse y está transformando profundamente todos los ámbitos de nuestro día a día, incluido el sistema judicial. Las ventajas derivadas de ello son numerosas, pero igual de abundantes pueden llegar a ser sus inconvenientes si no se respetan aquellos aspectos éticos y legales que ya están fuertemente arraigados en nuestra sociedad.

Ante la creciente utilización de algoritmos, herramientas actuariales y predicciones basadas en IA con el objetivo último de mejorar la toma de decisiones jurídico-penales y el funcionamiento del sistema de justicia, el presente estudio ha querido ofrecer

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

una primera aproximación a las percepciones que los ciudadanos tienen al respecto. En primer lugar, porque las personas no somos sino los sujetos destinatarios de las decisiones jurídico-penales y, si en el procedimiento de toma de tales decisiones se pueden ver afectados, e incluso vulnerados, ciertos derechos, resulta conveniente conocer la opinión de la sociedad sobre ello. En segundo lugar, porque, tal y como vienen sosteniendo autores con un consolidado prestigio, la percepción pública de la legitimidad y equidad de las instituciones y de la toma de decisiones legales afecta fundamentalmente al cumplimiento, respeto y compromiso de los ciudadanos con el sistema de justicia penal (Tyler & Jackson, 2014). Es decir, como se anunciaba al principio de este trabajo, la automatización de decisiones, las predicciones algorítmicas y el uso de la IA en general, solo triunfarán y serán legítimos si cuentan con la aprobación de la ciudadanía.

Al respecto, los resultados obtenidos reflejan que el nivel de aceptación existente entre los participantes es bajo y, por ende, que la ciudadanía se muestra reacia a que los jueces tomen sus decisiones en base a predicciones algorítmicas acerca del riesgo de reincidencia de un sujeto. En consonancia con esto, la muestra analizada percibe, de forma mayoritaria, que los jueces son frecuentemente imparciales a la hora de decidir, por lo que se muestran favorables a que sean estos quienes sigan decidiendo qué pena o medida imponer a un sujeto peligroso. El hecho de que los participantes dispusieran de información sobre la validez predictiva de la herramienta no ha dado lugar a diferencias significativas en el nivel de aceptación de su uso. No obstante, sería conveniente llevar a cabo futuros estudios en los que se evalúen otro tipo de variables que puedan influir en el grado de aceptación, como el tipo delito cometido, u ofrecer información sobre el historial delictivo del sujeto evaluado.

Todo apunta a que, aunque en otros campos de la sociedad la automatización de decisiones agilice los procedimientos y sea socialmente aceptada, en el ámbito penal no ocurre lo mismo. Sin embargo, advierte Miró Llinares (2020) que, el reconocimiento de todos los defectos y de todos los riesgos potenciales del uso de estas tecnologías no debe llevarnos necesariamente a conformarnos con el conocimiento crítico y el regodeo, ni mucho menos a abrazar la tecnofobia. La IA y los algoritmos de big data no son algo dado y autodefinido que sea imposible de cambiar y que, por tanto, deba rechazarse globalmente. Tampoco son algo totalmente neutro que pueda ser utilizado "bien o mal", sino que constituyen herramientas definidas en un contexto sociopolítico y con unos objetivos determinados que no pueden ser ignorados y sobre los que es necesario realizar una reflexión ética general, así como otras muchas específicas sobre los desarrollos concretos que se producen y cada una de sus funcionalidades. Todo ello debe hacerse con el conocimiento empírico de cómo se construyen y cómo funcionan, con la discusión sobre las bases teóricas en las que se fundamentan, con la constante identificación de sus efectos, etc.

