Este documento ha sido traducido por el Área de Servicios de Información, Traducciones y Lenguas Originarias de la Biblioteca del Congreso de la República con fines meramente informativos para los usuarios de la institución. Se trata de una traducción no oficial del texto en inglés «Impacto de los errores de la IA en un proceso con intervención humana»—publicado por Springer Nature— que no ha sido verificada por esta institución.[*]

Título del documento:

Inglés: «The impact of AI errors in a human-in-the-loop process», Agudo *et al*. *Cognitive Research: Principles and Implications* (2024) 9:1 https://doi.org/10.1186/s41235-023-00529-3. Artículo de acceso abierto.
N° de páginas: 16.
Fecha de documento: 2024-01-07.

Español: «Impacto de los errores de la IA en un proceso con intervención humana»
N° de páginas: 18.
Fecha de documento: febrero 2025

Institución: *SpringerOpen*. Las revistas y libros *SpringerOpen* están disponibles en línea de forma gratuita y permanente inmediatamente después de su publicación. Están sujetos a una revisión por pares de alto nivel y a servicios de autor y producción que garantizan la calidad y fiabilidad de la obra. Los autores que publican con SpringerOpen conservan los derechos de autor de su trabajo, licenciándolo bajo una licencia Creative Commons.
https://www.springeropen.com/

Para citar el artículo (en inglés): Agudo, U., Liberal, K.G., Arrese, M. *et al*. The impact of AI errors in a human-in-the-loop process. Cogn. Research 9, 1 (2024). https://doi.org/10.1186/s41235-023-00529-3

Autor del texto: Agudo *et al*. Cognitive Research: Principles and Implications (2024) 9:1 https://doi.org/10.1186/s41235-023-00529-3

---

[*] *N. de la T.: Documento traducido del inglés al español por el Área de Servicios de Información, Traducciones y Lenguas Originarias de la Biblioteca del Congreso de la República (MPZ).*

Agudo *et al.*
Cognitive Research: Principles and Implications (2024) 9:1
https://doi.org/10.1186/s41235-023-00529-3

Cognitive Research: Principles and Implications

ARTÍCULO ORIGINAL **Acceso abierto**

Impacto de los errores de la IA en un proceso con intervención humana

Ujué Agudo[1,2], Karlos G. Liberal[1], Miren Arrese1 y Helena Matute[2*]

Resumen

La toma de decisiones automatizada se ha hecho cada vez más común en el sector público. Por consiguiente, las instituciones políticas recomiendan la presencia de humanos en estos procesos de toma de decisiones como salvaguarda contra decisiones algorítmicas potencialmente erróneas o sesgadas. Sin embargo, la literatura científica sobre el desempeño del ser humano en el circuito no es concluyente sobre los beneficios y riesgos de dicha presencia humana, ni aclara qué aspectos de esta interacción humano-computadora pueden influir en la decisión final. En dos experimentos, simulamos un proceso automatizado de toma de decisiones en el que los participantes juzgan a múltiples acusados en relación con varios delitos, y manipulamos el tiempo en el que los participantes reciben apoyo de un supuesto sistema automatizado con inteligencia artificial (antes o después de emitir sus juicios). Nuestros resultados muestran que el juicio humano se ve afectado cuando los participantes reciben apoyo algorítmico incorrecto, particularmente cuando lo reciben antes de emitir su propio juicio, lo que resulta en una precisión reducida. Los datos y materiales para estos experimentos están ampliamente disponibles en Open Science Framework: https://osf.io/b6p4z/. El experimento 2 fue registrado previamente.

Palabras clave: Interacción humano-computadora, sesgo de automatización, IA, toma de decisiones, interacción humana, conformidad, inteligencia artificial

Antecedentes

La presencia de algoritmos de inteligencia artificial y sistemas automatizados en las decisiones del sector público (Araujo *et al*., 2020; Eubanks, 2018; O'Neil, 2016), como la asistencia social (Civio, 2022; De-Arteaga *et al*., 2020; López-Ossorio *et al.,* 2016), justicia (Casacuberta & Guersen-zvaig, 2018; Larson *et al.,* 2016; Martínez-Garay, 2016; Niiler, 2019), salud (Obermeyer *et al.,* 2019; Raghu *et al*., 2019) y educación (Alon-Barkat & Busuioc, 2022; Duncan *et al.,* 2020), es cada vez más común.

Así, muchos países ya utilizan sistemas automatizados de apoyo a las decisiones que a menudo se basan en inteligencia artificial (Solans *et al*., 2022). Ejemplos son Estados Unidos (Berkman

* Correspondencia: Helena Matute, matute@deusto.es
[1] Bikolabs/Biko, Pamplona, España
[2] Departamento de Psicología, Universidad de Deusto, Avda. Universidad 24, 48007 Bilbao, España

Klein Center, 2022), Reino Unido (Ministerio de Justicia, 2013), China (Wei, 2019), Estonia (Niiler, 2019), Argentina (Ministerio Público Fiscal de la Ciudad Autónoma de Buenos Aires, 2020), Polonia (Ministerstwo Sprawiedliwości, 2021) y España (Capdevila *et al*., 2015; Valdivia *et al.,* 2022) en el contexto judicial, que es el foco de este artículo. En estos casos, el sistema no suele tomar la decisión de forma totalmente autónoma, sino que apoya el proceso en la decisión humana (Araujo *et al.,* 2020) de diferentes maneras, ya sea recopilando y resumiendo la información necesaria para la decisión, o recomendando una decisión particular (Binns & Veale, 2021). Esta iniciativa de introducir humanos en un proceso de decisión automatizado se conoce en la literatura como interacción humana en el proceso. La idea es que esta presencia humana garantice una mejor decisión final debido a la supervisión humana del sistema y la intervención adecuada para prevenir o mitigar errores que pueda cometer el sistema automatizado (Ponce, 2022; Portela & Álvarez, 2022).

La legislación existente y las recomendaciones de políticas sobre sistemas de decisión automatizados (a menudo, designados por una variedad de términos como algoritmo, sistema de inteligencia artificial, tecnología de inteligencia artificial o robot; Comisión Europea, 2019) enfatizan el derecho de los ciudadanos a no estar sujetos a una decisión totalmente automatizada, y señalan la importancia de la presencia humana en el proceso como salvaguarda y protección ante una posible decisión algorítmica errónea o sesgada (Green, 2022; Portela & Álvarez, 2022).

Sin embargo, este enfoque no está exento de dificultades. Lograr una interacción adecuada entre humanos y sistemas automatizados es complejo porque requiere, entre otras cosas, que los humanos involucrados en el proceso tengan las habilidades, experiencia, motivación y tiempo para interpretar y gestionar de manera crítica la información proporcionada por el sistema (Ponce, 2022; Portela & Álvarez, 2022), que puedan comprender cómo funcionan estos sistemas y que sean capaces de discrepar de la decisión del sistema automatizado (Green, 2022) en caso de conflicto entre el juicio humano y la recomendación algorítmica (Valdivia *et al*., 2022).

Existe una gran cantidad de evidencia empírica que cuestiona la capacidad humana de no estar de acuerdo con una decisión automatizada o anularla. De hecho, durante más de dos décadas, la literatura científica ha señalado una tendencia humana a utilizar la información proporcionada por los sistemas de apoyo como un atajo para evitar buscar o procesar otra información relevante. De esta manera, las personas demuestran conformidad con la decisión del sistema o delegan su decisión al sistema. El exceso de conformidad humana cuando la evaluación del sistema es errónea, a menudo, en los campos de la ingeniería y la inteligencia artificial se denomina «sesgo de automatización» (Cummings, 2004; Lyell & Coiera, 2017; Mosier & Manzey, 2019; Parasuraman & Mustapha, 1996), y este efecto se ha documentado en ámbitos tan diversos como la aviación, la atención sanitaria, el ejército y el control de procesos (ver metaanálisis de Mosier & Manzey, 2019). Un ejemplo de este sesgo de automatización se encuentra en Lyell *et al.* (2017). En este estudio sobre la prescripción de medicamentos utilizando un sistema automatizado de apoyo a la toma de decisiones, los investigadores encontraron que cuando el sistema indicaba erróneamente que un medicamento no era apropiado para un paciente, los errores de prescripción aumentaban en un 56.9%.

En el sistema judicial, sin embargo, existen trabajos recientes que reportan un efecto menos consistente y robusto de este sesgo de automatización. Por un lado, hay casos de implementación de algoritmos que sugieren una excesiva conformidad humana con las decisiones del sistema, como en el caso de RisCanvi, el sistema utilizado para evaluar el riesgo de reincidencia de los reclusos en Cataluña, España. Según Saura y Aragó (2021), los funcionarios gubernamentales que utilizan RisCanvi no están de acuerdo con el algoritmo solo el 3.2% de las veces. Esto es así, aunque, como refleja el último informe general publicado sobre el rendimiento de este sistema, RisCanvi tiene una capacidad predictiva positiva del 18%, es decir, solo dos reclusos de cada diez acaban confirmando la predicción del sistema y reinciden tras ser catalogados de alto riesgo (Capdevila *et al.,* 2015; Martínez-Garay, 2016). Sin embargo, estos datos sobre la escasa capacidad predictiva de RisCanvi no se hacen visibles cuando se utiliza el sistema y, por lo tanto, es probable que sean desconocidos para los funcionarios gubernamentales que lo utilizan.

Por otro lado, varios estudios empíricos que utilizan modelos forenses de IA similares parecen sugerir resultados opuestos (Green & Chen, 2019a, 2019b, 2021; Grgic-Hlaca *et al*., 2019; Portela *et al.,* 2022; Skeem *et al.,* 2020). Por ejemplo, Grgic-Hlaca *et al.* (2019) llevaron a cabo un experimento en el que los participantes primero tuvieron que predecir, sin el apoyo de la IA, si algunos acusados reincidirían en

dos años. Luego, los investigadores mostraron a los participantes la predicción de reincidencia estimada por un programa informático, y se les pidió a los participantes que indicaran su predicción nuevamente. Los investigadores también mostraron a los participantes la tasa de precisión del programa de computadora (68%). Solo en una minoría de los casos los participantes ajustaron su predicción después de ver la estimación de la computadora, lo que muestra un bajo nivel de sesgo de automatización. Según los autores, el 32% de error informado de este sistema probablemente influyó en el bajo sesgo observado. Además, y como veremos más adelante, el hecho de que la predicción algorítmica se presentara después, y no antes, de que los participantes proporcionaran sus juicios, también puede haber sido un factor crítico en la baja conformidad observada en este estudio.

Además, en un experimento relacionado, Green y Chen (2019a) manipularon la raza del acusado para evaluar cómo estos datos afectaban la conformidad con el apoyo algorítmico. Descubrieron que el sesgo de automatización aumentaba cuando el algoritmo predecía un alto riesgo de reincidencia en los casos en los que el acusado era negro y un bajo riesgo de reincidencia en los casos en los que el acusado era blanco. Es decir, los participantes estuvieron de acuerdo con el algoritmo cuando confirmó sus propios prejuicios.

