INFLUENCIA DE LOS ALGORITMOS DE INTELIGENCIA ARTIFICIAL EN LAS DECISIONES HUMANAS EN EL ÁMBITO JUDICIAL

Experimentos en el contexto de un jurado popular simulado

PROGRAMA DE DOCTORADO: PSICOLOGÍA

Tesis doctoral realizada por Mario Álvarez Lafuente

Dirigida por la Dra. Helena Matute Greño

Bilbao, 5 de mayo de 2025

Deusto
Universidad de Deusto

Deusto
Universidad de Deusto

INFLUENCIA DE LOS ALGORITMOS DE INTELIGENCIA ARTIFICIAL EN LAS DECISIONES HUMANAS EN EL ÁMBITO JUDICIAL

Experimentos en el contexto de un jurado popular simulado

PROGRAMA DE DOCTORADO: PSICOLOGÍA

Tesis doctoral realizada por Mario Álvarez Lafuente

Dirigida por la Dra. Helena Matute Greño

El doctorando

Mario Álvarez Lafuente

Firmado digitalmente por Mario Álvarez Lafuente
Fecha: 2025.04.30 17:26:28 +02'00'

La directora

Helena Matute

Firmado digitalmente por Helena Matute
Fecha: 2025.04.30 18:05:25 +02'00'

Bilbao, 5 de mayo de 2025

Esta tesis se ha realizado en el marco del proyecto de investigación PSI2016-78818-R concedido a la Dra. Helena Matute por la Agencia Estatal de Investigación, el proyecto PID2021-126320NB-I00 concedido a los Dres. Helena Matute y Fernando Blanco, financiado por la Agencia Estatal de Investigación (AEI/ 10.13039/501100011033) y por FEDER Una manera de hacer Europa, así como la ayuda a equipos de investigación IT1696-22 concedida por el Departamento de Educación, Universidades e Investigación del Gobierno Vasco, a la Dra. Helena Matute. Además, Mario Álvarez ha recibido una Ayuda para la Formación de Personal Investigador (referencia BES-2017-081118), concedida por la Agencia Estatal de Investigación del Gobierno de España.

El Camino sigue y sigue
desde la puerta.
El Camino ha ido muy lejos,
y si es posible he de seguirlo
recorriéndolo con pie decidido
hasta llegar a un camino más ancho
donde se encuentran senderos y cursos.
¿Y de ahí a dónde iré? No podría decirlo.

La Canción de los Caminantes (Tolkien, 1954)

Agradecimientos

Esta tesis doctoral es para mí algo más que un trabajo académico. Empezar este doctorado fue toda una aventura para mí. Nunca había vivido fuera de Sevilla, ni siquiera había vivido independizado y, sin conocer a nadie y dispuesto a aprender todo lo posible, me vine a la otra punta de la península, más allá de Despeñaperros, cual pequeño y temeroso hobbit que no sabe lo que le depararía el camino, pero expectante y deseoso de nuevas vivencias que contar.

Dejé atrás a mucha gente, más nunca me abandonaron. Desde allí, fueron uno de mis pilares de apoyo. Sois muchos e intentaré nombraros a todos. En primer lugar, a mis padres Chari y Pepe y a mis hermanos Jose y Mari. También a mis cuñados Esther y Miguel y a todo el resto de mi familia. Aunque os daba miedo que me marchara lejos, nunca me frenasteis y me apoyasteis en todo momento. Muchas gracias. Os quiero.

No puedo olvidarme de mi otra familia, mi clan, mi gente rara. Me habéis aguantado mucho y nunca habéis fallado cuando os he reclamado: Lucas, Marga, Fati, Antonio, Bara, Dani, Fátima, Migue, Adri, Manuela, Lula. Gracias por el abrazo constante.

Al llegar a Bilbao me encontré desamparado al principio, pero encontré a muchísima gente en estos años que se han convertido en gente muy importante. Empezando por mis compañeros de LabPsico: Fer, Nela, Naroa, Marcos, Ujué, Aranza y Lucía. Muchos sois más que compañeros de trabajo, llegando a ser amigos e incluso compañeros de piso y os tengo que dar las gracias a todos por todos estos años. Conseguisteis que el día a día en el laboratorio (y fuera de él) fuera estupendo.

En Bilbao conocí a otras personas que son mi otro pilar hasta el día de hoy: Parmeeta, Mikel, Yule, Rebe, Antonio. Gracias por acogerme.

También hay muchos más a los que les tengo que dar las gracias por haber estado también ahí: a los Gatitos y Psicomemorias, los Camaradas Meditabundos y la Escisión, a los del Covent y a los de Meetup. Gracias a todos.

Y a quienes fueron parte del camino y, aunque ya no estén, dejaron su huella en mí: Vanessa, Óscar, Laura, Cris. También os lo debo a vosotros.

Por supuesto, mi gratitud infinita a Helena, por confiar en mí, por darme la oportunidad de formar parte de su equipo y por guiarme en este camino que, lejos de terminar, solo acaba de empezar. Para bien o para mal, siempre me recordarás como tu último alumno de doctorado. Gracias por todo.

Y, por último, a ti, Naira. Llegaste a mi vida por un cúmulo de casualidades casi imposibles y, desde entonces, has sido mi Constante. Esta aventura termina, pero ya sabemos cuál es la siguiente. ¿Y después? ¡Quién sabrá!

Índice

Aspectos éticos y ciencia abierta

La metodología de investigación utilizada en esta tesis fue aprobada por el Comité de Ética en la Investigación de la Universidad de Deusto (Ref: ETK-7/18-19) al considerar que se ajusta a los principios éticos y jurídicos exigidos y no conlleva ningún riesgo para los participantes. La participación de los participantes en los experimentos fue anónima y voluntaria y sus respuestas fueron recogidas vía internet con su permiso explícito, al aceptar su envío al final de cada experimento. Ninguna información personal fue recopilada.

Los datos en bruto de todos los experimentos de esta tesis se encuentran disponibles en abierto y pueden ser descargados desde Open Science Framework: https://osf.io/wr85g/?view_only=c09ad13968fc456a8c0fa991ee25758d

Resumen

Hablar de algoritmos y de inteligencia artificial (IA a partir de ahora) hace mucho que dejó de ser únicamente algo que leer en novelas de ciencia ficción. Los algoritmos forman parte de nuestra vida y nos influyen en muchos aspectos, desde elegir una canción o comprar un producto hasta que nos acepten un seguro privado o que nuestro barrio merezca más o menos presencia policial, siendo el ámbito judicial uno de los contextos donde están siendo cada vez más frecuentes. Sin embargo, los algoritmos pueden verse afectados por los sesgos de los seres humanos; y la confianza ciega en un algoritmo sesgado puede afectar a la toma de decisiones, que en justicia puede significar condenar erróneamente a una persona.

Esta tesis busca comprobar si la toma de decisiones basada en algoritmos sesgados o con errores puede, efectivamente, afectar a las decisiones tomadas por miembros de un jurado y encontrar alguna forma de minimizar esta influencia. Para ello, se presentarán una serie de experimentos realizados con el objetivo de comprobar si la recomendación errónea de un algoritmo puede afectar al veredicto emitido por los miembros de un jurado simulado (personas de la población general que podrían llegar a ser parte de un jurado popular), así como identificar otras variables que pueden afectar a estas decisiones, como el orden en el que se presenta la información o el hecho de presentar o no información a los participantes sobre la posibilidad de error en los algoritmos.

Parte I. Introducción

Capítulo 1. La Era de la Inteligencia Artificial

Desde los años 80 del siglo pasado, el imaginario colectivo ha estado lleno de ideas sobre cómo sería el mundo una vez que las IAs llegasen a ser una realidad. Películas como *The Terminator* (Cameron, 1984) o *WarGames* (Badham, 1983) nos muestran IAs (Skynet en *The Terminator* y Joshua en *Wargames*) que controlan arsenales militares que se vuelven en nuestra contra; mientras que otras como *Blade Runner* (Scott, 1982) o *Short Circuit* (Badham, 1986) presentan IAs con deseos, ambiciones y miedos que los acercan a los humanos.

Con lo planteado en estos títulos, y los cientos que vinieron después que trataban el tema, era lógico que todos esperáramos que durante los primeros años del nuevo siglo viviéramos el nacimiento y desarrollo de las IAs. Y ha pasado. No del modo catastrofista que nos planteaba el cine, pero sí que lo han hecho de una forma más silenciosa. Como algoritmos de IA.

Podemos encontrar algoritmos de IA en prácticamente la totalidad de los aspectos de nuestra vida. Por ejemplo, cuando usamos redes sociales como *Facebook* o *X* (anteriormente conocida como *Twitter*), hay un algoritmo de inteligencia artificial que analiza y decide qué información mostrarnos y recomendarnos (Diakopoulos & Koliska, 2017). Si usamos *Spotify*, también tenemos un algoritmo que nos recomienda música nueva en función de la que solemos escuchar (Bovenkamp, 2017). Si en algún momento hemos comprado en *Amazon*, seguro que nos llegan mensajes, correos o *banners* (piezas de publicidad integrada en páginas web en función de nuestras búsquedas en internet) en otras páginas recomendándonos productos que pueden interesarnos. Estas recomendaciones las realiza un algoritmo (Chen et al., 2016). Incluso si alguna vez hemos buscado el amor con alguna

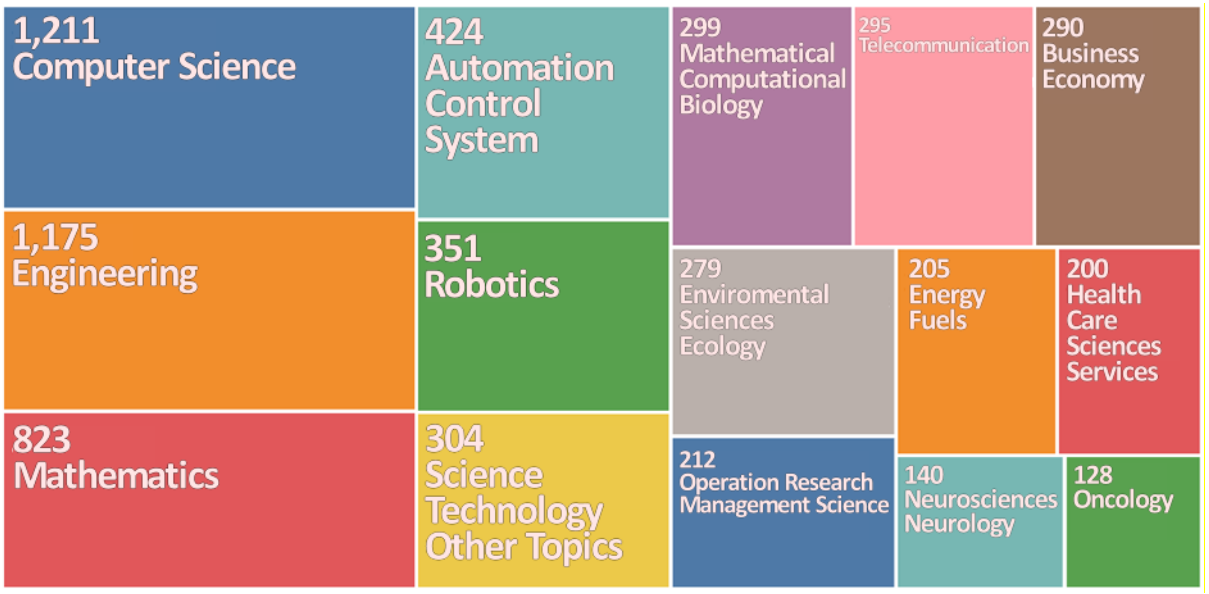
aplicación de citas, es un algoritmo el que decide a quién mostrarnos (Duportail, 2019).

Sin embargo, el uso de algoritmos de IA no se limita a situaciones tan comunes como son recomendaciones a la hora de comprar, encontrar pareja o disfrutar del tiempo de ocio. Estos algoritmos, pueden también denegarnos el acceso a la universidad (Chan, 2018) por considerar que somos un mal ciudadano (Bostman, 2017), decidir que no se nos conceda un seguro médico (Obermeyer et al., 2019) o un préstamo (Cheney, 2016), dictar sentencia en delitos leves (Roberts et al., 2021), decidir si nuestro barrio merece más o menos presencia policial (Lapowsky, 2018), si merecemos obtener la libertad condicional (Angwin et al., 2016) o el nivel de protección que necesitamos si somos víctimas de violencia de género (Instituto de Ciencias Forenses y de la Seguridad, 2018).

Como puede verse, el uso de los algoritmos se ha generalizado a multitud de aspectos de nuestra vida y del día a día. Sin embargo, parece que la mayoría de las investigaciones que se realizan sobre algoritmos están centradas en los aspectos técnicos de los mismos, dejando de lado aspectos como el efecto que tienen en nuestra vida o en nuestra forma de tomar decisiones. En la Figura 1 se muestran los resultados obtenidos por Agudo (2021) en *Web of Science* al realizar una búsqueda de artículos con los términos *Algorithm* y *Decision Making*. En ella podemos ver que la mayoría de los artículos obtenidos en aquel momento están dentro de las áreas de ciencia computacional, ingeniería y matemáticas. Lo más cercano a la psicología que se podía ver en los primeros puestos es neurociencias. Agudo indica que los artículos de Ciencias del Comportamiento podríamos encontrarlos en el puesto 16, mientras que de Psicología encontraríamos únicamente 105 artículos, y ocuparía el puesto 21.

Figura 1

Resultados obtenidos en Web of Science mostrando el número artículos publicados en 2021 que hayan usado los términos Algorithm y Decision Making con fecha de 21 de septiembre de 2021 (tomado de Agudo, 2021)



Si hacemos la misma búsqueda (a fecha de 07 de enero de 2025), podemos darnos cuenta de la importancia que está adquiriendo el tema. En este breve espacio de tiempo, hemos pasado de 6.336 artículos (en las quince categorías con mayor número de artículos) a 96.892 artículos relacionados con algoritmos y toma de decisiones (un incremento de 90.560 artículos), como puede verse en la Figura 2. Las categorías principales han cambiado bastante viendo la diversificación y la extensión del uso de algoritmos en todos los campos de la investigación, haciendo que encontremos ramas de la Psicología tales como Psicología Multidisciplinar (puesto 92 con 202 publicaciones), Psicología Experimental (puesto 95 con 194 publicaciones), Psicología Aplicada (puesto 125 con 111 publicaciones), Psicología (puesto 130 con

104 publicaciones) incluyendo las Ciencias del Comportamiento (en el puesto 135) con 93 publicaciones.

Figura 2

Resultados obtenidos en Web of Science mostrando el número de artículos publicados a día 7 de enero de 2025, que hayan usado los términos Algorithm y Decision Making

Qué es un algoritmo y qué es la IA

Llegados a este punto y antes de continuar, es necesario explicar y analizar una serie de términos; y empezaremos por el que más hemos usado hasta este momento y el que representa el primer paso en el desarrollo de las IAs: algoritmo.