De conformidad con lo anterior, para este autor se debe adoptar una actitud realista, éticamente crítica y empíricamente informada en relación con el uso y desarrollo de estas herramientas en el ámbito de la justicia penal. Una actitud que parte de reconocer que no "sabemos" sino que sólo "suponemos" lo que son y lo que implican estas tecnologías, pero que al mismo tiempo persigue determinar normativamente su diseño e

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

implementación social. Si bien, todo ello teniendo en cuenta sus posibilidades reales actuales, sus desarrollos más próximos y las potenciales consecuencias de ambos, así como los condicionantes contextuales. Podría definirse, en cierto modo, como una actitud pragmática en la medida en que la valoración que hacemos de la tecnología debe medirse teniendo en cuenta el éxito que tiene o puede tener, en función de la visión científica de la que partimos. Pero no sólo en lo que se refiere a la práctica judicial, es decir, identificando el posible éxito como una reducción de la reincidencia o de la delincuencia, sino también en la sociedad. Es decir, se trata de incluir en el juicio sobre la bondad o maldad de la tecnología los valores éticos sociales recogidos en la constitución, considerados esenciales por la ciudadanía y que podrían verse afectados positiva o negativamente por el uso de tales tecnologías. A su vez, esto genera la necesidad de analizar adecuadamente lo que está en juego y proponer que el resultado de las decisiones éticas que se adopten se incluya en el diseño de la tecnología.

De esta forma, la utilidad real, tanto en un sentido actual como en el sentido de poder determinar objetivos pragmáticos antes de la implantación de una tecnología, constituirá el baremo esencial que habrá de ser tenido en cuenta. Pero, además, esta se entenderá como una utilidad ética que va mucho más allá de la prevención del delito o de la seguridad: la consecución de objetivos preventivos en un estado democrático que no retroceda en materia de derechos o libertades (Miró Llinares, 2020). En definitiva, se trata de adoptar todas aquellas medidas que sean necesarias y que permitan mitigar los posibles inconvenientes derivados de la aplicación de estas tecnologías. Tales medidas deben abarcar desde la asignación de los recursos necesarios para garantizar su efectiva puesta en marcha, hasta el establecimiento de procedimientos que permitan obtener una mayor transparencia en la recolección de los datos, y garantizando, en todo caso, que no se produzcan sesgos ni desigualdades de ningún tipo. Sin tales salvaguardas es probable que la ciudadanía continúe mostrándose reacia a su incorporación en los procesos judiciales y que se reproduzcan las mismas tendencias que han existido durante décadas y que acaban afectando de manera desproporcionada a los ciudadanos más vulnerables.

Bibliografía

AMADOR HIDALGO, L. (1996). *Inteligencia artificial y sistemas expertos*. Universidad de Córdoba.

ANDRÉS PUEYO, A., Y ECHEBURÚA, E. (2010). Valoración del riesgo de violencia: instrumentos disponibles e indicadores de aplicación. *Psicothema, 22*(3), 403-409.

ANDRÉS PUEYO, A., Y REDONDO ILLESCAS, S. (2007). Predicción de la violencia: entre la peligrosidad y la valoración del riesgo de violencia. *Papeles del Psicólogo, 28*(3), 157-173.

ANDRÉS-PUEYO, A. (2017). Manual de evaluación del riesgo de violencia. Metodología y ámbitos de aplicación. Ismael Loinaz. *Anuario de Psicología Jurídica, 27*, 127-129.

ARAUJO, T., HELBERGER, N., KRUIKEMEIER, S., Y DE VREESE, C.H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY, 35*, 611-623. https://doi.org/10.1007/s00146-019-00931-w

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

BAKER, R. S. (2016) Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, 26 (2), 600-614.

BOTNICK, C. (2015). Evidence-Based Practice and Sentencing in State Court: A Critique of the Missouri System. *Wash. UJL & Pol'y, 49*, 159– 180.

CAPDEVILA, M., et al. (2015). Tasa de reincidencia penitenciaria 2014. Centro de Estudios Jurídicos y Formación Especializada, Generalitat de Catalunya.

COHEN, J. (1992). A power primer. *Psychological bulletin, 112*(1), 155.

COUSIN, G. (2005). Case study research. *Journal of geography in higher education, 29*(3), 421-427.