Es muy posible que estas contradicciones sobre el impacto del soporte del sistema automatizado en los procesos humanos se deban en parte a la amplia disparidad en los procedimientos metodológicos utilizados. Los trabajos existentes sobre este tema evalúan el papel de diferentes personas que toman decisiones que realizan diferentes tareas, en diferentes países, en diferentes dominios, en diferentes campos y con procesos de decisión muy diferentes. Los estudios también varían en términos de si los participantes están informados o no sobre la precisión predictiva del algoritmo; cómo se llama el algoritmo (sistema de apoyo a la decisión, programa informático, algoritmo o inteligencia artificial, entre otros); si el sistema proporciona o no soporte erróneo; si es necesaria o no una explicación de los criterios seguidos por el sistema; si los participantes reciben o no retroalimentación sobre cuán precisa fue su decisión, y en qué momento los participantes reciben apoyo del sistema, ya sea antes o después de emitir su propio juicio.

Por lo tanto, como ya hemos señalado, creemos que las diferencias en los resultados obtenidos en los estudios de sesgo de automatización en procesos con interacción humana pueden deberse en parte a la variedad de procedimientos metodológicos empleados en estos estudios y en esos modelos. Además, no todos esos estudios son un fiel reflejo de los procesos reales de toma de decisiones humanas utilizados en el sector público, por lo que es posible que no todos tengan el mismo valor ecológico desde una perspectiva aplicada. Por ejemplo, en casos de implementación real de sistemas de decisión automatizados en el sector público, el soporte del sistema generalmente se brinda al comienzo del proceso de decisión. En concreto, este proceso suele seguir la siguiente secuencia (Chong *et al*., 2022; Solans *et al*., 2022): Primero, el sistema evalúa la información disponible y muestra su valoración; luego, al humano se le dan solo algunas opciones: validar o modificar la evaluación del sistema. Esta secuencia implica que los tomadores de decisiones humanos nunca emiten explícitamente sus propios juicios, sino que simplemente validan o modifican las evaluaciones del sistema, y que el apoyo del sistema se reciba al inicio del proceso, estableciendo un orden en la presentación de la información, probablemente influye en el procesamiento de la información relevante para la toma de decisiones por parte de la persona (Marquardson & Grimes, 2018) y afecte la conformidad y precisión.

Un ejemplo de esta posible influencia sería el sesgo de anclaje (Rastogi *et al.,* 2022). El sesgo de anclaje es la tendencia a confiar excesivamente en una información que recibimos inicialmente (el ancla), de modo que luego tendemos a ajustar nuestro juicio final en función de ese punto de partida o ancla (Epley y Gilovich, 2006; Tversky y Kahneman, 1974). En el caso de los procesos en que interactúan seres humanos, el apoyo del sistema (por ejemplo, sugerir un riesgo alto, medio o bajo de reincidencia para un recluso) se presenta antes de que los humanos tomen sus decisiones. Esta decisión humana, como se señaló anteriormente, generalmente se limita a que los funcionarios gubernamentales confirmen o modifiquen la evaluación realizada previamente por el sistema. Así, este apoyo de la IA podría actuar como un ancla que condicione a la persona que deberá decidir, cuya decisión final sería simplemente un ajuste a la evaluación del sistema.

Por consiguiente, llevamos a cabo dos experimentos diseñados para probar si manipular el momento en el que se presenta el soporte del sistema en un proceso con interacción humana puede ayudar a aumentar la precisión de la decisión final y reducir la conformidad excesiva (es decir, el sesgo de la automatización)

cuando el sistema comete errores. Para recrear un proceso de decisión lo más cercano posible a los procesos reales implementados en el sector público, nuestros dos experimentos en el ámbito de la justicia simularon el sistema RisCanvi (Soler, 2013). Como se ha mencionado anteriormente, este sistema predice el riesgo de reincidencia de los reclusos en Cataluña, España, y lo usamos simplemente como ejemplo, porque incluye las características comunes a los otros sistemas descritos anteriormente: una secuencia específica en el proceso de decisión (primero el apoyo del sistema, luego validación o modificación por parte del decisor humano) y una interfaz interactiva muy simple que consta de solo dos botones, uno para validar y otro para modificar la evaluación del sistema. Por tanto, en este tipo de proceso con interacción humana, las personas que toman las decisiones no emiten explícitamente su propio juicio. En cambio, solo confirman o modifican la valoración previamente recibida del sistema. Creemos que esto probablemente podría favorecer la conformidad de las personas que deciden con la evaluación de la IA, lo que actuaría como ancla para la decisión humana. Como se mencionó anteriormente, las personas que usan RisCanvi están de acuerdo con el algoritmo el 96.8% de las veces (Saura & Aragó, 2021), a pesar de que el algoritmo tiene solo un 18% de poder predictivo positivo (Capdevila *et al.*, 2015; Martínez-Garay, 2016).

Existen pocos estudios de procesos con interacción humana que hayan manipulado el momento en que se recibe el soporte del sistema o que hayan presentado soporte algorítmico en diferentes momentos del proceso de decisión (Buçinca *et al.,* 2021; Echterhof *et al.,* 2022; Green & Chen, 2019b; Rastogi *et al.,* 2022; Vicente & Matute, 2023) que estudien si este apoyo puede provocar un efecto de anclaje en la decisión humana o afecta de alguna manera la decisión final. Por ejemplo, Green & Chen (2919b) realizaron un experimento en el que los participantes debían indicar en una escala del 0 a 100 la probabilidad de que varios reclusos no concurrirían al tribunal o que serían arrestados antes del juicio. En una de las condiciones experimentales, los participantes dieron su opinión antes de que se mostrara la evaluación algorítmica. Esta evaluación fue en ocasiones incorrecta, simulando el desempeño de sistemas del mundo real como el algoritmo COMPAS (Angwin *et al.,* 2016). En esta condición, en la que los participantes emitieron su juicio antes y después de recibir el apoyo algorítmico, se obtuvo mayor precisión, en comparación con otras condiciones en las que la evaluación algorítmica se proporcionó antes del juicio de los participantes o no se mostró en absoluto.

En otro contexto, Buçinca *et al.* (2021) realizaron un experimento en el que los participantes debían identificar el ingrediente con mayor contenido de carbohidratos en un plato de comida para reemplazarlo por otro plato con menos carbohidratos, pero de sabor similar. Descubrieron que la precisión y la conformidad de los participantes se vieron afectados por el momento en que recibieron apoyo erróneo de una IA. El desempeño de los participantes que emitieron su juicio antes de ver la evaluación incorrecta de la IA fue mejor que el de los participantes que vieron primero la evaluación de la IA. Además, el primer grupo también dio menos su conformidad que el grupo que recibió el apoyo erróneo de la IA antes de tomar la decisión. Aunque ninguno de los grupos que recibió el soporte incorrecto de la IA evitó por completo el sesgo de automatización, los autores sugieren que pedir a los participantes que emitan sus juicios antes de ver la evaluación incorrecta de la IA puede actuar como una función de forzamiento cognitivo. Esta función de forzamiento cognitivo obligaría a los usuarios de sistemas de apoyo en la toma de decisiones a pensar de manera más analítica e interrumpir el razonamiento rápido y heurístico que puede llevarlos a mostrar conformidad (Lambe *et al*., 2016).

Creemos que comprender el impacto de la interacción humana con la IA en los procesos de decisión automatizados puede conducir a decisiones más precisas y con menos sesgos de automatización, porque la actual falta de evidencia concluyente en esta área no está frenando la implementación de estos sistemas automatizados de apoyo a la toma de decisiones en el sector público, lo cual es motivo de preocupación. En consecuencia, como se mencionó anteriormente, realizamos dos experimentos, inspirados en sistemas de soporte de decisiones de IA del mundo real como RisCanvi. En estos experimentos, manipulamos el momento en que se reciben las evaluaciones algorítmicas en un proceso de interacción humana. Nuestro propósito era probar si esta manipulación podría contribuir a mejorar la toma de decisiones colaborativa entre humanos e IA al ayudar a reducir la conformidad cuando el sistema falla y aumentar la precisión de las decisiones. Es decir, nuestro propósito no era probar si los humanos son más o menos precisos que los sistemas automatizados al realizar sus evaluaciones. Nuestro objetivo era evaluar la secuencia estándar de toma de decisiones en procesos con intervención humana que, como se señaló anteriormente, consiste en que

el sistema primero muestra su evaluación y luego los humanos simplemente confirman o modifican esa evaluación, sin en ningún momento emitir expresamente su propio juicio. Consideramos que tal secuencia puede favorecer un sesgo de anclaje que podría afectar la precisión de las decisiones, incluso hasta el punto de conducir a una conformidad excesiva cuando el sistema falla. Si ese fuera el caso, creemos que cambiar el momento en el que se brinda el soporte de la IA y, como sugieren Lambe *et al.* (2016), obligar al ser humano a emitir explícitamente un juicio antes de recibir la evaluación del sistema, debería ser una buena estrategia para reducir el sesgo y, por tanto, aumentar la precisión y reducir la conformidad.

Experimento 1

Este experimento simula un proceso de intervención humana en el que los participantes reciben apoyo erróneo de un sistema de inteligencia artificial para decidir la culpabilidad de varios acusados. Nuestro propósito era probar si al forzar a los participantes a emitir su juicio explícitamente cuando aún no han recibido la evaluación errónea del sistema podría mejorar la precisión de la decisión, en comparación con dar como primer paso el soporte sesgado de la IA. Entonces, planteamos la hipótesis de que pedir a los tomadores de decisiones humanos que emitan su juicio antes de recibir la evaluación algorítmica mejoraría la precisión de su juicio y reduciría su conformidad con el soporte incorrecto de la IA, es decir, esto debería reducir su sesgo de automatización.

Método

Participantes

Reclutamos una muestra de 150 participantes (36.6% mujeres, 62.7% hombres, 0.7% no binarios), de 18 años o más (*M*=33.2, *SD*=11.4), a través de la plataforma Prolifc Academic. Dado que nuestro experimento, que se realizó en línea, se inspiró en el sistema de decisión automatizada utilizado en España, RisCanvi, reclutamos una muestra de este país. Para ello utilizamos los filtros «Nacionalidad: España» y «Primera lengua: español» en Prolifc. Si bien para realizar un experimento lo más similar posible al proceso de decisión del mundo real hubiera sido apropiado utilizar una muestra de funcionarios gubernamentales vinculados al ámbito penitenciario y judicial, optamos por una muestra de laicos. Esta decisión, además de facilitar el reclutamiento, estuvo respaldada por trabajos previos sobre sistemas automatizados de decisión en justicia, que afirman que el comportamiento de legos y profesionales no difiere (Green & Chen, 2021).

El análisis de sensibilidad para el tamaño de la muestra mostró que teníamos una capacidad del 80% para detectar efectos de tamaño pequeño a mediano (*w*=0.22). El programa en línea asignó aleatoriamente a cada participante a uno de dos grupos experimentales: SoporteIA→Juicio (*n*=76), o Juicio→SoporteIA (*n*=74).