Si somos nuevos en el tema y escuchamos la palabra algoritmo, posiblemente lo primero que se nos viene a la cabeza sea algo relacionado con la informática, pero

el término es mucho más antiguo. Todo se remonta al matemático del siglo IX Abu Abdallah Muḥammad ibn Mūsā al-Jwārizmī (Al-Juarismi en español) que fue quien introdujo los números indo-arábigos en Europa. Cuando su obra se tradujo al latín, se le dio el nombre de *Algorismus* (Chabert, 1999), de donde procede el término y que Ada Lovelace, pionera en el estudio de la computación, comenzó a utilizar ya en el S.XIX para hablar de todo método de cálculo automático (Essinger, 2015). Hasta este momento, el término algoritmo podría entenderse usando la comparativa con una receta de cocina: tenemos una serie de instrucciones y pasos que deben seguirse de una forma determinada para resolver un problema o conseguir un objetivo, como preparar un plato de comida.

Sin embargo, no es hasta 1948 (Jiménez, 2016) cuando nace el concepto moderno de algoritmo en el momento que Claude Shannon, un estudiante del Instituto de Tecnología de Massachusetts, desarrolla la Teoría de la Información y el término adquiere una importancia capital en la informática. A raíz de aquí, el concepto algoritmo comienza a hacer referencia a una serie de instrucciones y normas que se le dan a un ordenador y que este debe seguir para trabajar sobre un conjunto de datos y así resolver una tarea determinada (Lee, 2018).

Esto es lo que entendíamos desde hace unos años como algoritmo: una serie de instrucciones y normas dadas por un ser humano a un ordenador con las que este era capaz de resolver un problema. Pero, en los últimos años, con el avance de la tecnología, se ha empezado a hablar de algoritmos de IA (Duan et al., 2019), algunos de los cuales son incluso capaces de aprender las reglas que tienen que aplicar para resolver el problema sin que ningún ser humano las programe. De hecho, los algoritmos de IA poseen algunas capacidades avanzadas tales como la capacidad de aprender (Eiband et al., 2019). A menudo, también se les permite tomar decisiones

basándose en ese aprendizaje o, cuando menos, ayudar a las personas que deberán tomar la decisión, ofreciéndoles su consejo basado en ese aprendizaje.

Percepción humana de los algoritmos de IA

Una de las principales cualidades que tendemos a atribuir los humanos a los algoritmos de IA es que son completamente objetivos en sus decisiones (Araujo et al., 2020) y que, además, son más eficaces que nosotros a la hora de realizar ciertos tipos de tareas (Sundar & Kim, 2019). Sundar (2008) llamó a todo este proceso de atribución de los estereotipos de objetividad y perfección a los algoritmos *machine heuristic* (heurístico de la máquina). Junto a este término, debemos mencionar otro llamado sesgo de automatización (Mosier et al., 1996) que consiste en la creencia ciega de que la información que nos proporcionan los algoritmos es siempre veraz, llegando a veces a delegar por completo nuestras decisiones en la información que nos proporciona la máquina sin detenernos siquiera a cuestionarlas (Cummings, 2004).

Pero no solo atribuimos a los algoritmos este tipo de características. Al fin y al cabo, no dejan de ser programas, por lo que tendemos a atribuirles también los atributos de frialdad, inflexibilidad, falta de emociones y propensión a ser pirateados (Sundar, 2020). En relación con estos atributos, también nos encontramos el estereotipo de que, para las tareas que requieren juicios subjetivos o requieren capacidades emocionales, los humanos creemos que somos mejores (Agudo et al., 2022; Castelo et al., 2019; Lee, 2018). En algunos experimentos realizados sobre los estereotipos que tenemos de las máquinas (Pan et al., 2007) se encontró que, poniendo de ejemplo a *Google*, las personas tendemos a elegir siempre la primera búsqueda que nos recomienda el buscador a pesar de no ser siempre la más

relevante, lo que sugiere que confiamos en los algoritmos sobre qué información es la más adecuada respecto a la búsqueda que hemos realizado.

Con todas estas ideas y estereotipos sobre los algoritmos afianzadas en la sociedad, parece que ha surgido lo que se ha comenzado a llamar *algorithm appreciation* (Logg et al., 2019) o apreciación del algoritmo, que sería la preferencia por parte de los humanos de las recomendaciones dadas por los algoritmos sobre las de las personas. Esto queda patente en los experimentos de Logg y colaboradores (2019) donde los participantes confiaban más en la información o recomendación que les daba un algoritmo que en la que le proporcionaban otros humanos en tareas tales como determinar el peso de una persona mediante una fotografía, el futuro éxito de una canción en una lista de éxitos o la atracción que una persona sentiría hacia otra.

De todas formas, la apreciación del algoritmo no es absoluta ni funciona siempre de la misma forma. Por ejemplo, uno de los factores que influye en estos experimentos es la persona o grupo de personas con las que se compara el algoritmo (Önkal et al., 2009), de manera que, a veces, los participantes tienen más en cuenta la opinión de los expertos humanos que la del algoritmo. Otro factor que parece afectar es el tipo de tareas sobre las que se trabaja, siendo las tareas mecánicas (Araujo et al., 2020) más propensas a la confianza en el algoritmo que las tareas que requieren habilidades más subjetivas y específicamente humanas. Algunos estudios recientes (Marmolejo et al., 2025) también mencionan que factores tales como poseer conocimientos estadísticos ayudan a controlar la confianza en el algoritmo cuando se está trabajando con tareas de cierta importancia (como contratar a alguien), mientras que en tareas más mundanas (como recomendaciones de restaurantes) estos factores hacen que se confíe más en los algoritmos.

Con la información que ya tenemos, quedaría preguntarnos si son de verdad los algoritmos mejores que nosotros en todas esas habilidades que les atribuimos.

Los algoritmos de IA tienen sesgos

Solemos pensar que un algoritmo de IA es un programa aislado que lleva a cabo sus decisiones de manera neutral y objetiva (Kahneman et al., 2016), pero esto no es en absoluto así. Un algoritmo será tan objetivo como lo son el contexto o sistema social, cultural y político en el que se ha creado (Birhane, 2019), influyendo también los sesgos (entendidos como distorsiones o errores sistemáticos de juicio que se producen en la interpretación de la información o de la realidad) que presenten las personas que lo han diseñado y los que intervienen en su proceso de creación y desarrollo; además de la utilización que le dan los usuarios. Estos sesgos pueden ser incluidos en el sistema en cualquier momento, ya sea en la recogida o etiquetado de los datos que se usarán para entrenar al algoritmo, en el entrenamiento del mismo, en el análisis de la información final que proporciona el algoritmo (Xu & Doshi, 2019) e incluso en su interacción con los usuarios finales si se le permite al algoritmo seguir aprendiendo en esa fase.

Uno de los algoritmos del que más se han estudiado los sesgos que presenta y del que hablaremos luego más detalladamente es el algoritmo *COMPAS* (Angwin et al., 2016), usado en Estados Unidos para calcular el riesgo de reincidencia que tienen las personas solicitantes de libertad condicional. Esta información, presuntamente objetiva, presenta un sesgo racial que atribuye mayor riesgo de reincidencia a personas negras o latinas que a personas blancas (Larson et al., 2016). Otro caso, quizás más conocido por la opinión pública, es el de *Tay* (Vincent, 2016), una IA creada por *Microsoft* a la que se le dio el control de una cuenta de *Twitter* para estudiar cómo interactuaba y aprendía mediante las conversaciones con el resto de los

usuarios. En menos de un día (ver Figura 3), *Tay* aprendió a comportarse de forma racista, publicar mensajes de apoyo a Hitler y se volvió partidaria del genocidio, teniendo que ser retirada. El problema de *Tay* sería un caso de aprendizaje por reforzamiento del algoritmo en las redes sociales, ya que este aprendizaje erróneo se ha desarrollado en el momento que el algoritmo comenzó a interactuar con otros usuarios, algo que hoy en día no suele hacerse ya de manera tan abierta.

Figura 3

Capturas de Pantalla del Primer Tuit de la IA Tay de Microsoft y de un Tuit Pasadas Menos de 24 Horas

Otro tipo de sesgo sucede debido al tipo de información que se le proporciona al algoritmo durante el entrenamiento inicial. Por ejemplo, si queremos que aprenda a diferenciar entre un husky siberiano y un lobo (Ribeiro et al., 2016) tenemos que asegurarnos de darle material que no pueda producir confusión entre variables, pero esto suele ser difícil, y los errores suelen detectarse casi siempre a posteriori. Por ejemplo, cuando se diseñó un algoritmo para hacer esta tarea, todas las imágenes que se le proporcionaron para que aprendiera a distinguir lobos resultaron estar hechas en entornos nevados, por lo que el algoritmo lo que aprendió fue que, si había nieve en la

foto, era un lobo. Si no le damos material para que se evite la confusión entre variables, tendremos un algoritmo que detecta que un animal es un lobo solo analizando si el fondo es blanco, como puede verse en la Figura 4, lo que podría llevarle a grandes errores

Figura 4

Imagen proporcionada al algoritmo (panel izquierdo) y fragmentos que se observó que analizaba para identificar al animal (panel derecho)

Algo que debemos tener en cuenta es que estos sesgos, por lo general, son accidentales, pero también se han dado casos donde el algoritmo había desarrollado sesgos con el fin de maximizar los beneficios de la empresa. Por ejemplo, el algoritmo de asignación de asientos de algunas aerolíneas tenía como objetivo maximizar los beneficios y aprendió que separar a las personas que habían comprado los billetes juntos hacía que pagaran por sentarse juntos (Rogers, 2017).

En estos dos últimos casos, que se detectaron a posteriori, no era posible predecir estos sesgos por un problema que presentan la mayoría de los algoritmos de IA que discutiremos a continuación: funcionan como cajas negras.

Qué aprenden los algoritmos de IA

La mayoría de los algoritmos de IA funcionan como sistemas de caja negra (Burrell, 2016), donde desconocemos su proceso de aprendizaje e incluso el resultado del mismo. Es el caso de los algoritmos de aprendizaje automático, como el ejemplo del algoritmo de los lobos y perros (Ribeiro et al., 2016) que discutíamos anteriormente, donde solo investigándolos o auditándolos a posteriori podremos, y no siempre, hacernos a la idea de lo que realmente está haciendo el algoritmo. Es decir, el principal problema con estos algoritmos es que, con frecuencia, conforme van desarrollando sus propias reglas de aprendizaje, se vuelven tan complejos y abstractos que hace que no seamos capaces de comprenderlos (Challen et al., 2019; Merino, 2018). Este es también el caso del algoritmo *Deep Patient* (Knight, 2017) que fue capaz de predecir el diagnóstico de esquizofrenia, sin que nadie pudiera saber qué información procesó ni la forma en que lo hizo.

Sin embargo, la complejidad del sistema no es el único problema que nos podemos encontrar a la hora de analizar los algoritmos de IA. El otro gran problema que podemos encontrar es que la mayoría pertenecen a organismos privados que no permiten su estudio, como es el mencionado caso de *COMPAS*, utilizado para determinar el riesgo de reincidencia de los reclusos en algunas cárceles de Estados Unidos. Con *COMPAS*, los investigadores tuvieron que analizar de forma inversa los resultados proporcionados por el algoritmo en los últimos dos años para poder determinar que presentaba sesgos (Angwin et al., 2016). Otro caso similar es el de la *app* de citas *Tinder*, y que recoge Duportail (2019) en su libro donde indica que el

algoritmo calificaba como mejores candidatos a hombres con estudios superiores mientras que a las mujeres con estudios superiores les daba puntuaciones bajas, pero que nunca llegó a descubrir por completo el sistema de puntuaciones que usaba el algoritmo debido a que es uno de los mayores secretos de la empresa.

Estas dificultades para analizar los algoritmos pueden incluso costar vidas, como es el caso de los sanitarios de *Stanford Medicine* (Chen, 2020), donde un algoritmo decidió que los primeros sanitarios que debían vacunarse contra la *COVID-19* fueran los de más alto rango cuando eran los de bajo rango quienes estaban más en contacto con los enfermos. En otros contextos también pueden provocar errores que hagan que una persona inocente pueda ir a prisión, como los sistemas de reconocimiento facial que usan algunos cuerpos y fuerzas de seguridad de diversos países y el caso donde un hombre fue acusado en Detroit (Hill, 2020) porque un algoritmo de reconocimiento facial lo había identificado como alguien que había cometido un delito de hurto, teniendo solo en común ambas personas el color de la piel.

Este último error ocurre con cierta frecuencia debido a que los bancos de imágenes usados en los algoritmos de reconocimiento facial suelen estar compuestos, en su mayoría, por fotos de personas blancas, siendo más frecuentes los errores en personas cuyo color de piel no es blanco o de etnia europea (Buolamwini, 2016) pudiendo incluso no reconocerlas como seres humanos. Como puede verse, a pesar de que los algoritmos ya están en prácticamente todos los sectores de la sociedad (Kitchin, 2017; Neyland & Mollers, 2016; Willson, 2017), presentan aun multitud de sesgos y errores y, sin embargo, esto no hace que los eliminen de todos aquellos lugares donde determinan nuestras decisiones.

Un estudio realizado por nuestro equipo de investigación, que demuestra cómo las recomendaciones de la IA pueden influir en las decisiones de las personas, es el estudio de Vicente y Matute (2023), en el cual se pidió a los participantes que reconocieran una enfermedad en muestras ficticias de tejido humano siguiendo un criterio sencillo presentado mediante instrucciones escritas.

En el primer experimento, algunos participantes recibieron asistencia de una IA sesgada que cometía un error sistemático con un tipo de muestra. Estos participantes cometieron el mismo error que la IA, mientras que los participantes que no fueron asistidos por la IA no cometieron el error, ya que pudieron realizar la tarea correctamente. En el segundo experimento, se añadió una segunda fase de clasificación donde las muestras fueron clasificadas sin asistencia por ambos grupos. Los resultados mostraron que los participantes que fueron asistidos por la IA durante la fase inicial y que ahora no lo estaban, siguieron cometiendo el mismo sesgo, lo que sugiere que las personas podrían heredar el sesgo de la IA. En el tercer experimento, un grupo se entrenó solo antes de recibir la asistencia de la IA para probar si esto protegía a los participantes de la influencia de la IA. Los hallazgos mostraron que no fue así y que las personas cometieron el mismo error cuando la IA sesgada comenzó a asesorarlos.

Tal y como estamos viendo, podemos decir que la investigación sobre cómo afectan los algoritmos a las personas es algo muy necesario. Sin saber exactamente cómo funcionan, estamos dejando que los algoritmos nos aconsejen, o incluso decidan por nosotros, en prácticamente todos los aspectos de nuestra vida, ya sea en situaciones de ocio o algunas mucho más importantes y vitales. En el caso de esta tesis, el objetivo es indagar en la toma de decisiones y juicios de las personas en el contexto judicial cuando se encuentra presente un supuesto algoritmo de IA. Para

cumplir nuestro objetivo, hemos llevado a cabo una serie de experimentos donde un algoritmo ficticio actuaba como herramienta de apoyo a la toma de decisiones y su recomendación entra en confrontación con el resto de la información que los participantes tendrán disponible.