DE MICHELE, M., BAUMGARTNER, P., BARRICK, K., COMFORT, M., SCAGGS, S., Y MISRA, S. (2019). What do criminal justice professionals think about risk assessment at pretrial. *Fed. Probation, 83,* 32.

DIETVORST, B., SIMMONS, J.P., Y MASSEY, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *Journal of Experimental Psychology: General, 144* (1), 114-126. http://dx.doi.org/10.1037/xge0000033

FREEMAN, K. (2016). Algorithmic injustice: How the Wisconsin Supreme Court failed to protect due process rights in State v. Loomis. North Carolina Journal of Law & Technology, 18(5), 75.

GALLO, J. A. (2006). Errores y sesgos cognigtivos en la expansión del Derecho Penal. In *Derecho y justicia penal en el siglo XXI: liber amicorum en homenaje al profesor Antonio González-Cuéllar García* (pp. 31-48). Constitución y Leyes, COLEX.

GALLO, J.A. (2011). Las decisiones en condiciones de incertidumbre y el derecho penal. *Revista para el Análisis del Derecho, 4*.

GARRET, B.L., Y MONAHAN, J. (2020). Judging Risk. California Law Review ,108, 438-493.

GASEK, J. (2019). Community First: Why California's Elimination of Cash Bail May Have Missed the Mark. McGeorge L. Rev., 51, 1.

GRAY, PAMELA N. *Artificial legal intelligence*. Ashgate Publishing Company, 1997.

GUTHRIE, C., RACHLINSKI, J. J., Y WISTRICH, A. J. (2001). Inside the judicial mind. Cornell *Law Review, 86*, 777-830.

HARTLEY, J. (2004). Case study research. In. Catherine Cassel e Gilian Symon. Essential guide to qualitative methods in organizational research.

JEE, K., Y KIM, G-H. (2013). Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. *Healthc InformRes,* 19(2), 79-85.

KAHNEMAN, D. Y TVERSKY, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica, 47*(2), pp. 263-291.

KAHNEMAN, D. Y TVERSKY, A. (1983). Extensional vs. intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293-315

KAHNEMAN, D., SLOVIC, P., Y TVERSKY, A (Eds.) (1982). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

KAHNEMAN, D., Y TVERSKY, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

KAHNEMAN, D., Y TVERSKY, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-251.

KEHL, D., GUO P., Y KESSLER, S. (2017). Algorithms in the criminal justice system: assessing the use of risk assessments in sentencing. *Berkman Klein Center for Internet & Society*.

KIM, G-H., TRIMI, S., Y CHUNG, J-H. (2014). Big-Data Applications in the Government Sector. *Communications of the ACM*, 57(3), 78-85.

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

KYUNG LEE, M. (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

LOINAZ, I. (2017). *Manual de evaluación del riesgo de violencia*. *Metodología y ámbitos de aplicación*. Pirámide.

MARTÍNEZ GARAY, L. (2018). Peligrosidad, algoritmos y due process: el caso State vs. Loomis. *Revista de Derecho Penal y Criminología*, *20*, 485-502.

MARTÍNEZ GARAY, L. (2020). Evidence-based sentencing y evidencia científica. A la vez, algunas consideraciones sobre "políticas basadas en la evidencia" y el Derecho Penal. *Teoría y Derecho: revista de pensamiento jurídico, 20*, 238-277.

MARTÍNEZ GARAY, L., Y MONTES SUAY, F. (2018). El uso de valoraciones del riesgo de violencia en Derecho Penal: algunas cautelas necesarias. *Revista para el Análisis del Derecho, 2*, 2-47.

MCKAY, C. (2020). Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making. *Current Issues in Criminal Justice, 32*(1), 22-39.

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.

MIRÓ LLINARES, F. (2020). Predictive Policing: Utopia or Dystopia? On attitudes towards the use of Big Data algorithms for law enforcement. *Revista de Internet, Derecho y Política*, (30).