Diseño y procedimiento

Después de proporcionar cierta información demográfica básica (edad y sexo), todos los participantes leyeron las mismas instrucciones, donde se les indicaba que su tarea era evaluar la probabilidad de que varios acusados fueran culpables, basándose en los testimonios de los testigos. También les dijimos que contarían con el apoyo de un sistema de inteligencia artificial. A continuación, preguntamos a los participantes sobre su grado de confianza, tanto en sus propias capacidades como en el sistema de IA, dado que estos dos factores podrían afectar la aceptación o rechazo del soporte algorítmico, como señalan algunos investigadores (Chong *et al.*, 2022; Verde y Chen, 2019a). Por consiguiente, los participantes tuvieron que indicar cuán seguros estaban de que realizarían la tarea correctamente y de que el sistema de inteligencia artificial evaluaría adecuadamente la culpabilidad de los acusados. Luego, el experimento propiamente dicho comenzó.

El experimento consistió en tres pruebas para cada participante, y cada prueba constaba de tres pasos. La tabla 1 muestra un resumen de los tres pasos de cada prueba. En el paso 0, la computadora presentó a los participantes un caso penal para ser juzgado y los testimonios asociados a este. Para utilizar materiales estandarizados se utilizaron los casos penales del banco normativo de testimonios ForenPsy 1.0 desarrollado por Álvarez *et al*. (2023). Este banco incluye la descripción de tres casos penales (homicidio, amenazas y violación de propiedad privada) con 15 testimonios cada uno. En el estudio de Álvarez *et al.*, los 45 testimonios fueron clasificados por una muestra de participantes anónimos, quienes estimaron el grado de inocencia o culpabilidad que cada testimonio sugería sobre cada uno de los tres acusados.

Así, el paso 0 de cada juicio consistió en una descripción de uno de los tres casos penales de ForenPsy 1.0 (Álvarez *et al.*, 2023), junto con siete

Tabla 1: Resumen del diseño de los experimentos 1 y 2 que muestra los pasos de cada prueba

Grupo	Paso 0	Paso 1	Paso 2
SoporteIA →Juicio	Descripción de un caso penal y testimonios de testigos	Soporte IA (confirmar o modificar evaluación de IA	Juicio (sin que exista soporte de IA)
Juicio→SoporteIA		Juicio (sin que exista soporte de IA	Soporte IA (confirmar o modificar evaluación de IA

testimonios que sugerían inocencia o culpabilidad (ver fig. 1). Cinco de los siete testimonios apuntaban claramente a uno de los veredictos (inocencia o culpabilidad), y los otros dos apuntaban en la misma dirección, pero eran algo ambiguos, según la calibración de ForenPsy. La introducción de estos dos testimonios más ambiguos tenía como objetivo añadir realismo a los juicios. Algunos participantes consideraron los siete testimonios que sugerían inocencia, mientras que otros consideraron los siete testimonios que indicaban culpabilidad. El tipo de testimonios que recibió cada participante (es decir, inocencia o culpabilidad) fue aleatorizado. Además, el orden de presentación de cada caso y de caso testimonio se aleatorizó en cada prueba.

Nuestra principal manipulación experimental tuvo lugar en los pasos 1 y 2 de cada prueba. El orden en que ocurrieron estos dos pasos se invirtió para cada uno de los dos grupos diferentes (ver tabla 1). En el grupo SoporteIA⟶Juicio, durante el paso 1, a los participantes se les mostró la probabilidad de culpabilidad estimada por un sistema de inteligencia artificial (ficticio)[1]. La evaluación de IA de la culpabilidad del acusado que se mostró a los participantes solo podía tomar dos valores: alta probabilidad de culpabilidad o baja probabilidad de culpabilidad (ver fig. 2), por lo que podría ser congruente o contradictorio con el veredicto sugerido por los testimonios presentados durante el paso 0. En las dos primeras pruebas, la evaluación del sistema siempre fue correcta; esto es, siempre fue coherente con los testimonios presentados previamente que sugerían sea la inocencia o culpabilidad según la calibración de ForenPsy. Solo en la última prueba (en adelante, la prueba incorrecta) la evaluación del sistema fue errónea. En esta prueba el sistema siempre sugirió el veredicto contrario al sugerido por los testimonios. Por ejemplo, si los testimonios presentados habían sido calificados en ForenPsy 1.0 como indicativos de inocencia, el sistema sugirió una alta probabilidad de culpabilidad. Y si los testimonios habían sido calificados en ForenPsy 1.0 como sugerentes de culpabilidad, entonces el sistema indicaba una alta probabilidad de inocencia. Por tanto, la precisión de nuestro sistema ficticio fue del 66% (un error de cada tres), una tasa que no compartimos con los participantes porque los funcionarios gubernamentales que usan estos sistemas generalmente tampoco reciben esta información. Este nivel de precisión es muy similar al reportado por sistemas similares, como RisCanvi (Capdevila *et al.,* 2015) y COMPAS (Angwin *et al.,* 2016).

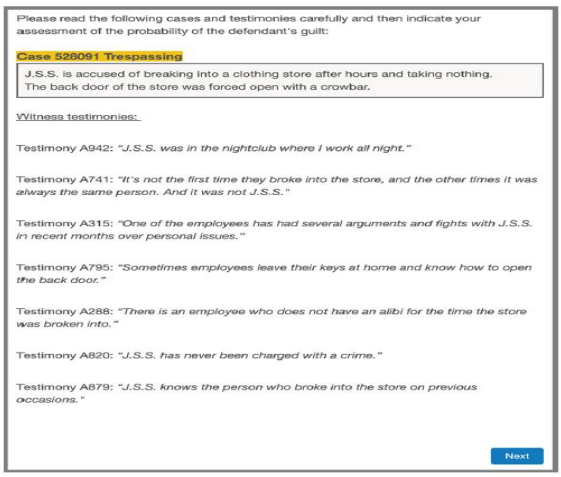

Fig. 1 Ejemplo de captura de pantalla del paso 0 (Descripción de un caso penal y declaraciones de testigos), en el experimento 1. *Nota*. En el experimento 1, cada caso penal mostraba siete testimonios (de inocencia en el ejemplo), mientras que en el experimento 2, cinco.

En la misma pantalla donde se mostraba la valoración del sistema, los participantes del grupo SoporteIA⟶Juicio debían elegir entre confirmar o modificar la valoración del sistema de IA, pulsando el

[1] Aunque la toma de decisiones automatizada en el sector justicia se utiliza tanto para decidir sobre acontecimientos pasados (por ejemplo, la condena de un acusado) como para estimar la probabilidad de acontecimientos futuros (por ejemplo, el riesgo de reincidencia del recluso), optamos por la primera opción en nuestro experimento, pues consideramos que sería más fácil para un participante no experto en temas de justicia decidir sobre una acción pasada antes que predecir una acción futura.

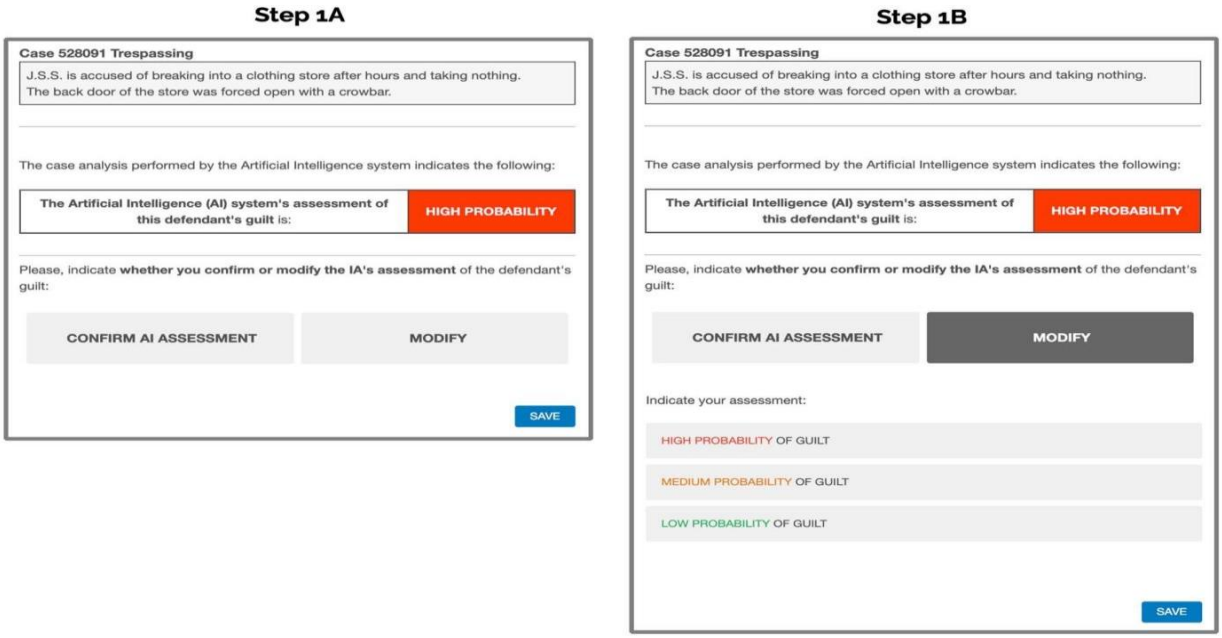

Fig. 2. El paso 1A muestra un ejemplo de una captura de pantalla del paso 1 (confirmar o modificar la evaluación de IA) en el Grupo SoporteIA→Juicio, en los experimentos 1 y 2). *Nota*. El paso 1B solo se mostró a los participantes que pulsaron el botón «Modificar». En ese caso, el cambio de 1A a 1B no implicó un cambio de pantalla. La nueva información se mostró debajo de los botones «Confirmar» y «Modificar» en la misma pantalla. El grupo Juicio→SoporteIA vio estas capturas de pantalla en el paso 2

botón correspondiente. Si los participantes decidían confirmar la evaluación de la IA, procedían directamente al paso 2. Si decidían modificar, aparecía una lista seleccionable debajo del botón y los participantes podían modificar la evaluación de la IA eligiendo una de estas tres opciones: alta, media o baja probabilidad de culpa (ver fig. 2). La razón por la que usamos una respuesta de tres opciones durante el paso 1, en lugar de usar una escala más continua y sensata, fue porque queríamos usar una medida lo más similar posible a la que comúnmente usan los sistemas de apoyo a las decisiones en la vida real. implementados en el sistema judicial (es decir, riesgo de reincidencia alto, moderado y bajo). Además, elegimos esta escala de tres puntos más realista, en lugar de simplificar la escala y utilizar solo las dos opciones que podía arrojar la evaluación de la IA (alta o baja probabilidad de culpabilidad), a fin de analizar si —en caso que los participantes no dieran conformidad al soporte incorrecto de la IA— esto implicaba que fueron precisos en su decisión (porque su veredicto fue congruente con lo indicado por los testimonios) o que no supieron si confirmar el apoyo de AI o el veredicto sugerido por los testimonios, por lo que no fueron precisos (porque eligieron la probabilidad media de culpa). Después de modificar la evaluación de la IA, estos participantes también procedieron al paso 2.