Capítulo 2. Algoritmos e IAs en Justicia

En el capítulo anterior hemos mencionado algunos ejemplos de algoritmos de IA que han mostrado sesgos, incluyendo algunos algoritmos de IA utilizados en el ámbito judicial y penitenciario. En este capítulo indagaremos un poco más en el campo de la justicia con el fin de comprender cómo ha ido evolucionando con el paso de los años y en qué situación se encuentra actualmente, así como qué podemos esperar en los próximos años.

La primera vez que aparecen modelos de IA en el ámbito judicial es a través de una serie de sistemas que asisten a los trabajadores jurídicos. Estos sistemas podían resolver cuestiones jurídicas mediante el análisis de información ya existente, aunque sin poder ofrecer un razonamiento jurídico (Ashley, 2017). Estos programas funcionaban igual que los sistemas de lógica computacional clásicos y no podían realizar argumentaciones sobre sus decisiones, por lo que, en un primer momento, en el contexto jurídico no tuvieron mucho éxito (Barona, 2021). Son los llamados sistemas expertos y, finalmente, dieron lugar a la principal aplicación de la IA en el área del derecho (Martínez, 2012).

Sistemas expertos en el sistema judicial

Los sistemas expertos no son algo que haya surgido en el campo jurídico. Los primeros modelos de sistemas expertos se pueden encontrar en campos como la medicina, la genética o la farmacia (Susskind, 1986). Son sistemas donde, inicialmente, un grupo de expertos humanos introducen el conocimiento que se quiere que el sistema experto posea para que ayude a los futuros usuarios. Para ello, los sistemas van haciendo preguntas a los usuarios hasta que son capaces de llegar a una respuesta válida para solucionar el problema que el usuario ha planteado.

Podríamos considerar a los sistemas expertos como los primeros algoritmos de IA modernos o los precursores de estos. Estos sistemas expertos presentan una serie de características, que son (Susskind, 1986):

1. Deben ser transparentes, por lo que deben poder dar una explicación del razonamiento que los ha llevado a sus conclusiones.
2. Llegan a sus conclusiones mediante el conocimiento extraído de la experiencia en un determinado campo.
3. Son flexibles, ya que es posible modificar sus bases de conocimiento si es necesario.

En conclusión, estos sistemas ayudan a los usuarios respondiendo preguntas jurídicas usando el conocimiento heurístico y el conocimiento formal de los expertos que las entrenaron.

El primer sistema experto jurídico podría considerarse el desarrollado por Sergot et al. (1986), cuyo objetivo fue convertir el *British Nationality Act* en una serie de condicionantes del tipo “si […] entonces […]” para determinar si una persona podía optar a la ciudadanía británica o no. Este modelo iba siguiendo un diagrama similar al de las ramas de un árbol de decisión en el que, ante una pregunta que se debía responder con un sí o un no, se avanzaba hasta la siguiente pregunta con el fin de determinar si las personas podían solicitar la ciudadanía británica o no. El problema que presentaba era que requería que la base de conocimiento fuera reescrita con cada modificación en la normativa.

Los siguientes modelos fueron los llamados positivistas explícitos subyacentes (Martínez, 2012). Estos fueron diseñados asumiendo que los usuarios aplican las normas jurídicas mediante la construcción de una serie de silogismos formados por

una norma jurídica (premisa mayor) y un suceso (premisa menor) y por la relación entre estas dos. El problema que presentaron estos modelos era que no podían hacer algo que los jueces humanos sí podían, que es recurrir a elementos más allá de las normas para emitir sus sentencias. Es decir, el problema que los juristas encontraron a estos modelos es que no eran capaces de salirse de lo que tenían programado para encontrar soluciones.

Otro tipo de modelos recibió el nombre de modelos de razonamiento legal basado en casos (Aleven, 2003). Estos tienen la particularidad de que, para resolver un problema, recurren a casos similares que tuvieran almacenados. Es decir, cuando se les presenta un problema, analizan toda la base de datos de casos de la que disponen y, cuando encuentran uno similar, analizan la resolución por si es aplicable también en este caso. De no ser así, pueden aplicar modificadores para adaptar la solución anterior al caso nuevo.

Habiendo comentado cómo se han ido desarrollando los sistemas expertos, a continuación, describiré los tres campos que se han propuesto en los que los sistemas expertos para la asistencia jurídica resultan especialmente útiles (Navas, 2017):

1. Jurisprudencia: la jurisprudencia consiste en todo el conjunto de fallos y sentencias dictados por los tribunales de justicia, por lo que en este contexto un sistema experto puede ayudar a los jueces a tener una línea de argumentación en su decisión, mientras que los abogados pueden utilizarlo para diseñar una defensa.
2. Análisis de documentos: aquí incluiríamos la extracción de información útil de un gran volumen de documentos, la presentación de sumarios de forma automatizada, así como la actualización constante de la información jurídica o el completado automático de formularios.

3. Códigos (legislativos, reglamentos o simples ordenanzas). La IA puede aplicarse a la elaboración de códigos relacionando normas entre sí, también ayuda a ver si la regulación de un supuesto concreto puede ser aplicada a otro supuesto y permite identificar normas que deberían modificarse para que el sistema tenga coherencia interna.

En comparación con los expertos humanos, los sistemas expertos tienen una serie de ventajas, como son el hecho de no tener exceso de confianza o sobreestimar sus propios conocimientos. Los sistemas expertos pueden ayudar a que la justicia llegue a cumplirse de forma más efectiva que actualmente, haciendo que haya menos conflictos sociales, los cuales muchas veces son provocados por el mal funcionamiento de la administración de justicia (Hernandez et. al., 2019). Sin embargo, esto no hace que los sistemas expertos, y los algoritmos de IA en general, estén libres de sus propios sesgos, sobre todo los depositados por los propios creadores de las bases de conocimiento, lo que hace necesario que los algoritmos sean transparentes y se presten a ser evaluados o verificados por comités de expertos ajenos a su creación. A continuación, presentamos algunos ejemplos de algoritmos de IA utilizados en el contexto judicial.

Sistemas Expertos Jurídicos. Ejemplos

Hay muchos sistemas expertos que ya se aplican o se aplicaron en varios sistemas judiciales. El primer ejemplo que vamos a mencionar es el sistema llamado *Split-Up* (Stranieri & Zeleznikow, 1998). Se trata de un sistema basado en reglas y se diseñó para ayudar en los casos de idoneidad parental y distribución de bienes en casos de divorcio en Australia. En estos casos, la Ley Familiar Australiana estipula una serie de factores (edad, salud, recursos económicos) a tener en cuenta a la hora de dictar sentencia. Lo que hicieron los desarrolladores de este sistema experto fue

establecer una jerarquía entre estos factores haciendo categorías (posibilidad de supervivencia, de empleo, obligaciones con los dependientes económicos, recursos financieros y contribución de los cónyuges al matrimonio) y subcategorías dentro de ellas. Este sistema permitía crear escenarios hipotéticos para los abogados y elaborar una predicción que les ayudara a la hora de llevar el caso. Por ejemplo, introducir que el cliente participó por igual en las tareas domésticas o que lo hizo de forma mayoritaria para así ver qué argumento beneficiaba en mayor medida a su cliente.

Otro sistema experto judicial es *Expertius* (Martínez, 2012), que fue desarrollado en México para que ayudara a decidir en juicios de manutención. Este sistema contaba con documentación más allá de la norma jurídica para poder llevar a cabo sus análisis, ya que contaba con el conocimiento que el operador le había proporcionado. Esta base de datos constaba de más de 400 expedientes judiciales sobre el tema. El proceso que seguía era el de buscar todos los argumentos posibles y la forma en la que estos pudieran ser refutados. También tenemos el caso de *AGATHA* (Chorley & Bench-Capon, 2006), primera herramienta que era capaz de desarrollar un argumento jurídico en base a los argumentos presentados en los casos que se le habían proporcionado para que aprendiera.

Por lo que sabemos, en España se utiliza hoy en día, al menos, dos sistemas expertos en el ámbito judicial. Uno de ellos es *VioGén* (Instituto de Ciencias Forenses y de la Seguridad, 2018). Este sistema tiene como función predecir y determinar el riesgo de reincidencia en los casos de violencia de género para así poder determinar las medidas de protección de la víctima. Algunas variables que tiene en cuenta son: el historial delictivo, la gravedad del primer hecho violento, la presencia de alteraciones mentales o el consumo de drogas. Sin embargo, parece que una auditoría externa ha detectado que *VioGén* comete errores de evaluación, y adolece de falta de

transparencia (del Castillo, 2022). Parece ser que no le da la suficiente importancia a la violencia psicológica y que, en la mayoría de los casos en los que solo se da de esta, los evalúa como de riesgo bajo, siendo imposible saber por qué debido a su opacidad. El otro sistema experto es *RisCanvi* (Soler, 2013), desarrollado en 2009 y que se encarga de ayudar a decidir sobre la puesta en libertad de presos, tal y como hace *COMPAS* en Estados Unidos.

Aunque los sistemas expertos son los algoritmos más frecuentes en justicia, también han empezado a aparecer sistemas más complejos. Un ejemplo es un algoritmo de IA que fue capaz de predecir el fallo emitido por los miembros del Tribunal Europeo de Derechos Humanos (Aletras et al., 2016). El porcentaje de acierto era de un 79% y las predicciones las realizó usando únicamente la información extraída de los juicios. Otro caso similar es el del algoritmo que fue capaz de predecir con un 70,2% de éxito el resultado del caso y con un 71,9% la votación de cada juez del Tribunal Supremo de Estados Unidos (Katz et al., 2017) en 28.000 casos y más de 240.000 votaciones que se remontaban desde principios del S.XIX, hasta 2015. Estos dos son ejemplos en los que el algoritmo analizaba, mediante aprendizaje automático, una gran cantidad de datos (legislación, jurisprudencia, protocolos) para llegar a una conclusión y una interpretación judicial.

Inteligencias Artificiales en los procesos judiciales

Con la irrupción de las IAs en los sistemas jurídicos, como por ejemplo *VioGén* (Instituto de Ciencias Forenses y de la Seguridad, 2018), no solo se está modificando la forma de trabajar y de actuar de los trabajadores jurídicos, sino que se está modificando el propio contexto jurídico, ya que estas nuevas herramientas pueden hacer que el sistema judicial cambie en busca de una mayor eficacia y eficiencia pudiendo, por ejemplo, requerir menos personal humano del que tiene ahora mismo.

Un ejemplo de ello lo encontramos en China (Paul, 2020). Desde 2017 tienen tres tribunales donde se llevan a cabo juicios virtuales sobre consumo, propiedad intelectual e industrial y que están dirigidos por una IA, que puede llegar a decidir en primera instancia. Estos tribunales están disponibles todos los días y a todas horas. Con respecto a lo mencionado sobre eficacia y eficiencia, los juicios en este sistema tienen una duración media de 28 minutos y un periodo de proceso de hasta 38 días, mientras que, en España, la media de un proceso civil son siete meses.

Otro ejemplo de una IA ya implementada en el ámbito judicial puede verse en Estonia (Bigas, 2019). En los juzgados de este país ya se han sustituido algunos puestos de funcionarios por un algoritmo que ha automatizado algunos procesos. Además de ello, tienen un proyecto para implementar un «Juez robot» que pueda llevar a cabo juicios pequeños similares a los de China. El proceso funciona de la siguiente manera: las dos partes del juicio subirán la documentación y toda la información relevante para el caso en una plataforma, donde la IA tomará una decisión que podrá ser apelada para que sea un juez humano quien tome la decisión.

En resumen, en este capítulo hemos intentado dejar patente que la IA y los algoritmos son ya una realidad en el ámbito jurídico. Además, cada vez están llevando a cabo tareas más complejas y autónomas. Sin embargo, también podemos observar que siguen cometiendo errores y fallos como es el caso del reciente caso detectado en España con el sistema *VioGén* (del Castillo, 2022). Por todo ello, resulta fundamental investigar cómo estos sistemas y sus posibles errores afectan a nuestra toma de decisiones en juicios y encontrar formas y métodos de reducir o evitar la automatización de nuestras decisiones, asumiendo que estas nuevas tecnologías van a ser una constante cada vez más frecuente.

Capítulo 3. Sesgos Humanos en Justicia

En los capítulos anteriores hemos descrito brevemente en qué consisten los algoritmos, las IAs y cómo se están integrando y forman parte ya del sistema judicial de muchos países. Y también hemos comentado que algunas de las ideas preconcebidas que muchas personas tienen sobre la introducción de los algoritmos en el sistema judicial es que van a mejorarlo debido a su objetividad (Araujo et al., 2020) y a su mayor eficacia (Sundar & Kim, 2019).

En apoyo a esta idea, hay que tener en cuenta que, en Justicia, como en cualquier otro aspecto de la vida, los humanos nos dejamos llevar por los sesgos que forman parte de nuestro día a día (Kahneman, 2011). Esto influye en ámbitos tan diversos como determinar si un medicamento es efectivo contra una enfermedad (Matute et al., 2022) o para decidir si una medida llevada a cabo por un partido político es efectiva o no, en función de nuestra ideología (Blanco et al., 2018).

Los jueces y jurados también tienen sesgos

Como ya discutimos anteriormente, los algoritmos presentan los sesgos que poseen aquellas personas que los diseñaron y las bases de datos con las que los entrenaron. Por ejemplo, el algoritmo *COMPAS* (Angwin et al., 2016) utilizado en EEUU para dictaminar la libertad condicional de los presos que la solicitaban mostraba un sesgo racista a la hora de presentar su valoración (Larson et al., 2016). Como es lógico, este sesgo también se ha encontrado en las personas que forman parte del sistema judicial, al igual que en todas las personas. Sommers y Ellsworth (2001) detectaron que, cuando los miembros del jurado eran personas blancas y el juicio en el que participaban como miembros de un jurado no estaba relacionado con racismo (es decir, no era un delito de odio racial), estos se mostraban más convencidos de la

culpabilidad de la persona acusada cuando esta era negra que cuando era blanca. Además, esta diferencia no se encontraba cuando el delito era específico de odio racial.

Podríamos pensar que son elementos personales tan importantes como, por ejemplo, la ideología, los que pudieran provocar que la decisión de jueces y jurados estuviera sesgada, pero, a veces, parecen ser hechos que resultan estar tan alejados del juzgado que jamás pensaríamos en ellos. Es el caso de la comida, como veremos a continuación.