MUÑOZ ARANGUREN, A. (2011). La influencia de los sesgos cognitivos en las decisiones jurisdiccionales: el factor humano. Una aproximación. *Indret: Revista para el Análisis del Derecho, 2.*

MYERS, D. G., Y LAMM, H. (1976). The group polarization phenomenon. *Psychological Bulletin, 83*, 602–27.

PIERSON, E. (2018). Demographics and discussion influence views on algorithmic fairness

ROBINSON (2013). *Intuitions of Justice and the Utility of Desert*. New York: Oxford University Press.

ROBINSON, P. H. (2000). Testing Lay Intuitions of Justice: How and Why? *Hofstra Law Review, 28.*

ROBINSON, P. H., Y DARLEY, J. M. (2007). Intuitions of justice: Implications for criminal law and justice policy. S. *Cal. L. Rev.,81*.

ROBINSON Y DARLEY (1995*). Justice, Liability, and Blame. Community Views and the Criminal Law*. Boulder: Westview Press.

SAXENA, N. A., HUANG, K., DEFILIPPIS, E., RADANOVIC, G., PARKES, D. C., Y LIU, Y. (2020). How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence*, *283*.

SCURICH, N., Y KRAUSS, D.A. (2019). Public's Views of Risk Assessment Algorithms and Pretrial Decision Making. *Psychology, Public Policy and Law, 26*(1), 1-32.

SELWYN, N. (2015). Data entry: towards the critical study of digital data and education. *Learning, Media and Technology*, 40(1),64–82.

SIEGEL, E. (2016). Predictive Analytics. The power to predict who will click, buy, lie or die. *New Jersey: John Wiley and Sons*.

SINGH, J. P., KRONER, D. G., WORMITH, J. S., DESMARAIS, S. L., Y HAMILTON, Z. (Eds.). (2018). Handbook of recidivism risk/needs assessment tools. John Wiley & Sons.

SKEEM, J., SCURICH, N., Y MONAHAN, J. (2020). Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law and human behavior.*

SOLER, C. (2013). RisCanvi. Protocolo de evaluación y gestión del riesgo de violencia con población penitenciaria. Foro internacional de buenas prácticas en prevención de la delincuencia

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

juvenil. Disponible online en https://www2.congreso.gob.pe/sicr/cendocbib/con4_uibd.nsf/138A3DBF8E8A85B905257C9F00803A14/$FILE/LinkClick6.pdf

SOLAR CAYÓN, J. I. (2020). La inteligencia artificial jurídica: nuevas herramientas y perspectivas metodológicas para el jurista. *Revus. Journal for Constitutional Theory and Philosophy of Law/Revija za ustavno teorijo in filozofijo prava*, (41).

STARR, S.B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stan L Rev,66*, 842-873

STEVENSON, M. T., Y DOLEAC, J.L. (2019). Algorithmic Risk Assessment in the Hands of Humans. *IZA-Institute of Labor Economics Discussion Paper Series, 12853*. Disponible en línea en https://www.iza.org/publications/dp/12853/algorithmic-risk-assessment-in-the-hands-of-humans

STEVENSON, M.T., Y DOLEAC, J.L. (2018). The Roadblock to Reform. *American Constitution Society*.

SUNDAR, S., Y NASS, C. (2001). Conceptualizing sources in online news. *Journal of Communication, 51*(1), 52–72.

SWANBORN, P. (2010). Case study research: what, why and how?. Sage.

TERRANOVA, V. A., WARD, K., SLEPICKA, J., & AZARI, A. M. (2020). Perceptions of Pretrial Risk Assessment: An Examination Across Role in the Initial Pretrial Release Decision. *Criminal Justice and Behavior, 47*(8), 927–942. https://doi.org/10.1177/0093854820932204

TYLER, T.R. Y JACKSON, J. (2014). Popular legitimacy and the exercise of legal authority: Motivating compliance, cooperation and engagement. *Psychology, Public Policy and Law, 20,* 78-95. DOI: http://dx.doi.org/10.1037/a0034514

VAN DIJCK, J. (2014). Datafication, dataism and dataveillance: big data between scientific paradigm and ideology. *Surveillance & Society, 12*(2), 197-208.