En el paso 2, a los participantes del grupo SoporteIA→Juicio se les dijo que tenían que indicar su juicio final sobre la culpabilidad del acusado. Este juicio final se proporcionó utilizando una lista seleccionable, que era idéntica a la utilizada cuando los participantes eligieron modificar la evaluación de la IA durante el paso 1 (es decir, probabilidad de culpabilidad alta, media o baja). Este paso puede parecer repetitivo, pero se agregó para (a) obtener al menos un juicio personal de todos los participantes (es decir, incluso de aquellos que eligen simplemente confirmar la evaluación de la IA en el paso anterior), y (b) igualar el número de veces que a los participantes de ambos grupos se les pidió que emitieran su juicio (ver tabla 1). Es decir, era importante que ambos grupos tuvieran el mismo tipo y número de pruebas para que solo hubiera un factor: el momento en que se mostrase la evaluación de la IA. Por lo tanto, si se observaron diferencias cuando los participantes emitieron sus juicios personales (sin que el apoyo de la IA estuviera presente), estas diferencias solo podrían atribuirse a que un grupo ya había recibido el apoyo de la IA en la fase anterior. Para que esta solicitud pareciera más natural, las instrucciones del paso 2 informaron a los participantes de este grupo que la evaluación que debían realizar era la que cerraba definitivamente el caso (ver fig. 3).

En el grupo Juicio→SoporteIA, la única diferencia fue que el orden de los pasos 1 y 2 se invirtió. Es decir,

durante el paso 1, los participantes de este grupo emitieron su juicio personal sobre la probabilidad de culpabilidad del acusado sin el apoyo de la IA, y utilizando la misma escala de tres puntos (probabilidad alta, media o baja de culpa) como el otro grupo. Esto fue diseñado como una propuesta de mejora a la secuencia de decisión habitual de los procesos humanos en el circuito que no piden explícitamente a los humanos que emitan su juicio antes de recibir la evaluación de la IA. Esperábamos que al obligar a estos participantes a emitir este juicio en un paso antes de ver la evaluación incorrecta de la IA pudiera mejorar la precisión y reducir la conformidad en su veredicto. A continuación, en el paso 2, se mostró la evaluación del sistema y los participantes de este grupo tuvieron la oportunidad de confirmarla o modificarla, al igual que los participantes del otro grupo durante el paso 1. Si decidían modificarlo, se les mostraba nuevamente la misma escala de tres puntos que en el paso anterior para que pudieran modificar la valoración de la IA según sus criterios.

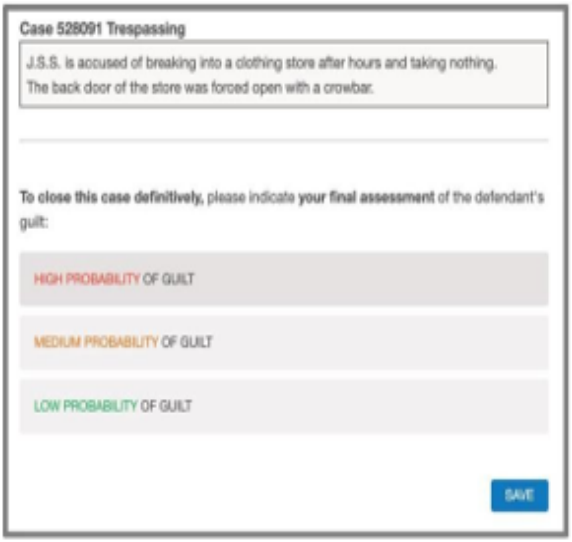

Fig. 3 Ejemplo de capturas de pantalla en el paso 2 (Juicio sin Soporte de IA) en el grupo SoporteIA→Juicio, en el experimento 1. *Nota*. El grupo Juicio→SoporteIA vio esta captura de pantalla en el paso 1. En ese caso, la frase «Cierre este caso definitivamente» no apareció en esta pantalla, sino en la correspondiente pantalla del paso 2.

La ausencia de un sistema automatizado real y el uso del conjunto de testimonios ForenPsy 1.0 nos permitió definir y controlar en detalle la apariencia, formato y errores del supuesto sistema algorítmico de soporte. Así, controlamos cuándo la evaluación de la IA era correcta (el apoyo de la IA era congruente con los testimonios) y cuándo la evaluación de la IA era incorrecta (la sugerencia del sistema de apoyo de la IA era contraria a la de los testimonios).

Una vez que los participantes completaron los tres pasos para cada uno de los tres casos penales que recibieron, se les preguntó nuevamente a todos los participantes sobre su confianza en sí mismos y en el sistema, utilizando las mismas preguntas que se utilizaron al comienzo del experimento. También se les preguntó si su trabajo o estudios estaban relacionados con la tecnología o el área de justicia[2]. Al finalizar, se informó brevemente a los participantes sobre el propósito real del estudio durante la etapa final de la reunión informativa.

Resultados y discusión

Precisión del juicio sin soporte de IA en el juicio incorrecto

Primero analizamos la precisión cuando los participantes emiten su juicio personal en la prueba incorrecta y sin soporte de la IA en ese momento. Es importante tener en cuenta que el paso en el que los participantes indicaron su juicio sin el apoyo de la IA difirió en función del grupo (ver tabla 1). Mientras que los participantes en el grupo Juicio→SoporteIA evaluaron la culpabilidad de los acusados por sí mismos en el paso 1, es decir, antes de recibir el apoyo incorrecto de AI en el siguiente paso, los participantes en el grupo SoporteIA→ Juicio lo hicieron en el paso 2, es decir, después de haber visto la evaluación de la IA en el paso 1. Esto nos permitió probar si juzgar un caso penal sin haber visto la evaluación de la IA incorrecta en ningún momento, en comparación con haberla visto en un paso anterior, daba como resultado un juicio más preciso.

Como esperábamos, los participantes del grupo Juicio →SoporteIA (es decir, el grupo que juzgó al acusado sin haber visto la evaluación incorrecta de la IA) fueron más precisos en las pruebas incorrectas que los participantes del grupo de SoporteIA→Juicio. Esto se puede ver en la fig. 4. Una prueba X^2 (chi al cuadrado) que analiza si los participantes acertaron o no en su juicio confirmó que las diferencias entre grupos eran

[2] Del número total de participantes, solo unos pocos informaron que tenían experiencia laboral o educativa en el ramo de la justicia (n = 2) o en tecnología (n = 35).

estadísticamente significativas, χ^2 (1) = 12.95, p<0.001, V de Cramer= 0.29. De todos los participantes en el grupo Juicio→SoporteIA, el 66.2% (49 de 74) proporcionaron juicios precisos, en comparación con el 36.8% de los participantes en el grupo SoporteIA→Juicio (28 de 76) que mostraron juicios precisos. Por lo tanto, parece que, como se esperaba, emitir su juicio personal antes de ver la evaluación incorrecta de la IA condujo a una mayor precisión, ya que los participantes en el grupo Juicio→SoporteIA fueron más precisos que los participantes en el grupo SoporteIA→Juicio.

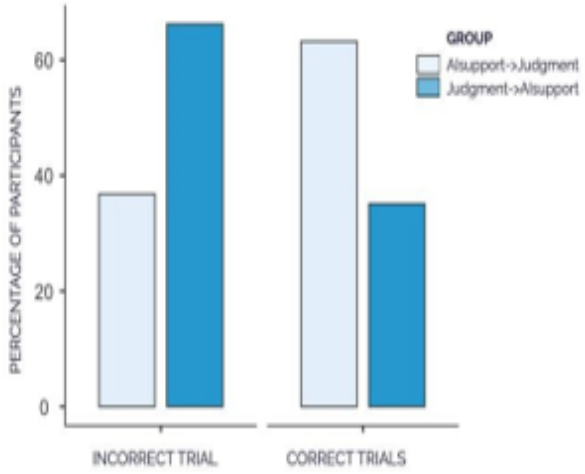

Fig. 4 Porcentaje de participantes con evaluaciones correctas en cada grupo por tipo de prueba (incorrecta o correcta), en el experimento 1. *Nota*. En las pruebas correctas, los testimonios y la evaluación de la IA que los participantes vieron en los pasos 1 y 2 (según el grupo) fueron congruentes; en la prueba incorrecta, fueron incongruentes. Los datos de precisión en las pruebas correctas representan el porcentaje de participantes cuyos juicios fueron precisos en ambas pruebas correctas.

Precisión de juicio sin soporte de IA en las pruebas correctas

A continuación, analizamos la precisión cuando los participantes emitieron sus juicios personales (sin el apoyo de la IA en ese momento) en las pruebas correctas, es decir, las dos pruebas en las que la IA sugirió el mismo veredicto que los testimonios. Nuevamente, en ambas pruebas, los participantes hicieron sus propios juicios personales ya sea en el paso 1, es decir, sin haber visto la evaluación correcta de la IA (grupo Juicio→SoporteIA), o en el paso 2, es decir, después de haber visto la evaluación de IA correcta. en el paso anterior (grupo SoporteIA→Juicio).

Para realizar este análisis, clasificamos a los participantes como precisos en las pruebas correctas si realizaron las evaluaciones correctas en ambas pruebas. Al contrario de lo que sucedió en la prueba incorrecta, encontramos que los participantes en el grupo SoporteIA→Juicio fueron generalmente más precisos que aquellos en el grupo Juicio→ SoporteIA. Según la chi al cuadrado, la asociación entre precisión y grupo fue estadísticamente significativa, χ^2 (1) = 11.8, p<0.001, V de Cramer=0.28. En el grupo SoporteIA→Juicio, el 63.2% de los participantes (48 de los 76 sujetos) acertaron en ambos casos correctos, mientras que solo el 35.1% de los participantes en el grupo Juicio→SoporteIA (26 de los 74) fueron precisos en ambos casos (ver fig. 4). Así, parece que haber recibido una evaluación correcta por parte de la IA en un paso previo a emitir su juicio personal, como es el caso en el grupo SoporteIA→Juicio, lleva a los participantes de este grupo a un aumento en la precisión de su juicio personal en las pruebas correctas.

Conformidad de la evaluación de la IA en la prueba incorrecta

A continuación, analizamos la conformidad de los participantes (es decir, el sesgo de automatización en este caso) cuando reciben apoyo de la IA y esta evaluación es incorrecta. Clasificamos a los participantes que mostraron conformidad en la prueba incorrecta si habían pulsado el botón para confirmar la evaluación de la IA, o si, a pesar de pulsar el botón para modificarla, finalmente seleccionaron la misma probabilidad de culpabilidad que la IA había sugerido (ver figura 2). Esto podía suceder en diferentes pasos dependiendo del grupo: paso 1 en el grupo SoporteIA →Juicio, y paso 2 en el grupo Juicio→SoporteIA. Esperábamos una conformidad menor por parte de los participantes en el grupo Juicio→SoporteIA en comparación con el otro grupo, porque estos participantes ya habían juzgado el caso por sí mismos en el paso anterior. Por tanto, esperábamos que su juicio previo evitara el efecto de anclaje que puede ocurrir cuando la evaluación de la IA se presenta primero, e incluso podía servir, según sugieren Lambe *et al*. (2016), como una función de forzamiento cognitivo. En resumen, esperaríamos que esta manipulación les facilitara la detección del error en la evaluación de la IA y redujera el posible sesgo de automatización.