En un estudio donde se analizaron más de 1100 juicios emitidos por ocho jueces de Israel (Danziger et al., 2011) en el que los reclusos solicitaban cambios en el tipo de condena que estaban cumpliendo, se llegó a la conclusión de que, conforme avanzaban los juicios a lo largo del día, los jueces tendían a emitir un juicio a favor del *status quo*, es decir, a no modificar las condenas o condiciones del recluso. Sin embargo, esta tendencia parecía cambiar si los jueces hacían un breve descanso y comían algo. Obviamente en este estudio, los propios autores no podían concluir que los jueces recuperaran su plena capacidad de resolver los problemas derivados de su trabajo solo con descansar, pero sí que las decisiones de los jueces, y de las personas en general, podían verse afectadas por múltiples variables ajenas al trabajo. Sin embargo, hay que comentar que este estudio recibió algunas críticas importantes respecto a su metodología. Weinshall-Margel y Shapard (2011) investigaron y llegaron a la conclusión de que los primeros casos que se ven en este tipo de tribunales son aquellos donde los solicitantes llevan abogado. Esto podría ser el motivo por el que hay más modificaciones de las medidas en los primeros casos.

Aun así, estas decisiones tomadas por jueces y jurados también pueden verse afectadas por otras variables que no tienen que ver con ellos mismos, sino con otros

aspectos de los juicios. En algunos estados de Estados Unidos, se requiere como prueba la grabación del interrogatorio donde los acusados confiesan, con el fin de poder comprobar si las confesiones se han hecho de forma coercitiva o bajo manipulación. Sin embargo, Lassiter y colaboradores (2002) realizaron un experimento donde a 42 participantes les mostraban el vídeo del interrogatorio y de la confesión, con la diferencia de que cada uno de los vídeos estaba grabado desde una perspectiva diferente: una de ellas enfocaba al acusado de frente y de cintura para arriba y mostraba la espalda del detective, mientras que la otra cámara mostraba a los dos participantes de perfil. En uno de los grupos, el que mostraba a los dos participantes de perfil, fue declarado culpable un 35% menos que en el otro, siendo esta la única variable modificada. Estos sesgos también fueron confirmados por otro experimento realizado unos años después (Ware et al., 2008)

En otra investigación, Pennington y Hastie (1992) pusieron a prueba a varios participantes que simulaban ser miembros de jurados. En una serie de experimentos evaluaron si, por ejemplo, se podía manipular a los participantes modificando la credibilidad de uno de los testigos. Cada participante veía cuatro testimonios; tres de ellos eran concordantes en la orientación de sus testimonios, mientras que un cuarto testimonio variaba el factor de credibilidad, así como la dirección de su testimonio. Por ejemplo, en uno de los casos, este testigo era la exmujer del acusado y era el único testigo que sugería la culpabilidad del mismo. La evaluación de la credibilidad de los testigos había sido baremada por una muestra de 432 participantes, donde cada testimonio era catalogado como de “baja credibilidad”, “alta credibilidad” o “nula credibilidad”. También manipularon las variables del orden y la forma en la que se contaba la historia, siendo modificada para ser contada como una historia o como una sucesión de testimonios. Este experimento encontró que la decisión de los jurados variaba en función de la forma y el orden en el que era contado el relato de los hechos.

Otro factor que parece afectar a las decisiones tomadas por los jurados es la información previa al juicio que se conoce de los acusados. En un experimento (Ruva & Guenther, 2014) se expuso a 320 participantes que simulaban ser miembros de un jurado a una información sobre el acusado una semana antes de un juicio por asesinato. La información podía ser negativa o neutra. Asimismo, también había un grupo que no recibía información previa sobre el acusado. Tras la segunda fase del experimento, que era la que simulaba el juicio, los participantes debían emitir un veredicto. Aquellos que previamente vieron información categorizada como negativa, tendían a dar una puntuación significativamente más alta de culpabilidad al acusado.

Los experimentos recientes realizados por nuestro equipo de investigación (Agudo et al., 2024) encontraron resultados relevantes sobre cómo influyen los posibles errores de los algoritmos de IA en el contexto legal y su influencia en los juicios humanos. En estos experimentos, se pidió a los participantes que decidieran si la recomendación dada por una IA en referencia a una decisión judicial debía ser aceptada o modificada. Esta recomendación de la IA se les presentó después de exponerles un caso y una serie de testimonios. Cada participante vio tres casos. En los dos primeros, la decisión del algoritmo fue consistente con lo que se esperaría tras leer los testimonios de los testigos. En el último, la recomendación del algoritmo fue errónea, es decir, contraria a lo que sugerían los testigos.

La diferencia entre los grupos experimentales fue que a un grupo se le dio la información sobre la recomendación del algoritmo antes de emitir su propio juicio y al otro grupo se le dio después. Los resultados indicaron que el efecto de la IA sobre los participantes fue mayor cuando la recomendación del algoritmo se presentaba antes de que emitieran su juicio, probablemente debido a un efecto de anclaje (Kahneman, 2011), lo que haría que la recomendación del algoritmo se convirtiera en el punto de

referencia para los miembros del jurado. Los experimentos de Agudo y colaboradores (2024) también mostraron que esta influencia de los errores de la IA en las decisiones humanas puede reducirse si se obliga a la persona a emitir su propio juicio antes de ver la recomendación de la IA para contrarrestar de esta forma el efecto de anclaje.

Influencia del orden

Además de estos experimentos, encontramos otros en los que se investigaron factores más relacionados con estudios clásicos de la psicología experimental como son el orden de presentación de la información (Wilson, 1971), así como los efectos de primacía (Pennington, 1982) y recencia (Furham, 1986). Estos experimentos son parte de la base sobre la que se sustentaron los experimentos realizados en esta tesis, por lo que nos detendremos en ellos con más detalle.

El efecto del orden en la presentación de la información en los juicios humanos y la toma de decisiones ha sido estudiado en varias ocasiones a lo largo de los años, encontrando resultados que indicaban que el efecto más frecuente era el efecto de primacía (Pennington, 1982) en el orden de presentación de los testimonios. Es decir, que las personas tienden a atribuir más peso a los primeros testimonios que se les presentan.

En el experimento de Pennington (1982), 192 participantes simularon ser miembros de un jurado. Para estudiar el efecto del orden, Pennington los dividió en cuatro grupos en los que variaba el orden de presentación de los testigos (si eran de la defensa o de la acusación) y el orden de presentación de los testimonios (de inocencia o de culpabilidad). El instrumento que utilizó consistió en unas transcripciones de los audios realizados por Sealy (1975) y por Sealy y Cornish (1973), que consistían en la simulación de un juicio, basado en un caso real, que grabaron mediante actores.

En los resultados, a pesar de esperar obtener un efecto de recencia, obtuvieron un efecto de primacía, consistente en que los testimonios presentados al principio parecían determinar el veredicto. Las posibles explicaciones que da Pennington (1982) en sus conclusiones al motivo de este efecto de primacía son dos. Por un lado, cree que el hecho de que toda la información les sea presentada a los participantes durante el breve lapso de tiempo de duración del experimento, cuando en la vida real los juicios duran mucho tiempo, podría quizás influir en que los primeros testimonios hubieran tenido más peso en los veredictos. Por otro lado, el caso que utilizó para el experimento consistió en un caso de agresión sexual, por lo que contempla la idea de que los participantes se hayan dejado llevar por una respuesta más emocional, especialmente ante los primeros estímulos recibidos.

En otra investigación, se encontró también un efecto de primacía, al mostrar que las probabilidades de que un acusado fuera absuelto en un tribunal podían depender de si el primer discurso lo realizaba el abogado de la defensa o el de la acusación, siendo más fácil que fueran absueltos si hablaba primero el abogado defensor (Wells et al., 1985). A una muestra de 291 estudiantes se les presentaba una transcripción de un juicio real donde los investigadores habían manipulado el orden de exposición del fiscal y el abogado, formando dos grupos. Al analizar los resultados, encontraron que el grupo donde leían primero el discurso del abogado defensor tendía a declarar inocente al acusado en mayor medida que el otro grupo.

Podría parecer, viendo los resultados de estos experimentos, que el efecto de primacía es bastante frecuente en el ámbito judicial. Sin embargo, hay otros experimentos donde el efecto encontrado es el de recencia, es decir, que la última información vista por los participantes es la que cobra más importancia. Furham (1986) encontró en sus experimentos que, cuando alteraba el orden de presentación de la

información de un juicio, la información presentada al final era la que más relevante resultaba. En uno de los experimentos, los participantes tenían que decidir la inocencia o culpabilidad de una persona en un caso de bigamia tras la presentación de siete testimonios. La diferencia entre dos de los grupos era el orden de los testimonios (si veían primero los testimonios de inocencia o los de culpabilidad). Tras ver todos los testimonios tenían que valorar en una escala de 1 a 9 la inocencia o culpabilidad del acusado. Los resultados mostraban que los juicios de los participantes coincidían con lo que veían al final del proceso, siendo idéntica la información recibida por ambos grupos salvo por el orden de presentación y se observó en los resultados un efecto claro de recencia.

Siguiendo con el efecto de recencia, en una serie de experimentos, Costabile y Klein (2005) encontraron que las pruebas presentadas al final del proceso judicial eran más determinantes a la hora de influir en el veredicto. En el primer experimento presentaron a 61 participantes el resumen de un juicio por homicidio. El material formaba parte del material elaborado por Kassin y Sommers (1997) y consistía en declaraciones de cinco testigos de los cuales, uno de ellos presentaba información crucial para el caso y era este testimonio el que variaba en posición (primero o quinto) entre los distintos grupos de participantes. Los resultados de este experimento mostraban que resultaba más relevante este testimonio cuando se presentaba en último lugar en vez de en primer lugar, provocando un efecto de recencia. En los siguientes tres experimentos de la investigación también encontraban efecto de recencia.

Como podemos observar, parece que el orden de presentación de la información en el ámbito judicial es realmente importante, aunque en la literatura que hemos podido revisar no queda claro el efecto que predomina, si primacía o recencia.

Tal vez el problema que ensombrece este tipo de investigaciones es que cada trabajo utiliza una serie de instrumentos distintos que pueden provocar diferencias en los resultados.

Influencia de la frecuencia del juicio

Un aspecto ligeramente relacionado con el orden de presentación de la información y con el que también trabajaremos en esta tesis es la frecuencia del juicio. Catena y colaboradores (1998) encontraron en una serie de experimentos que la frecuencia con que se pedía a los participantes que emitieran un juicio afectaba a la relación de causalidad entre dos sucesos. Estudios posteriores como los de Matute y colaboradores (2002) o el de Collins y Shanks (2002), encontraron entre sus resultados que el juicio que daban los participantes se iba modificando acorde a la información que habían visto justo antes de emitir un juicio, por lo que el orden interactuaba con la frecuencia. Collins y Shanks (2002) argumentaban que los participantes, a los que se les pedía ir emitiendo un juicio tras cada ensayo de la tarea, tenían en cuenta el último juicio emitido y cómo la nueva información afectaba a ese juicio funcionando ese último ensayo como una especie de ancla y generando un efecto de recencia.

En los experimentos de Vadillo y colaboradores (2004) encontraron resultados similares. En sus experimentos vieron que la frecuencia con la que se solicitaba una evaluación afecta a los juicios de los participantes. Cuando las evaluaciones se realizaban tras cada ensayo, los participantes mostraban efecto de recencia, equivalente al mostrado por los resultados ensayo a ensayo de Collin y Shanks (2002) y de Matute y colaboradores (2002) basándose en la información más reciente. En contraste, otros de los grupos debían emitir el juicio al final de la serie de ensayos y, a

diferencia de los primeros participantes, estos integraban toda la información vista, mostrando incluso un ligero efecto de primacía.

Otro artículo sobre la frecuencia del juicio es el de Catena y colaboradores (2004) que también encontró que cuando a los participantes se les presentaba una alta frecuencia de juicios, tendían a centrarse en la última respuesta dada y emitir su nuevo juicio teniendo sobre todo en cuenta esa información, mientras que cuando se les presentaba una baja frecuencia de juicio, mejoraban la precisión de la tarea que se les pedía ya que integraban mejor toda la información recibida.

Durante esta introducción, hemos hecho una revisión sobre la existencia de sesgos y errores tanto en personas, como en los propios algoritmos e incluso en las creencias que tienen las personas de los algoritmos. Además, hemos comentado que en el ámbito de la justicia también se cometen errores a pesar de la creencia de la objetividad de la justicia y que variables tales como el orden de presentación y la frecuencia con que se piden los juicios pueden estar influyendo en la obtención de unos resultados u otros. Por ello, en los experimentos de esta tesis trataremos de investigar estos tipos de errores, así como algunas de las variables que pueden influir en la obtención de resultados diferenciales y, por último, intentaremos también hallar una solución o una manera de reducir los errores.

Por ello, antes de centrarnos en la influencia de los algoritmos sobre los juicios de los jurados humanos, la siguiente parte de esta tesis doctoral consistió, en la realización de una serie de experimentos para conocer la influencia del orden y de la frecuencia del juicio en nuestros procedimientos de modo que pudiéramos neutralizar (o al menos conocer) su posible influencia en los experimentos posteriores de la tesis. A continuación, en la tercera parte de la tesis nos centramos en la elaboración y estandarización de un material que tuviera validez ecológica, que pudiera usarse con

participantes españoles, que pudiéramos usar en los demás experimentos de la tesis y que, además, pudiera ser compartido con otros investigadores que deseen replicar o profundizar en los hallazgos de esta tesis. Tras esos primeros experimentos y tras la creación de un instrumento estandarizado, nos centraremos, en la cuarta parte de la tesis, en unos experimentos que diseñamos con el objetivo de comprobar si los humanos (en nuestro caso, miembros de un jurado popular ficticio) podían dejarse influenciar por los posibles sesgos y errores en las recomendaciones del algoritmo, así como poner a prueba algunas posibles soluciones para minimizar esos errores. Finalmente, en la quinta y última parte de esta tesis abordaremos la discusión general.

Parte II. Experimentos

Capítulo 4. La Influencia del Orden y la Frecuencia del Juicio en las Decisiones Humanas en un Contexto Judicial

Como hemos mencionado en el capítulo anterior, se han realizado numerosas investigaciones que muestran que el orden en el que se presentaba la información en un juicio podía afectar a las decisiones tomadas por los miembros de un jurado aunque, por otro lado, en la literatura revisada, hay artículos que sugieren que tiene más peso en el juicio la información presentada en primer lugar, mientras que otros sugieren lo contrario (Costabile & Klein, 2005; Furham, 1986; Pennington, 1982; Wells et al., 1985).

También hemos visto que la frecuencia del juicio parece tener un efecto en las decisiones tomadas por las personas (Catena et al., 2004; Collins & Shanks, 2002; Vadillo et al., 2004;), creando una especie de ancla en la última respuesta dada cuando el juicio se solicita con frecuencia, pero no cuando se pide de manera global

Otro estudio que encontró resultados relacionados con la frecuencia del juicio fue el de Matute y colaboradores (2002). En ese artículo la tarea consistía en una simulación donde los participantes debían determinar si una medicina ficticia provocaba una reacción alérgica como efecto secundario o no. En este trabajo los autores quisieron comprobar si el orden de presentación de los ensayos, así como la frecuencia con la que a los participantes se les pedía que emitieran un juicio, podía afectar a los juicios emitidos por ellos en un contexto no judicial (entendiendo juicio como valoración o estimación subjetiva).