WASON, P. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129–140.

IUS ET SCIENTIA • 2021
Vol. 7 • N° 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

ANEXO I. Tabla resumen de las variables medidas en el estudio

Variable		Ítem	Escala
Sociodemográficas y de control	Sexo	¿Cuál es su sexo?	Masculino / Femenino
	Edad	¿Cuál es su edad?	-
	Estudios	¿Cuál es el nivel de estudios más elevado que ha alcanzado?	Educación primaria / Educación Secundaria Obligatoria / Bachillerato / Formación profesional / Grado o Licenciatura / Máster / Doctorado
	Estudios en Derecho	¿Tiene usted estudios en Derecho?	Sí / No
	Conocimiento sobre RisCanvi	¿Qué nivel de conocimiento diría usted que tiene acerca de la herramienta de valoración del riesgo dotada de IA "RisCanvi"?	0 = Nada 1 = Poco 2 = Medio 3 = Bastante 4 = Mucho
	Ideología política	Cuando se habla de política, se suelen utilizar los términos "izquierda" y "derecha". En una escala del 1 al 7 donde 1=Extrema izquierda y 7=Extrema derecha, ¿en qué punto de este eje político se situaría usted?	1 = Extrema izquierda 7 = Extrema derecha
Dependientes	Aceptación general del uso de IA (I)	En su opinión, ¿con qué frecuencia deberían los jueces basar sus decisiones en los resultados de este tipo de herramientas?	0 = Nunca 1 = Alguna vez 2 = Ocasionalmente 3 = Frecuentemente 4 = Siempre
	Aceptación general del uso de IA (II)	Y, en general, ¿cómo de aceptable le parece que se utilicen herramientas de valoración del riesgo automatizadas como RisCanvi en el sistema de justicia penal?	0 = Nada aceptable 1 = Poco aceptable 2 = Neutral 3 = Bastante aceptable 4 = Totalmente aceptable
	Aceptación específica del uso de IA (I)	¿En qué medida le parece aceptable que el Juez base su decisión de si poner en libertad o no al sujeto en el riesgo arrojado por la herramienta?	0 = Nada aceptable 1 = Poco aceptable 2 = Neutral 3 = Bastante aceptable 4 = Totalmente aceptable
	Aceptación específica del uso de IA (II)	Y, ¿en qué medida le parecería aceptable que el Juez tomara una decisión sobre la libertad de un sujeto contraria al riesgo mostrado por RisCanvi (i.e., poniendo en libertad a una persona sobre la que RisCanvi indica que tiene un riesgo alto de reincidir, o bien denegando la libertad a un sujeto sobre el que RisCanvi indica que tiene un riesgo bajo)?	0 = Nada aceptable 1 = Poco aceptable 2 = Neutral 3 = Bastante aceptable 4 = Totalmente aceptable

IUS ET SCIENTIA • 2021
Vol. 7 • Nº 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05

Variable		Ítem	Escala
Dependientes	Confiabilidad en los resultados de la herramienta	¿Cómo de confiable le parece el resultado arrojado por RisCanvi en este caso?	0 = Nada confiable 1 = Poco confiable 2 = Moderadamente confiable 3 = Bastante confiable 4 = Totalmente confiable
	Imparcialidad del Juez	Respecto de los jueces, en su opinión, ¿cree que los jueces son imparciales cuando toman decisiones judiciales en el ámbito penal?	0 = Nunca 1 = Alguna vez 2 = Ocasionalmente 3 = Frecuentemente 4 = Siempre
Independientes	Nivel de riesgo	Riesgo alto (manipulación) 0	Caso escenario
		Riesgo bajo (manipulación) 1	
		Sin información del riesgo 2	
	Información sobre la validez predictiva de RisCanvi	Sin información 0	Caso escenario
		Con información 1	

IUS ET SCIENTIA • 2021
Vol. 7 • N° 2 • pp. 57-87
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2021.i02.05