Construimos una tabla de contingencia para analizar la relación entre la conformidad de los participantes en la prueba incorrecta y el grupo, y encontramos que solo 25 de los 150 participantes validaron la evaluación errónea de la IA, por lo que solo el 16.7% mostró una conformidad excesiva con la IA. De estos 25 participantes, 10 pertenecían al grupo Juicio→

SoporteIA y 15 al grupo SoporteIA→Juicio. Esta diferencia es tan pequeña que no permite un análisis más detallado de este efecto de la conformidad[3].

A continuación, analizamos si esta falta de conformidad implica que los participantes fueron realmente precisos en sus juicios al confirmar o modificar la evaluación incorrecta de la IA, y si entre los grupos existe una diferencia en esta precisión. Así, comparamos su desempeño durante el paso en que estaba presente el soporte de IA y simplemente se les pidió que confirmaran o modificar la evaluación de la IA. Es decir, comparamos el paso 1 del grupo SoporteIA→Juicio con el paso 2 del grupo Juicio→SoporteIA. Esto nos permite probar si nuestra propuesta de forzar un juicio explícito al comienzo del proceso (como ocurre en el grupo Juicio→SoporteIA) mejora la precisión de la secuencia estándar de procesos de intervención humana que no preguntan por un juicio humano explícito antes de que se presente el soporte de IA (como se imita en el grupo SoporteIA→Juicio). Estamos interesados en esta comparación en lugar de comparar la decisión final entre los grupos en el paso 2, porque nuestra propuesta no es introducir un juicio explícito en ningún punto del proceso, sino forzarlo al inicio, a diferencia de la práctica habitual de presentar el soporte de IA al principio.

Según el chi al cuadrado, la asociación entre precisión y grupo no fue estadísticamente significativa, χ^2 (1) = 1.29, p = 0.256, V de Cramer = 0.09. En el grupo SoporteIA→Juicio, el 34.2% de los participantes (26 de los 76) acertaron en su decisión, mientras que el 43.2% de los participantes (32 de los 74) acertaron en el grupo Juicio→SoporteIA. Al parecer, aunque se fuerce el juicio al comienzo del proceso en el grupo Juicio→SoporteIA produce decisiones más precisas cuando no se ha visto la evaluación incorrecta de la IA, recibir este apoyo incorrecto en el siguiente paso perjudica el veredicto de los participantes al disminuir su precisión y alinearlo con los niveles del grupo SoporteIA→Juicio.

Cabe señalar que, para simular un proceso de decisión con intervención humana en la vida real en nuestro experimento, utilizamos una escala de tres niveles para solicitar a los participantes evaluaciones de culpabilidad de una manera similar a la utilizada por los algoritmos de soporte a las decisiones aplicando IA en los sistemas judiciales. Consideramos valiosa la contribución de este experimento precisamente porque hemos intentado simular un proceso de decisión real con la interacción humana aplicando IA. Sin embargo, éramos conscientes de que tal decisión procedimental implicaba elegir una escala con baja sensibilidad, por lo que decidimos utilizar una escala más sensible de 0 a 100 en nuestro siguiente experimento. Así, llevamos a cabo un nuevo experimento en el que modificamos algunas de las decisiones procedimentales anteriores, buscando una mayor robustez en los resultados a costa de una pequeña reducción en el valor ecológico del experimento.

Experimento 2

El objetivo de este experimento es replicar los resultados del experimento 1 y obtener resultados más robustos y generalizarlos a una muestra más grande. Para ello, realizamos tres modificaciones principales al experimento anterior. Primero, cambiamos la escala en la que los participantes hicieron sus evaluaciones de la escala de tres puntos utilizada en el experimento 1 (que simula sistemas de apoyo a la toma de decisiones de IA del mundo real) a una escala más estándar de 0 a 100 utilizada en la investigación psicológica (ver fig. 5). Este cambio permite mediciones más sensatas y también facilita el uso de análisis estadísticos más robustos.

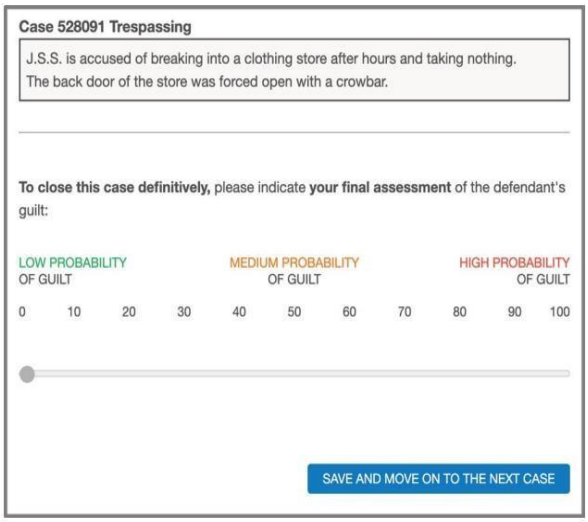

Fig. 5 Ejemplo de una captura de pantalla en el paso 2 (Juicio sin soporte de IA) en el grupo SoporteIA→Juicio,

[3] Además, estos datos de conformidad baja no nos permitieron analizar si la autoconfianza de los participantes o la confianza en el sistema contribuiría a incrementar los sesgos de automatización.

en el experimento 2. *Nota*. Para indicar cuándo se completaron los pasos de cada caso, en este experimento 2 cambiamos la palabra del botón «Guardar» del experimento 1 por «Guardar y pasar al siguiente caso». El grupo Juicio→SoporteIA vio esta captura de pantalla en el paso 1. En ese caso, la frase «Para cerrar este caso definitivamente» de las instrucciones no apareció en esta pantalla, sino en la pantalla del paso 2.

En segundo lugar, aumentamos a nueve (en lugar de tres) el número de casos penales que deben juzgarse para poder aumentar también la sensibilidad de esta manera. Esto también nos permitirá comparar los juicios de los participantes en tres casos incorrectos en lugar de solo un caso incorrecto, y en seis casos correctos en lugar de dos. El índice de precisión se mantiene en un nivel real del 66%. Además, también eliminamos los testimonios ambiguos de los casos presentados y utilizamos solo los cinco testimonios de cada caso que apuntaban más claramente a la inocencia o la culpabilidad. Decidimos eliminar la ambigüedad porque si incluso cuando los materiales son muy fáciles y el veredicto es obvio, los participantes se dejan engañar por el soporte erróneo de la IA, entonces tendríamos evidencia clara de un problema grave, con personas siguiendo los errores de la IA incluso en casos que podrían resolver fácilmente por sí solos.

Por último, ampliamos la muestra de participantes no solo en número (260 participantes en este experimento), sino también en diversidad: aunque mantuvimos el español como idioma del estudio, abrimos la participación a personas de cualquier país. Este experimento fue registrado previamente en https://aspredicted.org/ph9br.pdf.

Al igual que en el experimento 1, esperábamos que los participantes que recibieran el apoyo erróneo de IA al comienzo del proceso (grupo SoporteIA→ Juicio) mostrarían mayor conformidad y menor precisión que aquellos que emiten su propio juicio antes de recibir el apoyo incorrecto de IA (grupo Juicio→SoporteIA). Además, esperábamos una mayor precisión en los juicios de los participantes del grupo Juicio→SoporteIA sobre pruebas incorrectas en comparación con los juicios del grupo SoporteIA→Juicio.

Método

Participantes

Reclutamos una muestra de 260 participantes (42.3% mujeres, 54.6% hombres, 3.1% no binarios), de 18 años o más (*M*=30.7, *DS*=9.23), a través de la plataforma Prolifc Academic. Utilizamos el servicio de selección interno de Prolifc para reclutar esta muestra específica: los participantes mayores de 18 años que no habían participado previamente en otros experimentos realizados por nuestro equipo de investigación en la plataforma Prolifc, con el español como primera lengua, pero de cualquier país. Así, las nacionalidades más representadas fueron mexicana (40% de los participantes), española (37.7%) y chilena (8.5%), pero también hubo participantes de Italia, Estados Unidos, Venezuela, Perú y Colombia, entre otros.

El análisis de sensibilidad para el tamaño de la muestra mostró que teníamos un poder del 80% para detectar efectos pequeños (*d*=0.10). Como en el experimento anterior, los participantes fueron asignados aleatoriamente a uno de los dos grupos experimentales: SoporteIA→Juicio (*n*=132), o Juicio→SoporteIA (*n*=128).

Diseño y procedimiento

El diseño y procedimiento fueron muy similares a los del experimento anterior. En primer lugar, los participantes leyeron las instrucciones y brindaron su edad y sexo. Esta vez no les preguntamos sobre su confianza en sus propias capacidades y en las capacidades del sistema, ni al principio ni al final del estudio, porque estas medidas no afectaron los resultados del experimento 1.

Luego presentamos los casos que se debían juzgar, con algunos cambios respecto al experimento anterior que describimos a continuación. Esta vez, cada participante vio nueve juicios (nueve casos penales) en lugar de tres. Para utilizar los materiales estandarizados de ForenPsy 1.0 ya desarrollados y probados por Álvarez *et al*. (2023), habíamos utilizado en el experimento 1 solo tres casos penales (dos casos correctos y un caso incorrecto). Sin embargo, para ganar sensibilidad en el experimento 2, decidimos aumentar el número de casos a nueve. Estas nueve pruebas se presentaron agrupadas por tipo de delito (tres por homicidio, tres por amenaza y tres por allanamiento de morada), y tanto estos grupos de delitos como las pruebas dentro de estos grupos se presentaron en orden aleatorio. De las nueve pruebas, tres fueron los mismos casos utilizados en el experimento 1, basado en el conjunto estandarizado ForenPsy 1.0 (Álvarez *et al*., 2023). Las otras seis pruebas se crearon utilizando el modelo de lenguaje grande ChatGPT4 AI (OpenAI, 2023), que editamos para mayor claridad y coherencia.

Así, en cada uno de los nueve juicios, el paso 0 consistió en un artículo de portada que describía un caso penal y cinco testimonios que indicaban claramente un veredicto de inocencia o culpabilidad para el acusado. Para que el veredicto sea más obvio para los participantes y así controlar mejor cuándo la evaluación de la IA sería incorrecta, eliminamos los dos testimonios ambiguos por caso que habíamos usado en el experimento 1, por lo que usamos solo cinco testimonios por caso en este experimento. Para ponderar los testimonios de los seis nuevos casos creados para este experimento, realizamos un estudio previo con una muestra de 52 voluntarios[4] para calibrar y seleccionar los testimonios que indicaban con mayor claridad un veredicto de inocencia o culpabilidad.

Al igual que en el experimento 1, la proporción de casos con apoyo erróneo de la IA fue del 33%, es decir, tres de los nueve casos (uno de homicidio, uno de amenazas y uno de allanamiento de morada) en este experimento fueron incorrectos. Aunque el orden en el que se presentaron las pruebas para cada delito fue aleatorio, forzamos que el primer juicio que cada participante tuvo que considerar, no fuera nunca un juicio incorrecto para generar cierta confianza en el sistema de IA.