En el primer experimento de este artículo (Matute et al., 2002), en la primera mitad de los ensayos, la administración de la medicina iba seguida de la reacción alérgica, mientras que en la segunda mitad no. Los investigadores encontraron que,

cuando el juicio se solicitaba tras cada ensayo a los voluntarios, éste iba siendo modificado en cada ensayo y que, debido a que los últimos ensayos no presentaban reacción alérgica, la media del último juicio resultaba inferior, es decir, los participantes consideraban que el medicamento no causaba efectos secundarios comparado con aquellos en los que la secuencia de ensayos se presentaba en el orden inverso, es decir, primero los ensayos sin reacción alérgica y al final los ensayos en los que se daba la reacción alérgica. Esto es indicativo de un efecto de recencia cuando el juicio se pedía en cada ensayo, que sin embargo no se daba cuando el juicio debía ser emitido de forma global al final de todos los ensayos. En este último caso, toda la información que habían visto (tanto positiva como negativa) la integraban antes de emitir un juicio, que resultaba tener un valor medio. Otros trabajos en los que se ha obtenido también un efecto de recencia (entendido como el efecto en el que la información recibida al final de una serie de ensayos adquiere mayor importancia que la información recibida en los primeros ensayos) en los experimentos que utilizaban un juicio ensayo a ensayo son, por ejemplo, los realizados por López y colaboradores (1998).

Con el objetivo de conocer cómo influyen estas variables en nuestros procedimientos, decidimos llevar a cabo un acercamiento entre el contexto judicial y los procedimientos de algunos experimentos anteriores realizados en nuestro laboratorio. Para ello, tomamos como base el trabajo de Matute y colaboradores (2002) y decidimos adaptar la metodología de este experimento al contexto judicial.

Aunque el contexto judicial es bastante complejo, decidimos empezar la investigación con un experimento similar a este, debido a la experiencia de nuestro grupo con esta metodología, pero cambiando el contexto médico por el contexto judicial de modo que pudiéramos utilizarlo posteriormente en los demás experimentos

de la tesis. La ventaja de utilizar una metodología experimental que ya conocíamos era que se había trabajado mucho con ella y se conocían los resultados que deberíamos esperar en estos primeros experimentos si nuestra adaptación del procedimiento, instrucciones, etc., funcionaba adecuadamente en el contexto judicial. Por este motivo, podíamos predecir en mejor medida los posibles resultados y detectar los posibles errores que pudiéramos tener en la adaptación del procedimiento.

Experimento 1-A. Influencia del orden y frecuencia del juicio

El objetivo de este experimento fue evaluar si tanto el orden en el que se iban presentando los estímulos durante un proceso judicial como la frecuencia con que se pedía el juicio podía afectar al veredicto emitido por los participantes y cómo lo hacía. Como ya hemos mencionado, el contexto en este caso era la simulación de un juicio, por lo que los estímulos eran testimonios de inocencia o culpabilidad. Esperábamos replicar en el contexto judicial los resultados de Matute et al. (2002) en el contexto médico, en los que se observaba que los participantes iban modificando su juicio con cada nuevo ensayo, mostrando un efecto de recencia, y que, por otro lado, integraban toda la información cuando tenían que emitir el juicio global.

Método

Participantes y Materiales. El número de participantes fue de 90 (40% mujeres, 59% hombres, 1% otro; edades comprendidas entre 18 y 69 años, *M* = 29.9; *DT* = 8.89) reclutados en la Universidad de Deusto. Un análisis de sensibilidad muestra que, con el tamaño de muestra actual (n=90), obtenemos una potencia de 0,80 para detectar un efecto de tamaño pequeño (*d* = 0.208) en las diferencias entre grupos. El experimento se realizó presencialmente en las cabinas individuales con ordenadores del Laboratorio de Psicología Experimental.

Los participantes se asignaron aleatoriamente a uno de los siguientes tres grupos: Grupo Control (*n* = 34); Grupo T(I)-T(C) (*n* = 30) y Grupo T(C)-T(I) (*n* = 26). La nomenclatura de los grupos indica si los testimonios de inocencia [T(I)] se presentaban antes que los testimonios de culpabilidad [T(C)], después, o en orden aleatorio (control). El experimento se realizó a través de un programa HTML dinámicamente modificado por Java Script.

Diseño y Procedimiento. El procedimiento utilizaba una tarea estándar de contingencias similar a la utilizada por Matute y colaboradores (2002) en contexto médico pero adaptada al contexto judicial. La sesión estuvo formada por 40 ensayos, 20 en los que los testimonios indicaban culpabilidad y 20 en los que los testimonios indicaban inocencia. Los participantes se asignaron aleatoriamente a tres grupos, cuya diferencia es el orden de presentación de los ensayos, tal y como se muestra en la Tabla 1. En el Grupo T(I)-T(C) se presentaban primero los 20 ensayos de inocencia, en el Grupo T(C)-T(I) primero los 20 de culpabilidad y en el Grupo Control se presentaban al azar. Cada participante debía emitir un juicio sobre la culpabilidad del acusado tras cada uno de los 40 ensayos. Este juicio ensayo a ensayo era la variable dependiente crítica del experimento.

Además de esto, también se les pidió que emitieran un Juicio Global al final de la serie de 40 ensayos, con el fin de comprobar si eran capaces de integrar toda la información, tal y como se mostraba también en el trabajo de Catena y colaboradores (1998), en el de Collins y Shanks (2002), y en el de Matute y colaboradores (2002). Para recoger tanto los juicios de cada ensayo como el Juicio Global utilizábamos la pregunta “¿Hasta qué punto consideras a la persona acusada inocente o culpable?”, seguido de una escala que iba desde -100 (inocente) hasta 100 (culpable).

Tabla 1

Resumen del Diseño del Experimento 1-A

Grupo	Fase 1	Fase 2	Juicio crítico	Juicio adicional
Control	T(I)/T(C)		Ensayo	
T(I)-T(C)	T(I)	T(C)	a	Global
T(C)-T(I)	T(C)	T(I)	ensayo	

Nota. T= Testimonios; I= Inocencia; C= Culpabilidad; / = orden aleatorio

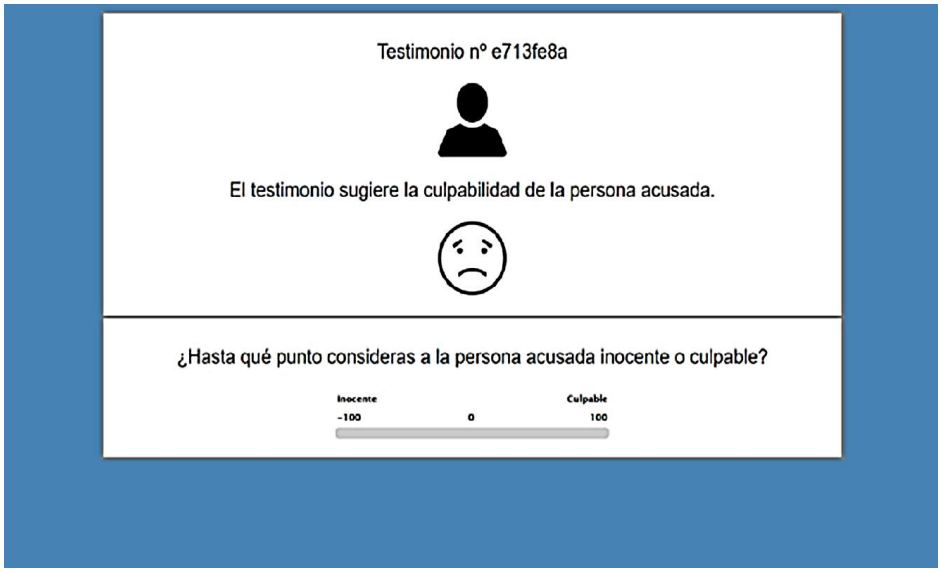
Las variables independientes fueron: el orden de presentación de los testimonios de inocencia (I) y culpabilidad (C), así como el momento en el que se emitió la respuesta, siendo los juicios ensayo a ensayo, y no el juicio global, aquellos en los que esperábamos observar un efecto de recencia si replicábamos los resultados de Matute y colaboradores (2002) en este nuevo contexto. En el Juicio Global no lo esperábamos porque este funcionaría, al menos en principio, como integrador de todos los ensayos.

Al empezar la tarea, a los participantes se les planteó una historia con el siguiente texto: *"Imagina que formas parte de un jurado. Durante todo el proceso judicial se te irán presentando una serie de testimonios que servirán para apoyar la teoría de la acusación (culpable) o de la defensa (inocente). Aunque el testimonio de los testigos no es suficiente para condenar a una persona a prisión, nos interesa tu opinión personal tras escuchar a los testigos. Cada vez que un testimonio te sea presentado, tendrás que indicar hasta qué punto te inclinas a pensar que esa persona es culpable o inocente."*

A continuación, tenían lugar los ensayos sucesivos en los que mostrábamos los testimonios y solicitábamos a los participantes su juicio ensayo a ensayo, utilizando para ello pantallas similares a la que se muestra en la Figura 5.

Figura 5

Ejemplo de Ensayo de Culpabilidad y Pregunta de Juicio Ensayo a Ensayo en el Experimento 1-A



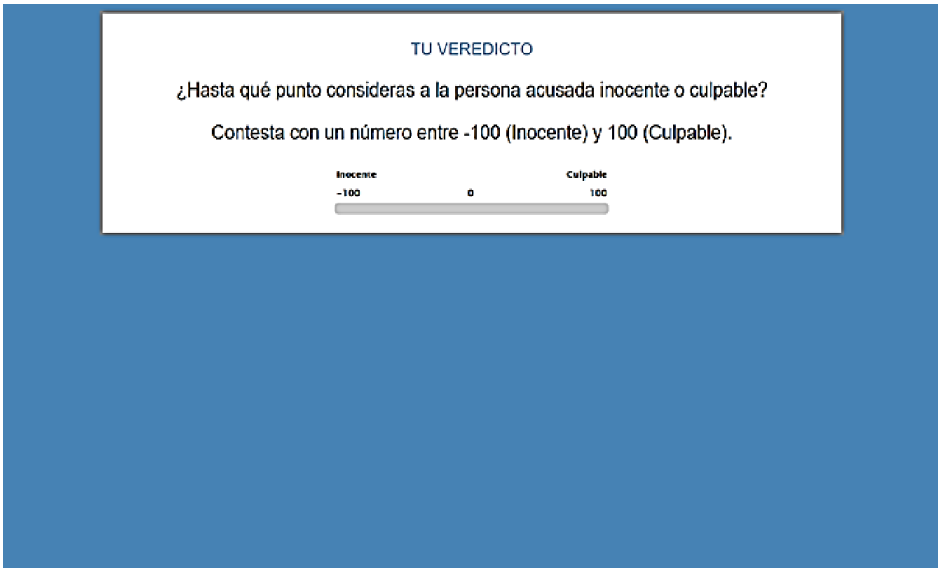
En cada ensayo debían valorar hasta qué punto consideraban a esa persona inocente o culpable mediante una puntuación de -100 a 100, como hemos mencionado anteriormente. Tras cada ensayo, se presentaba una pantalla donde los participantes tenían que pulsar para poder continuar viendo más ensayos.

Finalmente, después de haber visto los 40 ensayos se les explicaba que a continuación tendrían que emitir un informe final considerando toda la información que

habían visto durante toda la tarea, en concreto: *"Ya has terminado de realizar la primera parte. Ahora te pediremos que emitas un informe final y definitivo teniendo en cuenta absolutamente toda la información que has recibido sobre la persona durante el juicio".* Y, tras ello, se les mostraba una pantalla como la de la Figura 6, donde se les volvía a preguntar hasta qué punto consideraban a la persona acusada inocente o culpable, teniendo que valorarla de nuevo en una escala de -100 a 100.

Figura 6

Pantalla de Juicio Global Durante Experimento 1-A

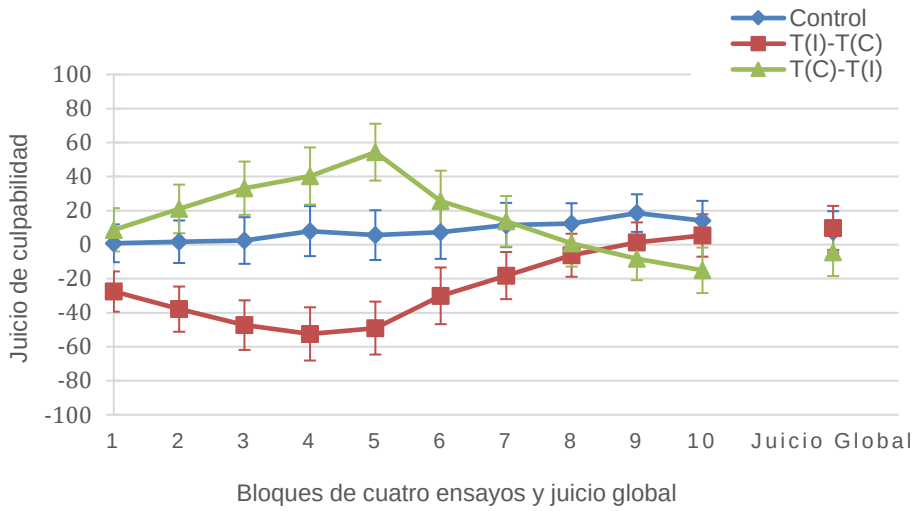


Resultados y Discusión

Los ensayos se agruparon en 10 bloques de cuatro ensayos para poder analizar mejor la evolución de los juicios a lo largo del experimento, como puede verse en la Figura 7.

Figura 7

Juicio de Culpabilidad de los Grupos por Bloque de Ensayos y en el Juicio Global



Nota: T: Testimonios; I: Inocencia; C: Culpabilidad; las barras de error indican el error típico

El Grupo Control se mantiene en unas puntuaciones centrales durante todos los ensayos, mientras que el grupo con testimonios iniciales de inocencia puntúa en la primera mitad cada vez más hacia -100 (inocente), y el grupo con testimonios iniciales de culpabilidad puntúa en la primera mitad cada vez más hacia el 100 (culpable). Esta tendencia comienza a cambiar tras el Bloque 5 (ensayo 20) donde los testimonios que ven los participantes pasan a ser de sentido contrario, lo que se refleja en el cambio que sufren sus juicios, que comienzan a ser valorados en el sentido contrario al que habían estado siendo valorados hasta ese momento, llegando a un punto central de la escala en los últimos bloques.