La manipulación experimental tuvo lugar en los pasos 1 y 2. En el grupo SoporteIA→Juicio, en el paso 1, los participantes consideraron la evaluación del sistema de IA (baja o alta probabilidad de culpabilidad, como en el experimento 1). En los casos incorrectos, la evaluación de IA fue incongruente con los testimonios presentados anteriormente y sugirió el veredicto opuesto. Al igual que en el experimento 1, en la misma pantalla se preguntó a los participantes si querían confirmar la evaluación de la IA o modificarla. Si decidieran modificarla, podrían utilizar la misma escala de tres puntos que en el experimento anterior (alta, media o baja probabilidad de culpabilidad). Luego, en el paso 2, se pidió a los participantes que indicaran su evaluación final. En este experimento, este juicio se indicó en una escala más estándar de 0 a 100 (donde 0 era baja probabilidad de culpabilidad y 100 era alta probabilidad de culpabilidad; ver fig. 5).

En el grupo Juicio→SoporteIA, el orden de los pasos 1 y 2 se invirtió: los participantes emitieron su juicio sobre la probabilidad de culpabilidad del acusado en la escala de 0 a 100 puntos en el paso 1 y luego vieron la evaluación de la IA, y la confirmaron o modificaron. como veredicto en el paso 2. Finalmente, todos los participantes indicaron si su trabajo o estudios estaban relacionados con la tecnología o justicia, y fueron interrogados e informados sobre el verdadero propósito del estudio.

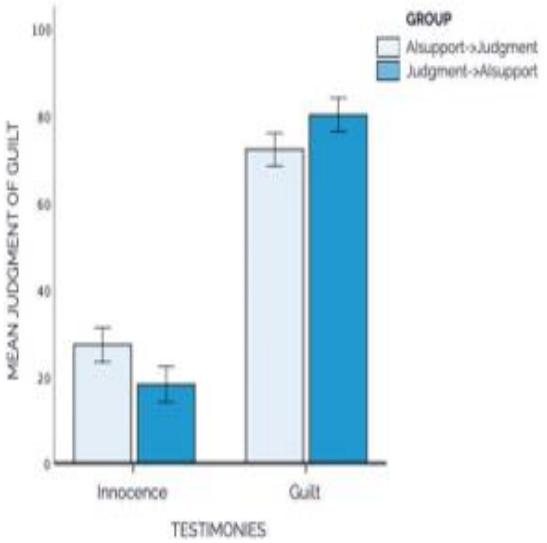

Fig. 6 Juicio de culpabilidad medio en juicios incorrectos, por tipo de testimonios y grupo, en el experimento 2. Nota. Barras de error IC del 95%

Resultados y discusión

Precisión de juicio sin soporte de IA en las pruebas incorrectas.

Primero analizamos la diferencia de los juicios entre grupos cuando los participantes juzgaron a los acusados en las pruebas incorrectas (tres casos incorrectos, uno por cada uno de los tres tipos de delitos). Dado que utilizamos una escala de 0 a 100 para medir los juicios de los participantes en este experimento, nos centramos en la diferencia de juicios entre grupos para analizar la precisión.

[4] La muestra de 52 participantes (69.2% hombres, 28.8% mujeres, 2% no responde) estuvo compuesta por empleados de la misma compañía (una consultora de tecnología española) y todos eran mayores de 18 años (*M* = 35.6, *SD* = 8.92).

La figura 6 resume los resultados. Como se puede ver en esta figura, la precisión media de los juicios en el grupo Juicio→SoporteIA es mayor que en el grupo SoporteIA→Juicio, independientemente de si los participantes consideraron testimonios que apuntaban a la inocencia o la culpabilidad. Estas impresiones fueron confirmadas por un ANOVA mixto 2 (testimonios: inocencia, culpabilidad) × 2 (grupo: SoporteIA→Juicio, Juicio→SoporteIA) con la media de los juicios en las pruebas incorrectas como variable dependiente[5]. Este ANOVA mostró un efecto principal de los testimonios, $F(1, 188)=636.4$, $p<0.001$, $\eta^2_p=0.772$; y ningún efecto principal para el grupo, $F(1, 188)=0.104$, $p=0.747$, $\eta^2_p=0.001$. Sin embargo, y como esperábamos, observamos una interacción de testimonios x grupo, $F(1, 188)=16.3$, $p<0.001$, $\eta^2_p=0.080$.

Las comparaciones *post hoc* posteriores (corrección de Tukey) mostraron que, como esperábamos, los participantes en el grupo Juicio→SoporteIA juzgaron a los acusados en los casos incorrectos como menos culpables cuando los testimonios indicaban inocencia ($M=18.1$, $SD=15.2$, $t(369)=-3.26$, $p=0.007$, $d=-0.53$) que los participantes en el grupo SoporteIA→Juicio ($M=28.2$, $SD=22.5$). Además, cuando los testimonios indicaron culpabilidad, el grupo Juicio→SoporteIA juzgó que los acusados en los casos incorrectos eran más culpables ($M=79.8$, $SD=15.7$, $t(369)=2.83$, $p=0.025$, $d=0.48$) que el grupo SoporteIA→Juicio ($M=70.8$, $SD=21.0$). Estos resultados indican que los participantes en el grupo Juicio→SoporteIA fueron más precisos en sus juicios. Así, parece que aquellos participantes que emitieron sus juicios después de haber recibido el apoyo incorrecto de la IA se vieron influidos negativamente por este, y su precisión se vio reducida.

Precisión de juicio sin soporte de IA en las pruebas correctas

A continuación, analizamos los juicios de los participantes en las pruebas correctas (seis de los nueve casos utilizados en este experimento). En el experimento 1, encontramos un resultado opuesto en las pruebas correctas en comparación con el ensayo incorrecto: los participantes en el grupo SoporteIA→Juicio fueron más precisos en estos ensayos que aquellos en el grupo Juicio→SoporteIA.

Debido a que cambiamos la escala de juicio en el Experimento 2 para tener una medida más sensible, ahora pudimos realizar un ANOVA mixto de 2 (testimonios: inocencia, culpa) × 2 (grupo: SoporteIA→Juicio, Juicio→SoporteIA) con los juicios medios en las pruebas correctas como variable dependiente. Encontramos un efecto principal de los testimonios, $F(1, 249)=2453.4$, $p<0.001$, $\eta^2_p=.908$); sin efecto principal del grupo, $F(1, 249)=0.43$, $p=0{,}515$, $\eta^2_p=0.002$, y una interacción de testimonios x grupo, $F(1, 249)=5.28$, $p=0.022$, $\eta^2_p=0.021$. Para buscar posibles diferencias entre grupos cuando los testimonios indicaban inocencia y cuando indicaban culpabilidad, realizamos comparaciones *post-hoc*, con corrección de Tukey. Estos no mostraron diferencias estadísticamente significativas entre grupos cuando los testimonios indicaron inocencia ($t(491)=1.28$, $p=0.575$, $d=0.16$; Juicio→SoporteIA, $M=18.9$, $SD=12.8$; grupo SoporteIA→Juicio, $M=16.9$, $SD=11.5$), ni cuando los testimonios indicaron culpabilidad ($t(491)=-2.15$, $p=0.139$, $d=0.23$; Juicio→SoporteIA, $M=77.8$, $SD=16.1$; grupo SoporteIA→Juicio, $M=81.3$, $SE=13.8$). Por lo tanto, no encontramos diferencias entre los grupos cuando los participantes emitieron sus juicios sin el apoyo de la IA en las pruebas correctas. Parece que la diferencia entre grupos observada en las pruebas correctas del experimento 1 desaparece cuando se utilizan más casos y una escala más estandarizada y sensible en este experimento.

Conformidad con la evaluación de la IA en las pruebas incorrectas

También analizamos la conformidad de los participantes con la evaluación incorrecta de la IA. Clasificamos a los participantes que están conformes con la evaluación de IA si mostraron conformidad en al menos una de las tres pruebas incorrectas. Esperábamos que el grupo Juicio→SoporteIA mostrara menos conformidad que el grupo SoporteIA→ Juicio. Aunque no habíamos encontrado esta diferencia entre los grupos en el experimento 1 debido a la baja tasa de conformidad en ambos grupos en ese estudio, esperábamos observar este resultado en el experimento 2, a medida que aumentamos el tamaño de la muestra y el número de pruebas incorrectas.

[5] Las tres pruebas incorrectas de cada participante se presentaron aleatoriamente con algunas de ellos que sugerían inocencia y otras culpabilidad. Esta aleatoriedad implica que algunos participantes recibieron testimonios de inocencia en los tres casos incorrectos, otros de culpabilidad en los tres casos, y algunos con testimonios de inocencia y otros de culpabilidad. Esto se refleja en los grados de libertad del ANOVA. En el grupo SoporteIA→Juicio, el número total de juicios con

Realizamos una prueba de chi-cuadrado para analizar la diferencia entre grupos en la conformidad de los participantes. No encontramos una diferencia estadísticamente significativa respecto a conformidad entre los grupos, χ^2 (1) = 0,37, p = 0.545, V de Cramer = 0.04. En SoporteIA→Juicio, el 24.2% de los participantes mostraron conformidad en al menos una de las tres pruebas incorrectas en que la evaluación de IA fue incorrecta (32 participantes de 132); un porcentaje muy similar se observó en el grupo Juicio→SoporteIA, donde el 21.1% de los participantes mostraron conformidad (27 participantes de 128).

Por último, aunque no encontramos diferencias estadísticas entre los grupos de conformidad, debemos señalar que el soporte erróneo de la IA volvió a afectar negativamente la precisión de las decisiones en este experimento. Comparamos el porcentaje de participantes que acertaron en los tres casos en los que recibieron la evaluación de IA incorrecta (en el paso 1 en el grupo SoporteIA→Juicio y en el paso 2 en el grupo Juicio→SoporteIA), y encontramos que solo el 29.5 de participantes en el grupo SoporteIA→Juicio (39 de 132) y el 31.3% de los participantes en el grupo Juicio→SoporteIA (40 de 128) fueron precisos. No hubo diferencias significativas en la precisión entre los grupos en su paso asistido por IA, χ^2 (1) = 0.09, p = 0.765, V de Cramer = 0.02. Nuevamente, aunque los participantes en el grupo Juicio→SoporteIA emitieron juicios más precisos que el otro grupo durante el paso en que se dio el soporte de IA (que para ellos tuvo lugar al comienzo de la tarea), cuando recibieron apoyo de IA incorrecto, la precisión de sus decisiones se vio perjudicada al nivel de la del otro grupo.

Discusión general

Aunque no existe un consenso claro sobre si los sistemas de decisión automatizados con procesos de intervención humana contribuyen a una mejor toma de decisiones, su uso está aumentando en muchos campos diferentes, incluido el sector público. La presente investigación fue diseñada para comprender mejor cómo un sistema de decisión automatizado puede impactar las decisiones en el contexto legal. Nuestros experimentos sugieren que el juicio humano puede verse afectado dependiendo del momento en que se recibe apoyo de la IA.

En los casos en los que la valoración de la IA es incorrecta, parece que el veredicto humano será más preciso si se emite antes de recibir el apoyo erróneo de la IA. Tanto en el experimento 1 como en el experimento 2, encontramos que cuando los participantes emitieron su juicio antes de recibir el apoyo de IA incorrecto (paso 1 en el grupo Juicio→SoporteIA), su juicio estaba más cerca del indicado por los testimonios que cuando los participantes emitieron su juicio después de haber recibido apoyo erróneo de IA (paso 2 en el grupo SoporteIA→Juicio).