Estas impresiones las confirmamos con un ANOVA mixto con las variables Grupo y Bloque sobre los juicios ensayo a ensayo. Aunque el ANOVA no mostró un efecto principal para los bloques, $F(10, 870) = 1.36$, $p = .194$, $\eta^2g = .009$ sí que mostró efecto principal de la variable Grupo, $F(2, 87) = 18.00$, $p<.001$, $\eta^2g = .157$ así como interacción entre Bloque y Grupo, $F(20, 870) = 15.11$, $p<.001$, $\eta^2g = .157$. Al realizar, a continuación, las comparaciones Post-hoc con la corrección de Tukey observamos lo siguiente: en el Bloque 1 ya encontramos diferencias entre los dos grupos experimentales, con $t(87) = -4.0960$, $p = .030$; mientras que el Grupo Control no presenta diferencias con el Grupo T(I)-T(C), $t(87) = 3.4237$, $p = .191$, ni con el Grupo T(C)-T(I), $t(87) = -0.9209$, $p = 1.000$. Estas diferencias entre los dos grupos experimentales están presentes también en el Bloque 2 $t(87) = -5.9180$, $p < .001$, Bloque 3 $t(87) = -7.3734$, $p < .001$, Bloque 4 $t(87) = -7.9222$, $p<.001$, Bloque 5 $t(87) = -8.8706$, $p < .001$ y Bloque 6 $t(87) = -4.4646$, $p = .009$. A partir del Bloque 6, deja de haber diferencias entre ellos desde el Bloque 7 $t(87) = -3.0729$, $p = .393$ hasta el Bloque 10 $t(87) = 2.1926$, $p = .943$. Esto puede deberse a que es durante el Bloque 6 cuando empieza a presentarse información contraria a la que han estado viendo desde el principio y empiezan por tanto a partir de ese punto a converger más los juicios hacia las puntuaciones intermedias. El Grupo Control no difiere de los anteriores.

Por otro lado, al comparar los juicios globales de los dos grupos experimentales, encontramos que las puntuaciones son prácticamente iguales: $t(87) = 1.4923$, $p = 1.000$; lo que nos indica que los participantes integran al finalizar toda la información recibida durante todo el proceso si se les solicita el juicio global al final. Esto se corresponde con lo observado en el experimento de Matute y colaboradores (2002), donde a aquellos participantes a los que se les pedía un Juicio Global, tendían a integrar toda la información. El Grupo Control, tal y como esperábamos tampoco difiere de los otros dos grupos en el Juicio Global.

Asimismo, vemos que la curva de aprendizaje de los grupos es también de forma similar a los grupos del experimento de Matute y colaboradores (2002) en el que los participantes iban ajustando el juicio ensayo a ensayo según lo observado en el ensayo anterior y a los que se les mostraba también casos de signo contrario en la segunda mitad del experimento. Esto resultó interesante ya que parece indicar que el orden en el que se presenta la información en un juicio puede afectar al veredicto, especialmente si se emiten juicios y opiniones con frecuencia, sin esperar a la valoración global. Esto confirma los resultados de los experimentos mencionados en el capítulo anterior. No obstante, en el presente experimento los resultados no son tan marcados como en el de Matute y colaboradores (2002), donde las líneas de los grupos acababan cruzándose y dando juicios claramente opuestos en el último Bloque. En el experimento actual, aunque la tendencia es la misma, las diferencias de los últimos bloques son menos marcadas, dado que se quedan en un punto medio y no llegan a convertirse en opuestas. Aunque esto no es una diferencia crítica para nuestro propósito, en el siguiente experimento exploramos algunas cuestiones procedimentales que pudieron estar detrás de esta pequeña discrepancia, y qué podríamos mejorar para conseguir resultados más robustos en sucesivos experimentos.

Experimento 1-B. Influencia del orden y frecuencia del juicio con escala unidireccional

En el Experimento 1-A comprobamos que el orden de presentación de la información puede influir en el veredicto emitido por los participantes cuando se solicitan los juicios ensayo a ensayo, confirmando lo observado anteriormente tanto en experimentos realizados en el contexto judicial como los realizados por Matute y colaboradores (2002) en contexto médico con un procedimiento similar al que estamos

usando en esta tesis doctoral, aunque los resultados de nuestro experimento resultaron menos marcados. Mientras que en el experimento de Matute y colaboradores los grupos llegaban a ocupar al final de la tarea posiciones opuestas a sus puntuaciones iniciales dentro de la gráfica de puntuaciones, en el Experimento 1-A los grupos se quedaron en los últimos ensayos en una posición intermedia de la escala de puntuaciones.

Tras analizar el experimento y sus resultados, pensamos que es posible que la escala de valoración bidireccional que habíamos utilizado (de -100 a 100) pudiera haber afectado al juicio emitido por los participantes, ya que en la mayoría de los experimentos previos se utilizaba una escala unidireccional (de 0 a 100), que quizás fuera algo más sensible. Existe bibliografía que muestra que los dos tipos de escala funcionan de manera diferente, y sugiere que a veces pueden resultar mejores las escalas bidireccionales (Ng et al., 2023). Esto puede ser así debido a que en las escalas unidireccionales se elimina la opción a los participantes de valorar negativamente, algo que en ocasiones podría ser necesario.

Sin embargo, en nuestro caso podía tener más sentido, y ser más fácilmente comprensible para los participantes una escala en la que la culpabilidad va de 0 a 100, en lugar de una en la que va de -100 a 100. De hecho, el valor de cero en la mitad de la escala bidireccional podía haber inducido algún tipo de error y que algunos participantes hubieran confundido un valor de cero de culpabilidad con una valoración de inocencia, que sin embargo estaba etiquetada en el valor -100. Además, en los experimentos originales de Matute y colaboradores (2002), la escala que usaban era unidireccional. Por ello, decidimos replicar el Experimento 1-A cambiando la escala de culpabilidad a una escala de 0 a 100.

Método

Participantes y Materiales. Reclutamos a 97 participantes (81.44% mujeres, 18,66% hombres, edades 18-22, *M* = 18.7, *DT* = 0.875) que eran parte del alumnado de primer curso del Grado de Psicología de la Universidad de Deusto en la asignatura de Procesos Psicológicos Básicos II. Un análisis de sensibilidad muestra que, con el tamaño de muestra actual (*n* = 97), obtenemos una potencia de 0,80 para detectar un efecto de tamaño bajo, *d* = 0.18, en las diferencias entre grupos. Los participantes realizaron el experimento como parte de la actividad docente. No estaban obligados a enviar los datos una vez completada la tarea, aunque como parte de la evaluación debían realizar un pequeño informe sobre lo que habían aprendido durante una clase práctica en la que, al finalizar la tarea, recibieron una explicación completa y didáctica sobre el experimento.

La asignación a cada uno de los tres grupos se hizo de forma aleatoria: Grupo Control (*n* = 26), Grupo T(I)-T(C) (*n* = 34) y Grupo T(C)-T(I) (*n* = 37). El experimento se realizó a través de un programa HTML dinámicamente modificado por Java Script y diseñado para el experimento.

Diseño y Procedimiento. El único cambio respecto al experimento anterior fue que sustituimos la escala bidireccional (-100 a 100) del Experimento 1-A, por una unidireccional de 0 (inocente) hasta 100 (culpable), tal y como mostramos en las Fig. 8 y 9.

Figura 8

Ejemplo de Ensayo de Inocencia y Juicio Mostrado en el Experimento 1-B

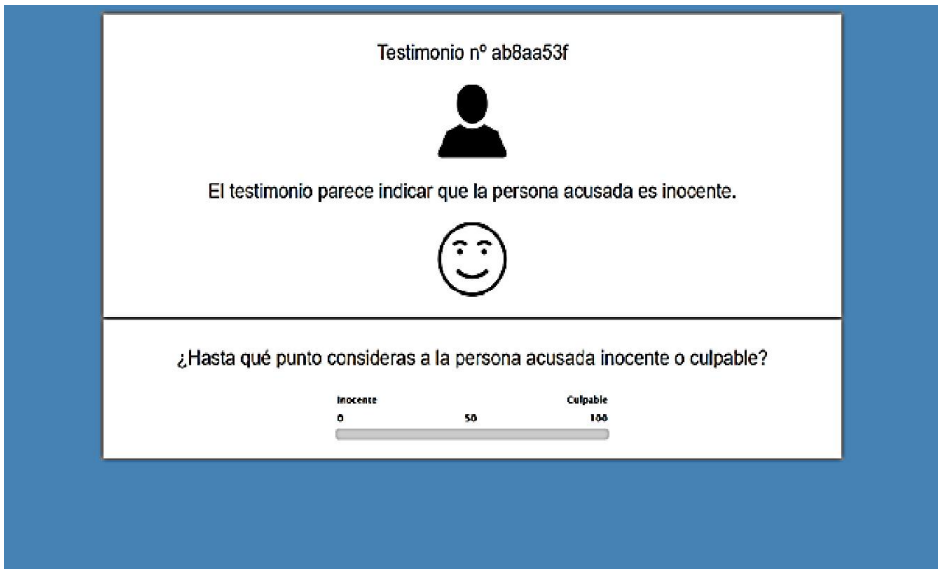
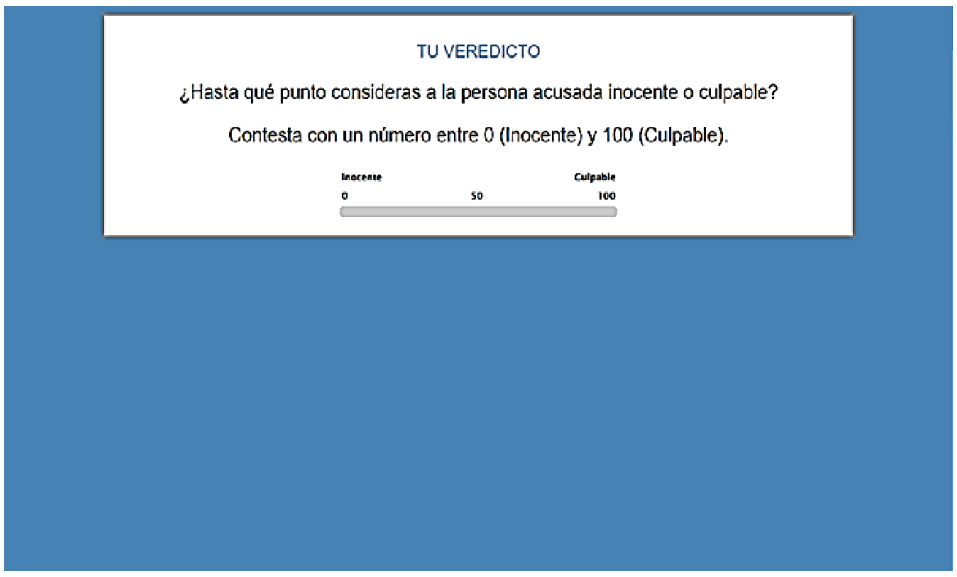


Figura 9

Pantalla de Juicio Global Durante Experimento 1-B

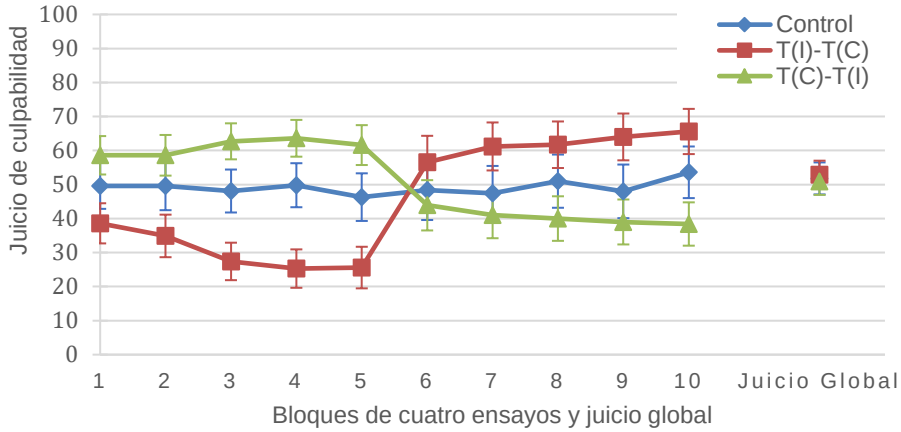


Resultados y Discusión

Al igual que en el Experimento 1-A, los ensayos fueron agrupados en 10 bloques de cuatro ensayos cada uno para poder analizar mejor la evolución de los juicios a lo largo del experimento. Como podemos observar en la Figura 10, sucede algo similar al anterior experimento, aunque de forma más marcada.

Figura 10

Juicio de Culpabilidad de los Grupos por Bloque de Ensayos y en el Juicio Global



Nota: T: Testimonios; I: Inocencia; C: Culpabilidad; las barras de error indican el error típico

Realizamos un ANOVA mixto con las variables Grupo y Bloque sobre los juicios ensayo a ensayo, que mostró un efecto principal para los bloques, $F(10, 940) = 2.37$, $p = .009$, $\eta^2g = .017$, así como interacción entre Bloque y Grupo con $F(20, 940) = 25.58$, $p < .001$, $\eta^2g = .261$. Sin embargo, no encontramos efecto principal en la variable

Grupo $F(2, 94) = 1.29$, $p = .279$, $\eta^2g = .009$. En este experimento las comparaciones Post-Hoc con la corrección de Tukey mostraron resultados similares a los del Experimento 1-A entre los grupos experimentales durante los Bloques del 1 al 5, encontrando diferencias significativas en todos ellos: Bloque 1, $t(94) = -4.792$, $p = .003$; Bloque 2, $t(94) = -5.358$, $p < .001$; Bloque 3 $t(94) = -9.058$, $p < .001$; Bloque 4 $t(94) = -9.596$, $p < .001$ y Bloque 5, $t(94) = -8.330$, $p < .001$.

Llegados a este punto, en el Experimento 1-A, en el Bloque 6 había diferencias significativas y luego dejaba de haber diferencias entre los dos grupos experimentales. Sin embargo, en este Experimento, sucede lo contrario. En el Bloque 6 no encontramos aún diferencias significativas, $t(94) = 2.316$, $p = .901$ pero, a partir del Bloque 7, empezamos a encontrar diferencias significativas entre los dos grupos experimentales en todos los bloques: Bloque 7, $t(94) = 4.050$, $p = .033$; Bloque 8, $t(94) = 4.495$, $p = .008$; Bloque 9, $t(94) = 5.133$, $p < .001$ y Bloque 10, $t(94) = 5.799$, $p < .001$. Es justo en el Bloque 6 donde empiezan a recibir información contradictoria, lo que hace que las puntuaciones sean similares en ese bloque. Conforme van pasando los bloques y siguen viendo nueva información, se distancian en la dirección contraria. Es decir, se muestra un efecto del orden más acusado que en el Experimento 1-A y en este nuevo experimento las diferencias resultan estadísticamente significativas no solo en la primera mitad del experimento sino también en la segunda. Por tanto, parece que el cambio a una escala de 0 a 100 sí ha tenido un efecto sobre los resultados, siendo más sensible esta escala al resultado esperado. El Grupo Control muestra diferencias significativas con respecto a alguno de los grupos en algunos de los bloques, como en el Bloque 3 con el Grupo T(I)-T(C), $t(94) = 7.515$, $p < .001$; el Bloque 4 con el Grupo T(C)-T(I), $t(94) = 9.116$, $p < .001$ y con el Grupo T(I)-T(C), $t(94) = -3.299$, $p = .012$. Llegados a este punto deja de tener diferencias significativas con ellos. El Grupo

Control muestra una puntuación media durante todo el proceso sin ninguna diferencia significativa entre sus bloques de ensayos ni con respecto al Juicio Global.