En los casos en los que la evaluación de la IA es correcta, descubrimos que, aunque recibir el apoyo correcto de la IA en un paso anterior parecía hacer que los juicios en el grupo SoporteIA→Juicio fueran más precisos que aquellos en el grupo Juicio→SoporteIA en el experimento 1, esta diferencia no fue estadísticamente significativa en el experimento 2, que incluyó una muestra más grande y heterogénea, un mayor número de pruebas y una escala más sensible para medir los juicios que el experimento 1. Esto sugiere que el apoyo correcto de la IA en un proceso con intervención humana no es tan beneficioso como podría parecer, mientras que el soporte incorrecto de la IA es fundamental porque aumenta el error humano. Por lo tanto, nuestros experimentos muestran un posible efecto de anclaje del apoyo incorrecto de la IA en la decisión humana. Recibir el apoyo incorrecto al comienzo del proceso perjudicó el juicio explícito posterior de los participantes en el grupo SoporteIA→Juicio.

También cabe señalar que no encontramos excesiva conformidad de los participantes con el apoyo erróneo de la IA en cualquiera de los experimentos. Es posible que los participantes fueran conscientes de la incongruencia entre el veredicto indicado por los testimonios y la evaluación incorrecta de la IA, lo que les evitó el sesgo de automatización. Este resultado, que va en la línea de trabajos más recientes sobre este tema (De-Arteaga *et al.,* 2020; Grgic-Hlaca *et al*., 2019; Portela *et al*., 2022), contrasta con la alta conformidad que algunos sistemas automatizados, como RisCanvi, se muestran fuera del laboratorio (por encima del 95%; Saura & Aragó, 2021). Se necesitan esfuerzos de investigación futuros para comprender

con testimonios de inocencia y algunos de culpabilidad. Esto se refleja en los grados de libertad de ANOVA. En el grupo SoporteIA→Juicio, el total de número de pruebas con testimonios de inocencia presentados a los participantes fue n=111; en el grupo Juicio→SoporteIA el número de pruebas con testimonios de inocencia presentados a los participantes fue n=112 y el número de pruebas con testimonios que sugerían culpabilidad fue n=110.

tales discrepancias y su impacto. Es importante destacar que esta falta de conformidad no implicó una decisión más acertada. En ambos grupos, en el paso en el que debían confirmar o modificar la evaluación incorrecta de la IA, la tasa de precisión fue baja, incluso para el grupo que previamente se había visto obligado a emitir su juicio y lo había hecho con precisión (es decir, el grupo Juicio→SoporteIA). Estos resultados sugieren que, si bien puede parecer aconsejable que las personas involucradas en procesos de intervención humana informen explícitamente sus juicios al principio, en lugar de simplemente supervisar la IA y confirmar o modificar las evaluaciones de la IA, los errores de la IA es muy probable que comprometan su decisión final incluso cuando esos errores de IA ocurren después de un juicio humano preciso. Por lo tanto, probablemente sea mejor que el juicio humano ocurra primero y luego la IA, en vez que el humano proporcione una segunda opinión, para detectar (y advertir) de posibles errores humanos. Pero claro, seguiremos necesitando un tercero (un auditor humano externo o un comité humano) que tenga la última palabra y sea capaz de analizar críticamente cualquier posible discrepancia en esta colaboración entre humanos y la IA.

Es importante señalar que nuestros experimentos tuvieron como objetivo simular sistemas del mundo real, como RisCanvi, para recrear el proceso estándar de este tipo de sistemas automatizados. No fueron experimentos para evaluar RisCanvi. De hecho, RisCanvi tiene más intervención humana en el proceso de decisión (por ejemplo, se requiere de un humano al momento de seleccionar la información que el sistema deberá considerar o no y cuando un funcionario de gobierno decide modificar el riesgo estimado por el sistema, lo cual requiere validación final por parte de una persona diferente; Portela & Álvarez, 2022), que otros sistemas automatizados de decisión conocidos (como COMPAS o el sistema automatizado de control fronterizo de Frontex; Portela & Álvarez, 2022). Estas diferencias en la cantidad y el propósito de la intervención humana en los diversos sistemas de decisión automatizados existentes pueden afectar la precisión de la toma de decisiones al utilizar esos sistemas, así como la conformidad humana con el soporte del sistema.

En nuestros experimentos, probablemente, habríamos podido obtener niveles más altos de conformidad por parte de los participantes, por ejemplo, introduciendo testimonios más ambiguos en los casos para que el veredicto fuera menos obvio. Sin embargo, nuestro objetivo no era alcanzar un alto nivel de conformidad. Decidimos eliminar la ambigüedad porque si incluso cuando los materiales son muy fáciles y el veredicto es obvio, los participantes se dejarían engañar por el soporte erróneo de la IA, entonces tendríamos evidencia clara de un problema grave, con personas siguiendo los errores de la IA incluso en los casos en que podrían resolverlo fácilmente por sí solos.

Los experimentos actuales muestran que ciertos detalles en la interacción entre humanos y la IA, como el momento de presentación de la evaluación de la IA y si se pide o no a los humanos que emitan sus juicios explícitamente, pueden tener un impacto importante en las decisiones. Curiosamente son detalles que no se suelen tener en cuenta cuando se evalúa la conveniencia de implementar estos sistemas, o cuando estos algoritmos son auditados, interna o externamente, ya que en estos casos el foco suele ponerse en los aspectos técnicos del desempeño del algoritmo en sí y no en su interacción con las personas que lo utilizan (Buçinca *et al*., 2021; Verde, 2022). Nuestra investigación destaca la necesidad de considerar no solo los aspectos técnicos, sino, lo más importante, la interacción humano-IA al evaluar o auditar estos sistemas, porque aspectos como el momento en que los funcionarios gubernamentales reciben la evaluación de IA, ya sea que emitan o no su juicio antes de ver la evaluación de la IA, o si son conscientes o no, por ejemplo, de la tasa de error del sistema, puede tener un impacto muy grande en las decisiones tomadas con intervención humana. De hecho, esos aspectos determinarían si el apoyo de la IA prepara las decisiones humanas en la dirección valorada por la IA (cuando la IA actúa primero) o si simplemente se le deja el papel de brindar una segunda opinión después de que el humano ya haya emitido un juicio. Nuestros experimentos también sugieren que es importante que los humanos involucrados en estos procesos tengan las habilidades, experiencia y tiempo para interpretar y gestionar la información proporcionada por el sistema (Ponce, 2022; Portela & Álvarez, 2022), para ser informados sobre la tasa de error de los sistemas que supervisan, y tener la capacidad (y los incentivos) para ser críticos y estar en desacuerdo con las decisiones del sistema cuando sea necesario. Como ya se ha demostrado, a menudo se subestima la influencia del apoyo y la recomendación algorítmica en las decisiones humanas, tanto públicas como privadas (Agudo & Matute, 2021).

Aunque en esta investigación hemos utilizado la precisión como medida para evaluar el proceso de participación humana, las decisiones en el campo de la justicia deben ser justas, corregibles y éticas, entre

otras cosas, además de ser precisas (Green y Chen, 2019b). Por lo tanto, también es necesario considerar todos estos aspectos, y no solo la precisión predictiva del sistema, al determinar qué tan beneficioso sería implementar sistemas de decisión automatizados en el dominio público. En este sentido, creemos que es necesario un análisis crítico sobre la conveniencia de establecer procesos con intervención humana; no porque creamos que es mejor tomar decisiones de forma totalmente automatizada, sino porque consideramos que el papel del ser humano en un proceso automatizado esconde ciertos escollos que es necesario investigar en detalle.

En primer lugar, si bien la supervisión humana se propone como salvaguarda en decisiones automatizadas de alto riesgo (Comisión Europea, 2019), existen cuestiones críticas como las mencionadas anteriormente, relacionadas con la experiencia de estos humanos, el tiempo del que disponen, la responsabilidad y agencia que tienen, su motivación e incentivos, etc., que pueden convertir los procesos humanos en procesos 'cuasi automatizados' donde el humano no aporta casi nada (Wagner, 2019) e incluso proporciona una falsa sensación de seguridad (Ponce, 2022), como sugieren nuestros resultados. De hecho, vale la pena señalar que cuando los sistemas son precisos, generalmente se celebra su éxito, enfatizando el papel crucial del algoritmo en esa precisión. Sin embargo, cuando se equivocan, se culpa a los humanos por su falta de supervisión o su sesgo de automatización.

Es importante destacar que se ha observado que es más difícil para los humanos supervisar y juzgar la precisión de la predicción de un algoritmo que hacer sus propias evaluaciones y predicciones (Green, 2022). Por lo tanto, como se señaló anteriormente, es posible que en lugar de tener humanos supervisando las decisiones de la IA, una mejor estrategia podría ser dejar que los humanos tomen las decisiones, mientras se utilizan herramientas de IA para brindar una segunda opinión y alertar a los humanos de posibles errores humanos que puedan ocurrir, y luego tener un auditor humano o un comité humano que analice cualquier posible discrepancia entre humanos y IA.

Nuestros resultados contribuyen a una cantidad cada vez mayor de evidencia científica que sugiere que, antes de implementar estos sistemas, es necesario considerar qué decisiones tiene sentido automatizar y cómo hacerlo. De ser así, se debe tener en cuenta, por ejemplo, si se trata de una decisión en la que la precisión de los sistemas automatizados es muy superior a la precisión humana, o si, por el contrario, se trata de una decisión donde se deben tener en cuenta una amplia variedad de criterios (como hemos mencionado, en el caso de las decisiones judiciales, que deben ser justas, corregibles y éticas, además de precisas), por lo que no es adecuado el uso de sistemas automatizados. No todas las decisiones son aptas para la automatización, ni parece deseable que como humanos dependamos de procesos de toma de decisiones en los que la persona involucrada en esos procesos sea un claro candidato a ser señalado como el error del sistema.

Declaración de importancia

Los algoritmos de inteligencia artificial y los sistemas automatizados de apoyo a las decisiones son cada vez más comunes en el sector judicial, pero no están exentos de errores y sesgos. Por esta razón, se han propuesto procesos con intervención humana como garantía para salvaguardar la justicia e integridad de tales decisiones. Sin embargo, se sabe poco sobre cómo pueden reaccionar los humanos a los sistemas de apoyo a las decisiones de IA y qué condiciones garantizan que sean capaces de detectar errores de la IA y rechazar el apoyo de la IA cuando sea apropiado. Los experimentos fueron diseñados para comprender mejor cómo un sistema de decisión automatizado puede impactar las decisiones humanas en el contexto legal, y si este proceso puede mejorarse manipulando algunas variables, como el tiempo en el que tienen lugar los diferentes pasos del proceso. Nuestros resultados muestran la importancia de que los humanos que participan en el proceso de decisión informen su juicio antes de recibir apoyo de la IA. En los casos en que la evaluación de la IA sea incorrecta, el juicio humano es más preciso si se emite antes de recibir la evaluación de la IA errónea. Por lo tanto, el momento en que se proporcione el apoyo de la IA puede ser de gran importancia. Nuestros resultados contribuyen a una cantidad cada vez mayor de evidencia científica que sugiere que es importante comprender cómo la IA afecta las decisiones humanas y considerar qué decisiones tiene sentido automatizar y cómo hacerlo.