Por otro lado, al comparar los dos grupos experimentales en la valoración del Juicio Global, encontramos que las puntuaciones de ambos grupos son prácticamente las mismas en la zona central de la escala, $t(94) = 0.693$, $p = 1.000$, lo que nos indica que los participantes integran toda la información recibida durante todo el proceso cuando se les solicita que lo integren al finalizar, al igual que hace el Grupo Control, que tampoco difiere de los anteriores en el Juicio Global. Esto se corresponde con lo observado en el Experimento 1-A, así como en el de Matute y colaboradores (2002), donde a aquellos participantes a los que se les pedía un Juicio Global integraban toda la información. Los resultados de este experimento indican que el orden en el que se presenta la información en un juicio puede afectar al veredicto emitido, así como parece apreciarse un efecto de recencia al tener importancia la información que ven los participantes en los ensayos más recientes. Sin la pregunta explícita de integración del Juicio Global, los participantes priorizan la información recibida en la parte final de la tarea.

Discusión del Capítulo 4

En esta serie experimental pudimos comprobar y replicar en un contexto judicial los resultados observados por Matute y colaboradores (2002) en un contexto médico. En nuestros experimentos, el orden de los testimonios cuando se solicitaba un juicio ensayo a ensayo iba modificando la valoración de los participantes, que parecían dejarse llevar por los últimos ensayos mostrando así un efecto de recencia. Además, vimos que la valoración, conforme van pasando los bloques, se va modificando acorde al juicio anterior y que una pregunta final actúa como método integrador de toda la información anulando en cierto modo el efecto de recencia, como también muestran

investigaciones previas (Catena et al., 2004; Collins & Shanks, 2002; Vadillo et al., 2004).

En el Experimento 1-A no pudimos llegar a encontrar con nitidez estos resultados pues los últimos juicios emitidos por los dos grupos experimentales se quedaban en un punto medio de la escala, lo que nos llevó a volver a realizar el experimento, pero modificando la escala. A pesar de que algunos autores (Ng et al., 2023) han sugerido la conveniencia de usar escalas bidireccionales, nosotros optamos por utilizar una escala unidireccional en lugar de bidireccional por los motivos que mencionamos en la introducción de este capítulo y muy especialmente porque pensamos que el resultado menos marcado del Experimento 1-A podía deberse a que la puntuación de cero en la escala bidireccional podía inducir algún tipo de error y que los participantes confundieran esta puntuación con una valoración de inocencia.

El Experimento 1-B efectivamente sugiere que la escala unidireccional resultaba ser más sensible a la hora de recoger los juicios ensayo a ensayo y, en este caso, sí que recogimos un cruce claro en las valoraciones que confirmaban la presencia de un efecto de recencia en los juicios ensayo a ensayo, que desaparecía cuando pedíamos a los participantes realizar un Juicio Global integrando toda la información vista hasta el momento.

Sin embargo, un aspecto que nos preocupaba del experimento era que resultaba poco ecológico. Esto era una ventaja a nivel experimental porque facilitaba el control de las variables, pero lo alejaba de la situación que podían vivir los miembros de un jurado en la realidad, lo que nos condujo a nuestro siguiente experimento.

Capítulo 5. ForenPsy, Batería de Testimonios para Experimentación

Experimento 2. Supuestos judiciales y testimonios estandarizados

Una vez observado en los primeros experimentos que en el contexto judicial parece que las decisiones pueden verse afectadas por factores ajenos a la información dada como, por ejemplo, el orden en el que se presentan los testimonios, decidimos continuar investigando esta línea. Sin embargo, con el fin de seguir con los experimentos y tratando de hacer que se parecieran más a una situación real y necesitando sin embargo unos estímulos controlados, diseñamos la siguiente batería de testimonios, que llamamos ForenPsy, y que, para poder utilizarla, fue necesario realizar una estandarización previa de los ítems creados.

Método

Participantes y materiales. Reclutamos a un total de 60 personas (47% mujeres, 12% hombres, 42% no reportado; edades 18-58, *M* = 30, *DT* = 0.6). Todos ellos fueron reclutados en la Universidad de Deusto en Bilbao.

Los participantes realizaron la tarea con papel y lápiz de forma individual y autónoma en una sala tranquila de la universidad, junto a un experimentador que los acompañó por si surgía alguna duda.

El banco de testimonios consistió en 45 testimonios ficticios agrupados en tres supuestos judiciales: 15 de homicidio, 15 de amenazas y 15 de allanamiento. Estos tres delitos fueron elegidos porque son algunos de los que en España pueden recurrir a la presencia de un jurado popular (Ley Orgánica 5/1995, de 22 de mayo,

del Tribunal del Jurado). Esto resultaba importante para intentar hacer más ecológico el banco de testimonios y los futuros experimentos.

Para cada uno de los supuestos judiciales creamos una historia donde dábamos información sobre un suceso ficticio en el que una persona era sospechosa de haber cometido un delito. Para cada historia, creamos 15 testimonios provenientes de supuestos testigos llamados a declarar. Estos testimonios fueron diseñados de tal manera que unos sugerían inocencia y otros culpabilidad. El objetivo de esta estandarización fue comprobar si, efectivamente, los participantes valoraban cada testimonio de la forma que habíamos previsto. El Apéndice A muestra los tres supuestos judiciales y los 15 testimonios creados para cada supuesto.

Para la preparación de todos los textos, seguimos una serie de criterios. El primero de ellos es que los nombres, tanto de los acusados como de los testigos, eran solamente iniciales. Esto se hizo para evitar que otras variables demográficas tales como el género, la nacionalidad o la clase social influyeran en la decisión como demostraron, por ejemplo, Sealy y Cornish (1973) en jurados populares.

El segundo es que los testimonios fueron diseñados teniendo en cuenta algunas variables psicolingüísticas que pudieran afectar a su comprensión. Especialmente, los testimonios no diferían significativamente en longitud (*M* de número de letras = 63, *DT* = 18.8, rango = 64-116; *F(*14, 30) = 0.91, *p* = .561, y la facilidad con que se podía leer y comprender el texto o legibilidad del texto (*M* = 89.1, *DT* = 13.5, rango = 64-116; *F(*14, 30) = 1.05, *p* = .431, por lo que según los criterios de Fernández (1959) fueron considerados como textos de dificultad fácil.

Diseño y Procedimiento. Solicitamos a los participantes que leyeran cada uno de los tres supuestos judiciales seguido de los 15 testimonios correspondientes a cada uno de ellos. Mientras tanto, los participantes completaron dos tareas. Por un lado, debían evaluar cada testimonio como exculpatorio o inculpatorio; por otro lado, se les preguntó también por la importancia subjetiva que tendría para ellos ese testimonio.

Estas tareas son similares a las empleadas en estudios previos que simulan un contexto judicial (Engel et al., 2020), donde los investigadores utilizaban una historia escrita junto a una serie de testimonios de testigos que complementaban la historia. Para la primera pregunta que se les hacía de cada testimonio, se instruyó a los participantes a evaluar cada testimonio como exculpatorio o inculpatorio. Para la segunda pregunta, referente a la importancia subjetiva de la tarea del testimonio, se pidió a los participantes que calificaran la importancia del testimonio para emitir el veredicto en una escala tipo Likert (de 1, *muy poco,* a 5, *mucho*). La tarea la realizaban los participantes en un cuestionario creado para este propósito. Toda la sesión duró aproximadamente 10 minutos para cada participante.

Para evitar la influencia del orden de presentación de la información, realizamos seis modelos diferentes del cuestionario y los distribuimos al azar entre los participantes. En estos modelos se contrabalanceó el orden de presentación de los supuestos judiciales y los 15 testimonios de cada uno tal y como se muestra en la Tabla 2.

Tabla 2

Resumen del Contrabalanceo del Orden de Presentación de la Información en el Experimento 2

Modelo de cuestionario	Orden de supuestos	Orden de Testimonios
1	Hom - Ame – All	1-2-3-4-5-6-7-8-9-10-11-12-13-14-15
2	Ame - All – Hom	15-14-13-12-11-10-9-8-7-6-5-4-3-2-1
3	All - Hom – Ame	8-10-15-7-4-11-1-3-5-2-13-6-14-12-9
4	Hom - All – Ame	9-12-14-6-13-2-5-3-1-11-4-7-15-10-8
5	Ame – Hom – All	13-4-11-8-3-15-1-7-14-2-12-9-6-10-5
6	All – Ame - Hom	5-10-6-9-12-2-14-7-1-15-3-8-11-4-13

Nota. Hom: Homicidio; Ame: Amenazas; All: Allanamiento. Los testimonios concretos se representan aquí por el número que tienen asociado en el modelo del Apéndice A, que se corresponde con el Modelo 1.

Los participantes realizaron la tarea de forma individual en una sala de la universidad destinada a la misma, para evitar posibles distractores, mientras que un experimentador los acompañaba para resolver dudas en el caso de que las hubiera. Las variables dependientes fueron las siguientes: (a) Porcentaje de acuerdo: Se calculó el porcentaje de participantes que emitieron el juicio esperado para cada testimonio. Consideramos juicio esperado aquel juicio (de inocencia o culpabilidad) emitido por los participantes que coincidía con el juicio preestablecido o previsto por los experimentadores al momento de diseñar el testimonio. Por lo tanto, se espera que sea el juicio más frecuente entre los participantes. (b) Relevancia subjetiva: Para cada testimonio se calculó el valor medio en la escala de 5 puntos de la pregunta sobre la importancia subjetiva que conceden los participantes a ese testimonio. Los valores pequeños indican una relevancia baja y los valores grandes indican una relevancia alta del testimonio para determinar el veredicto. Esta medida se calculó independientemente de que el juicio emitido por el participante coincidiera o no con el juicio esperado.

Resultados y Discusión

Se examinaron las respuestas a cada testimonio con el fin de proporcionar datos sobre: a) el porcentaje de los participantes cuyo juicio concordaba con el Juicio Esperado por los experimentadores (inocencia o culpabilidad); y b) la puntuación media de la escala sobre la importancia subjetiva de ese testimonio.

En primer lugar, encontramos un acuerdo medio del 83% entre los juicios de los participantes y el Juicio Esperado. Es decir, como media, el 83% de las valoraciones de los participantes a los testimonios fueron tal y como esperábamos que fueran valorados. Solo las historias de amenaza presentan un acuerdo medio inferior al 80% siendo el acuerdo medio en los casos de allanamiento de 80.11% y en los casos de homicidio del 85.99%, tal y como se muestra en la Tabla 3.

Tabla 3

Porcentaje de Acuerdo entre el Juicio Esperado y el Juicio Emitido en cada testimonio, Porcentaje de Acuerdo Medio de cada Tipo de Historia y Puntuaciones Medias del Grado de Relevancia

	Homicidio			Amenazas			Allanamiento		
Test.	J.E.	%	Rel.	J.E.	%	Rel.	J.E.	%	Rel.
1	Culp.	88.33	3.32	Culp.	73.33	2.55	Culp.	80.00	2.95
2	Culp.	91.67	3.42	Culp.	70.00	2.55	Culp.	91.67	3.35
3	Inoc.	98.33	4.00	Culp.	73.33	2.77	Inoc.	98.33	3.90
4	Culp.	83.33	2.97	Culp.	88.33	3.43	Inoc.	48.33	2.32
5	Culp.	78.33	2.30	Inoc.	83.33	2.93	Inoc.	88.33	2.22
6	Culp.	93.33	3.67	Inoc.	85.00	2.38	Culp.	81.67	3.03
7	Culp.	86.67	3.28	Culp.	88.33	3.07	Culp.	83.33	3.25
8	Culp.	70.00	2.60	Inoc.	73.33	2.45	Inoc.	81.67	2.73
9	Culp.	83.33	2.75	Inoc.	83.33	2.50	Inoc.	80.00	2.33
10	Inoc.	100.00	3.25	Inoc.	60.00	2.52	Inoc.	96.67	3.07
11	Inoc.	98.33	3.00	Inoc.	63.33	3.00	Inoc.	86.67	3.02
12	Inoc.	95.00	2.82	Culp.	78.33	2.78	Inoc.	100.00	2.43
13	Inoc.	88.33	2.68	Culp.	76.67	2.53	Culp.	56.67	2.65
14	Culp.	85.00	2.83	Culp.	76.67	2.53	Culp.	75.00	2.77
15	Inoc.	50.00	2.50	Inoc.	80.00	2.58	Culp.	53.33	2.93
Media		85.99	3.03		76.89	2.70		80.11	2.86

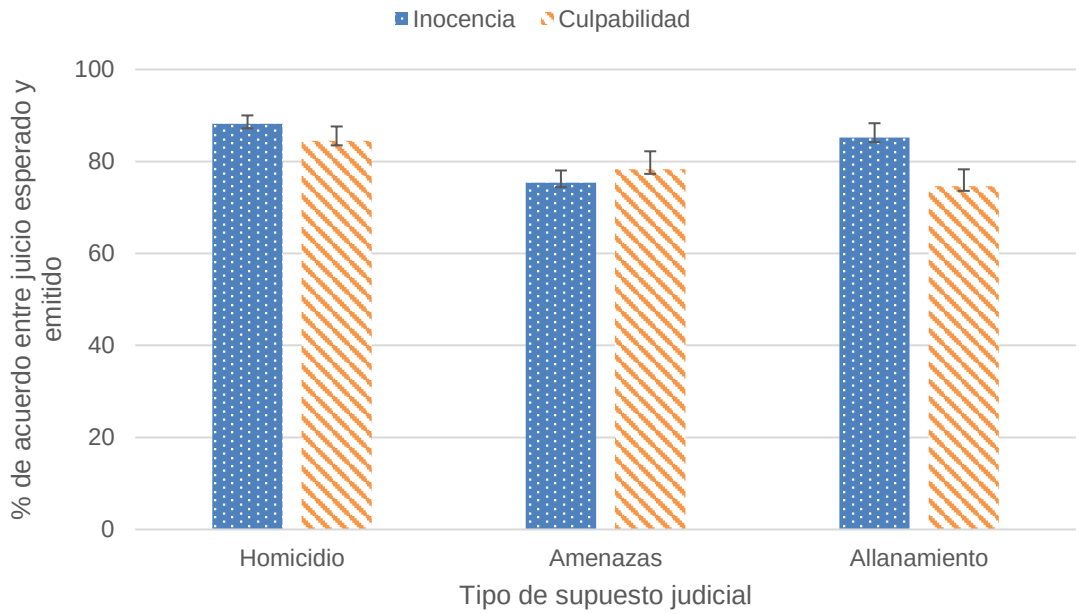
Nota. Test.: Testimonio; %: Porcentaje de acuerdo; J.E.: Juicio Esperado; Rel: Relevancia Subjetiva promedio; Culp.: Culpabilidad; Inoc: Inocencia

La Tabla 3 muestra también los porcentajes de acuerdo de cada testimonio individualmente. Como puede verse, la mayoría de los testimonios superan el 70% de acuerdo. Solamente seis testimonios tienen porcentajes de acuerdo inferiores al 70% y solo uno de ellos (el testimonio 4 del caso de allanamiento) obtiene un porcentaje de acuerdo inferior al 50%.