Contribuciones del autor

Concibió y diseñó el experimento: UA, KGL, HM, MA; Software experimental: UA. Realizó el experimento: UA. Analizados los datos: UA. Escribió el manuscrito: UA, HM, KGL. Revisado el manuscrito: UA, KGL, HM, MA.

Fondos

El apoyo a esta investigación ha contado con la Beca PID2021-126320NB-I00 del Gobierno de España MCIN/AEI/10.13039/501100011033 y con la Beca FEDER (Fondo Europeo de Desarrollo Regional), así como con la Beca IT1696-22 del Gobierno Vasco, ambas concedidas a SM. Los financiadores no tuvieron ningún papel en el diseño del estudio, la recopilación y el análisis de datos, la decisión de publicar o la preparación del manuscrito.

Disponibilidad de datos y materiales

Los datos y materiales para este experimento están disponibles gratuitamente en Open Science Framework: https://osf.io/b6p4z/. Preinscripción El experimento 2 se preinscribió en https://aspredicted.org/ph9br.pdf.

Declaraciones

Aprobación ética y consentimiento para participar

El Comité de Revisión Ética de la Universidad de Deusto aprobó el procedimiento para este experimento. La participación fue anónima y los participantes enviaron sus respuestas de forma voluntaria. No se recopiló información personal.

Consentimiento para publicación

Los autores autorizan a CR:PI a publicar esta investigación.

Intereses en competencia

Los autores no declaran tener intereses en competencia.

Recibido: 10 de abril de 2023. Aceptado: 12 de diciembre de 2023. Publicado: 7 de enero de 2024

Referencias

Agudo, U., & Matute, H. (2021). The influence of algorithms on political and dating decisions. PLoS ONE, 16(4), e0249454. https://doi.org/10.1371/ journal.pone.0249454

Alon-Barkat, S., & Busuioc, M. (2022). Human-AI interactions in public sector decision-making: 'Automation Bias' and 'Selective Adherence' to algorithmic advice. *Journal of Public Administration Research and Theory*. https://doi.org/10.1093/JOPART/MUAC007

Álvarez, M., Martínez, N., Agudo, U., & Matute, H. (2023). ForenPsy 1.0. Retrieved from https://osf.io/detn4/

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Araujo, T., Helberger, N., Kruikemeier, S., de Vreese, C. H., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, 35(3), 1–13. https://doi.org/10.1007/ S00146-019-00931-w

Berkman Klein Center. (2022). Risk assessment tool database. Berkman Klein Center. https://criminaljustice.tooltrack.org/

Binns, R., & Veale, M. (2021). Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR. *International Data Privacy Law*, 11(4), 319–332. https://doi.org/10.1093/IDPL/IPAB020

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21. https://doi.org/10.1145/3449287

Capdevila, M., Blanch, M., Ferrer, M., Pueyo, A., Framis, B., Comas, N., Garrigós, A., Boldú, A., Batlle, A., & Mora, J. (2015) Tasa de reincidencia penitenciaria 2014. Centre d'Estudis Jurídics y Formació Especialitzada de la Generalitat de Catalunya. https://cejfe.gencat.cat/web/.content/home/recerca/cataleg/crono/2015/taxa_reincidencia_2014/tasa_reincidencia_2014_cast.pdf

Casacuberta, D., & Guersenzvaig, A. (2018). Using Dreyfus' legacy to understand justice in algorithm-based processes. AI & Society, 1–7. https://doi. org/10.1007/s00146-018-0803-2

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. Computers in Human Behavior, 127, 107018. https://doi.org/10.1016/J.CHB. 2021.107018

Civio. (2022). *La Justicia impide la apertura del código fuente de la aplicación que concede el bono social*. https://civio.es/novedades/2022/02/10/la-justicia-impide-la-apertura-del-codigo-fuente-de-la-aplicacion-que-concede-el-bono-social/

Cummings, M. (2004). Automation bias in intelligent time critical decision support systems. In AIAA 1st Intelligent systems technical conference. American institute of aeronautics and astronautics. https://doi.org/10.2514/6.2004-6313

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in- the-loop: Decisions in the presence of erroneous algorithmic scores. In *2020 CHI Conference on human factors in computing systems*, 1–12. https://doi.org/10.1145/3313831

Duncan, P., McIntyre, N., & Levett, C. (2020). Who won and who lost: when A-levels meet the algorithm. *The Guardian*. https://www.theguardian.com/education/2020/aug/13/who-won-and-who-lost-when-a-levels- meet-the-algorithm

Echterhoff, J. M., Yarmand, M., & McAuley, J. (2022). AI-moderated decision-making: Capturing and balancing anchoring bias in sequential decision tasks. In *Proceedings of the 2022 CHI conference on human factors in computing systems (CHI '22),* 161, 1–9. https://doi.org/10.1145/3491102.3517443

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. Psychological Science, 17(4), 311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Commission. (2019). Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trust worthy-ai

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law and Security Review,* 45. https://doi.org/ 10.1016/j.clsr.2022.105681

Green, B., & Chen, Y. (2019a). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Conference on Fairness, accountability, and transparency*, 90–99. https://doi.org/10.1145/3287560.3287563

Green, B., & Chen, Y. (2019b). The principles and limits of algorithm-in-the-loop decision making. ACM on Human-Computer Interaction, 3(CSCW). https:// doi.org/10.1145/3359152

Green, B., & Chen, Y. (2021). Algorithmic risk assessments can alter human decision-making processes in high-stakes Government Contexts. *ACM on Human–Computer Interaction,* 5(CSCW2). https://doi.org/10.1145/34795 62

Grgic-Hlaca, N., Engel, C., & Gummadi, K. P. (2019). Human decision making with machine assistance: An experiment on bailing and jailing. SSRN *Electronic Journal*. https://doi.org/10.2139/SSRN.3465622

Lambe, K. A., O'Reilly, G., Kelly, B. D., & Curristan, S. (2016). Dual-process cognitive interventions to enhance diagnostic reasoning: A systematic review. *BMJ Quality & Safety,* 25(10), 808–820. https://doi.org/10.1136/bmjqs-2015-004417

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

López-Ossorio, J. J., González Álvarez, J. L., & Andrés Pueyo, A. (2016). Eficacia predictiva de la valoración policial del riesgo de la violencia de género. *Psychosocial Intervention*, 25(1), 1–7.

https://doi.org/10.1016/J.PSI.2015.10.002
Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. Journal of the American Medical Informatics Association, 24(2), 423–431. https://doi.org/10.1093/jamia/ocw105
Lyell, D., Magrabi, F., Raban, M. Z., Pont, L. G., Baysari, M. T., Day, R. O., & Coiera, E. (2017). Automation bias in electronic prescribing. *BMC Medical Informatics and Decision Making*, 17(1). https://doi.org/10.1186/S12911-017-0425-5
Marquardson, J., & Grimes, M. (2018). Supporting better decisions: How order effects influence decision support system alignment. *Interacting with Computers*, 30(6), 469–479. https://doi.org/10.1093/iwc/iwy022
Martínez-Garay, L. (2016). Errores conceptuales en la estimación de riesgo de reincidencia: La importancia de diferenciar sensibilidad y valor predictivo, y estimaciones de riesgo absolutas y relativas. *Revista Española De Investigación Criminológica*, 14, 1–31.
Ministerio Público Fiscal de la Ciudad Autónoma de Buenos Aires. (2020). Innovación e inteligencia artificial. https://mpfciudad.gob.ar/institucional/2020-03-09-21-42-38-innovacion-e-inteligencia-artificial
Ministerstwo Sprawiedliwości. (2021). Algorytm SLPS. https://www.gov.pl/web/sprawiedliwosc/algorytm towards-a-meaningful-human-oversight-of-automated-decision-making-systems/
Ministry of Justice. (2013). Offender assessment system (OASys). Data.Gov.Uk. https://www.data.gov.uk/dataset/911acd3c-495f-48ca-88b6-024210868b 06/offender-assessment-system-oasys
Mosier, K. L., & Manzey, D. (2019). Humans and automated decision aids: A match made in heaven? Human performance in automated and autonomous systems, 19–42. https://doi.org/10.1201/9780429458330-2
Niiler, E. (2019). Can AI Be a Fair Judge in Court? Estonia thinks so. WIRED. https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax23402
O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishing Group.
OpenAI. (2023). ChatGPT (Jun 22 version) [Large language model]. https://chat.openai.com/chat
Parasuraman, R., & Mustapha, M. (1996). Automation and human performance. CRC Press.
Ponce, J. (2022). Reserva de humanidad y supervisión humana de la inteligencia artificial. El Cronista Del Estado Social y Democrático De Derecho, 100, 58–67.
Portela, M., & Álvarez, T. (2022). Towards a meaningful human oversight of automated decision-making systems. Digital Future Society. https://digitalfuturesociety.com/report/
Portela, M., Castillo, C., Tolan, S., Karimi-Haghighi, M., & Pueyo, A. A. (2022). *A comparative user study of human predictions in algorithm-supported recidivism risk assessment*. arXiv. https://doi.org/10.48550/arxiv.2201.11080
Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., & Mullainathan, S. (2019). *The algorithmic automation problem: Prediction, triage, and human effort*. ArXiv. https://doi.org/10.48550/arXiv.1903.12220
Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *Proceedings of the ACM on Human–computer Interaction*, 6(CSCW1), 1–22. https://doi.org/10.1145/3512930
Saura, G., & Aragó, L. (2021). Un algoritmo impreciso condiciona la libertad de los presos. *La Vanguardia*. https://www.lavanguardia.com/vida/20211206/7888727/algoritmo-sirve-denegar-permisos-presos-pese-fallos.html
Skeem, J., Scurich, N., & Monahan, J. (2020). Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law and Human Behavior*, 44(1), 51–59. https://doi.org/10.1037/LHB0000360
Solans, D., Beretta, A., Portela, M., Castillo, C., & Monreale, A. (2022). *Human response to an AI-based decision support system: A user study on the effects of accuracy and bias*. arXiv. https://doi.org/10.48550/arXiv.2203.15514
Soler, C. (2013). *RisCanvi. Protocolo de evaluación y gestión del riesgo de violencia con población penitenciaria* [PowerPoint slides]. Slideplayer. https://slideplayer.es/slide/7242758/
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157), 1124–1131. https://doi.org/10.1126/scien ce.185.4157.1124
Valdivia, A., Hyde-Vaamonde, C., & García-Marcos, J. (2022). *Judging the algorithm: A case study on the risk assessment tool for gender-based violence implemented in the Basque country*. arXiv. https://doi.org/10.48550/arXiv. 2203.03723
Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports,* 13, 15737. https://doi.org/10.1038/s41598-023-42384-8
Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet*, 11(1), 104–122. https://doi.org/10.1002/poi3.198
Wei, J. (2019). China uses AI assistive tech on court trial for first time. *China-Daily*. https://www.chinadaily.com.cn/a/201901/24/WS5c4959f9a3106 c65c34e64ea.html

Publisher's Note