La Figura 11 muestra el porcentaje de acuerdo entre el Juicio Esperado por los investigadores y el juicio emitido por los participantes para cada supuesto judicial y tipo de testimonio.

Figura 11

Porcentaje de Acuerdo entre el Juicio Esperado y el Juicio Emitido



Nota. Las barras de error indican el error típico

Llevamos a cabo una serie de análisis ANOVA 3 x 2 (Supuesto judicial x Juicio Esperado) con el fin de conocer si el supuesto judicial (homicidio, amenazas u allanamiento) y el juicio esperado (inocencia o culpabilidad) tenía un efecto en las variables de porcentaje de acuerdo esperado y relevancia subjetiva. El ANOVA mostró que el supuesto judicial no afectó significativamente al porcentaje de acuerdo esperado $F(2, 39) = 1.965$, $p = .154$ ni a la puntuación de relevancia subjetiva del testimonio $F(2, 39) = 2.244$, $p = .120$. El tipo de testimonio (inocencia o culpabilidad) tampoco afectaba al porcentaje de acuerdo esperado $F(2, 39) = 1.036$, $p = .315$, ni a la puntuación de relevancia subjetiva $F(2, 39) = 0.985$, $p = .327$. Esto sugiere que los 45 testimonios funcionan de manera adecuada y similar, independientemente del supuesto judicial o si los testimonios fueron de inocencia o culpabilidad.

Discusión del Capítulo 5

La creación de este instrumento surgió como una necesidad de tener una herramienta que pudiéramos utilizar en los futuros experimentos de esta tesis. Con los resultados obtenidos, podemos asegurar que este instrumento es útil para el propósito que queríamos, ya que los participantes interpretan la mayoría de los testimonios creados exactamente como esperábamos y con un elevado porcentaje de acuerdo. Teniendo este instrumento diseñado y estandarizado ya fue posible seguir con una nueva línea de experimentos donde usamos estos testimonios como base para investigar en el contexto judicial.

Por otro lado, este instrumento ha sido rediseñado y mejorado en una segunda versión (Álvarez et al., 2025) donde los testimonios fueron reevaluados y reconfirmados con una muestra mucho mayor, además de haber creados seis casos nuevos. Además, este instrumento está en abierto y puede ser consultado y ampliado por cualquiera que lo desee a través de OSF.

Capítulo 6. Influencia de los Algoritmos en los Juicios Basados en Testimonios

Experimento 3. Influencia de algoritmos para modificar un juicio basado en testimonios

Habiendo comprobado en los Experimentos 1-A y 1-B que variables como el orden de presentación de los testimonios y la frecuencia con la que se pide el juicio pueden afectar a una decisión tomada en un juicio, decidimos seguir adelante con esta misma línea de investigación por lo que, con esta serie experimental buscamos extender los resultados obtenidos en la Serie Experimental 1 a un contexto en el que se utilizan IAs de apoyo a la decisión en el ámbito judicial. Además, con la creación de la batería ForenPsy en el capítulo anterior, ya disponíamos de una herramienta que nos permitía realizar experimentos de una manera más ecológica que los de la Serie Experimental 1.

Como hemos mencionado anteriormente en esta tesis, se está produciendo un auge en el uso de los algoritmos y las IAs en el ámbito judicial que, lejos de desaparecer, irá en aumento, y existe evidencia de que pueden influir en las decisiones (Agudo et al., 2024). Además, hemos mencionado que el uso de IAs puede tener varios efectos en los usuarios como son la apreciación del algoritmo (Logg et al., 2019) o el heurístico de la máquina (Sundar, 2008), que pueden provocar que los humanos tomemos decisiones sesgadas debido a atribuciones erróneas que hacemos de las IAs, como ya vieron también Agudo y colaboradores (2024).

En este experimento quisimos comprobar si la recomendación de un algoritmo podía alterar el veredicto emitido previamente por una serie de personas que simulaban

participar en un juicio como miembros de un jurado. Para ello, planteamos a los participantes un caso judicial con una serie de testimonios indicativos de culpabilidad o inocencia dependiente del grupo. Tras emitir su veredicto, les mostrábamos el veredicto de un supuesto algoritmo, que siempre era contrario a lo que sugieren los testimonios y se les ofrecía la posibilidad de cambiar su veredicto.

Método

Participantes y Materiales. En este experimento participaron 62 personas (82.3% mujeres, 17.7% hombres) estudiantes de primer curso del Grado de Psicología de la Universidad de Deusto con edades comprendidas entre los 18 y los 27 años (M = 18.9, DT = 1.55) y fueron asignados aleatoriamente a los grupos Ti-Ac (n = 30) y Tc-Ai (n = 32). Un análisis de sensibilidad muestra que, con el tamaño de muestra actual (n = 62), obtenemos una potencia de 0.80 para detectar un efecto de tamaño mediano (d = 0.36) en las diferencias entre grupos. Los participantes eran estudiantes de la asignatura Procesos Psicológicos Básicos II y realizaron el experimento como parte de la actividad docente. No estaban obligados a enviar los datos una vez completada la tarea, aunque como parte de la evaluación debían realizar un pequeño informe sobre lo que habían aprendido durante una clase práctica en la que, al finalizar la tarea, recibieron una explicación completa y didáctica sobre el experimento. El estudio fue realizado vía internet.

Los materiales utilizados fueron los testimonios estandarizados del instrumento ForenPsy desarrollado en el Experimento 2, correspondientes al supuesto judicial de amenazas y una tarea diseñada para los experimentos de esta tesis llamada *Tarea del Jurado.* Elegimos el supuesto de amenazas porque era el que dio resultados más estables y del que podíamos extraer siete testimonios robustos de cada tipo (inocencia y culpabilidad). De los 15 testimonios que se

estandarizaron del supuesto de amenazas en el Experimento 2, seleccionamos siete de cada tipo (inocencia o culpabilidad) para que fueran el mismo número de estímulos de cada tipo y descartamos el que menos claro era en su valoración. El experimento se realizó a través de *Google Forms* y los participantes fueron asignados aleatoriamente a los distintos grupos mediante un *script* diseñado en el laboratorio.

Diseño y Procedimiento. Los participantes se asignaron aleatoriamente a dos grupos que diferían en el tipo de testimonios que veían, tal y como se muestra en la Tabla 4. Todos leían primero el supuesto judicial. A continuación, el grupo Ti-Ac veía testimonios de inocencia y, posteriormente, el algoritmo (ficticio) le indicaba que su valoración era de culpabilidad; y el grupo Tc-Ai veía testimonios de culpabilidad y, posteriormente, el algoritmo le indicaba que su valoración era de inocencia. La sección de la tarea donde los participantes visualizaron los testimonios estaba formada por siete ensayos de entrenamiento presentados en el mismo orden para todos los participantes. La información sobre la valoración del algoritmo era siempre la contraria a lo correspondiente a los testimonios, es decir, si los testimonios indicaban inocencia, el algoritmo valoraría al acusado como culpable y viceversa. Cada participante debía emitir un juicio indicando el grado en que consideraba culpable al acusado en dos ocasiones. La primera ocasión (J1) era tras la presentación de los testimonios y la segunda ocasión (J2) tras la presentación de la valoración del algoritmo, tal y como se muestra en la Tabla 4, Fases 2 y 4. Para registrar el juicio de los participantes utilizamos una escala de 0 a 10 donde 0 era Definitivamente Inocente y 10 era Definitivamente Culpable.

Tabla 4

Resumen del Diseño del Experimento 3

Grupo	Fase 1	Fase 2	Fase 3	Fase 4
Ti-Ac	Ti	J1	Ac	J2
Tc-Ai	Tc	J1	Ai	J2

Nota. T: Testimonio; A: Algoritmo; i: Inocencia; c: Culpabilidad; J: Juicio

La información se les proporcionaba a los participantes en una serie de pantallas donde, primero, veían la información del caso, y a continuación veían los siete testimonios. Tras emitir el primer juicio veían la información que proporcionaba el algoritmo tras haber analizado, supuestamente, la información del caso. Esta información era precedida por una explicación sobre qué son los algoritmos de IA, que estos se están usando en justicia y se les ponía una noticia sobre *COMPAS*, el algoritmo utilizado en Estados Unidos para determinar la libertad provisional de algunos reclusos (véase Angwin et al., 2016; Larson et al., 2016) como ejemplo. A continuación, se solicitaba su segundo juicio sobre la culpabilidad del acusado.

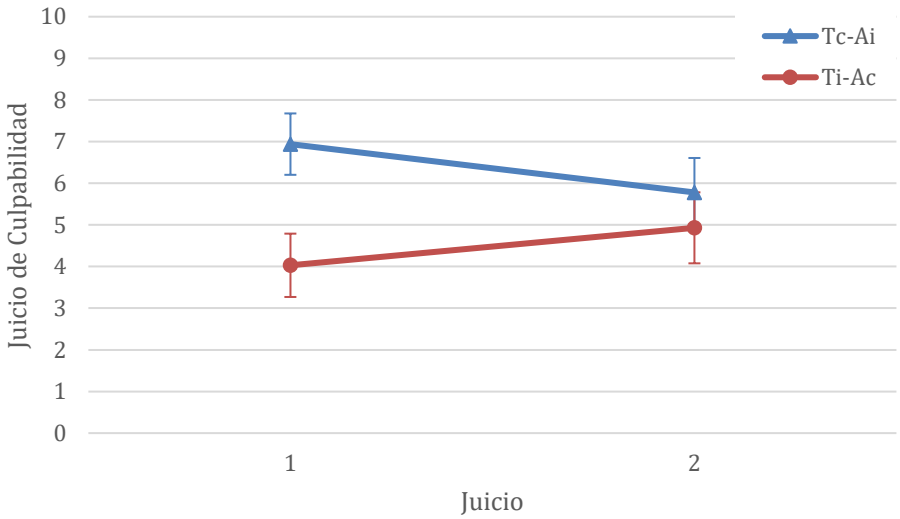
Para medir los dos juicios, se preguntaba a los participantes: "*Con toda la información que tienes y sabiendo que en el mundo real el testimonio de los testigos no es suficiente para condenar a una persona a prisión, en una escala de 0 a 10, donde 0 es Definitivamente Inocente y 10 es Definitivamente culpable, ¿cómo considerarías al investigado?".* En los extremos de cada escala aparecían sus valores, teniendo a la izquierda la etiqueta "Inocente" y a la derecha "Culpable".

Resultados y Discusión

En primer lugar, encontramos que el veredicto que daban los participantes inicialmente ante los testimonios (Juicio 1) concuerda con las valoraciones dadas a los testimonios en el experimento que hicimos para calibrar los testimonios de ForenPsy en el Experimento 2, tal y como puede verse en la Figura 12. Es decir, el juicio de culpabilidad es más elevado cuando los testimonios son de culpabilidad que cuando son de inocencia.

Figura 12

Juicios Medios de los Grupos en los Juicios 1 y 2 en el Experimento 3



Nota. T: Testimonio; A: Algoritmo; I: Inocencia; C: Culpabilidad; Las barras de error indican el error típico

En un primer análisis, observamos que el Grupo Ti-Ac, que recibió testimonios de inocencia, presentó una puntuación *M* = 4.03 (*DT* = 0.388) en el Juicio 1, mientras que el Grupo Tc-Ai, que recibió testimonios de culpabilidad, presentó una puntuación *M* = 6.94 (*DT* = 0.376). Es decir, tal y como esperábamos los testimonios que fueron estandarizados como testimonios de inocencia se valoraban con menor nivel de culpabilidad que los de culpabilidad en el primer juicio.

A continuación, realizamos un ANOVA mixto 2 x 2 (Grupo x Juicio) con el fin de determinar si se producían diferencias debido a la recomendación del algoritmo entre el Juicio 1 y el Juicio 2. En él no encontramos efecto principal en la variable Juicio, $F(1, 60) = .505$, $p = .480$, $\eta^2g = .001$, pero sí encontramos un efecto principal en la variable Grupo $F(1, 60) = 11.9$, $p = .001$, $\eta^2g = .151$, así como interacción entre las variables Juicio y Grupo, $F(1, 60) = 32.734$, $p<.001$, $\eta^2g = .051$. En primer lugar, al realizar las comparaciones Post-hoc con la corrección de Tukey, encontramos una diferencia significativa entre los dos grupos durante el Juicio 1, $t(60) = 5.38$, $p<.001$, lo que indica que los testimonios diseñados para ForenPsy funcionan adecuadamente.

También encontramos que la información contraria que proporcionaban los algoritmos sí modificaba posteriormente el juicio, haciendo que los dos grupos diesen puntuaciones más centrales en el Juicio 2. Es decir, al analizar las diferencias entre el Juicio 1 y el Juicio 2 en cada grupo, los que basándose en los testimonios de inocencia decían que el sujeto era inocente, al ver que el algoritmo lo contradecía, e indicaba culpabilidad, tendieron a suavizar su juicio de inocencia aumentando la culpabilidad, $t(60) = -3.47$, $p = .005$. Y lo contrario ocurría en el otro grupo, en el que habían visto testimonios de culpabilidad (y juzgado como tales) pero ahora el algoritmo indicaba lo contrario y por tanto los participantes se inclinaron hacia un

juicio de culpabilidad más moderado $t(60) = 4.61$, $p<.001$. Además, al comparar los dos grupos en el Juicio 2, no encontramos diferencias entre ambos grupos, $t(60) = 1.40$, $p = .505$; de modo que las diferencias que se habían observado inicialmente como respuesta al hecho de haber recibido testimonios diferentes, desaparecían al proponer el algoritmo una valoración contraria.

Estos resultados sugieren que la recomendación errónea o sesgada de un algoritmo podría llegar a influir sobre el veredicto de un jurado en casos en los que se podría presentar su análisis como una prueba forense. Ya lo hemos podido comprobar en la vida real con el caso de COMPAS (Angwin et al., 2016; Larson et al., 2016), donde un grupo de expertos estaban siendo afectados por un algoritmo sesgado a la hora de decidir sobre la libertad de una persona. En nuestro caso, una sola frase que indicaba la supuesta recomendación de un algoritmo sirvió para suavizar la decisión tomada y dirigirla hacia un veredicto concreto tras haber visto siete testimonios que indicaban el veredicto contrario.

Experimento 4. Influencia de un algoritmo y de alertar de errores en algoritmos sobre los juicios basados en testimonios

En el Experimento 3 mostramos que un algoritmo podía afectar a la valoración de unos participantes de un jurado popular simulado. Siguiendo con la línea del Experimento 3, en este experimento quisimos responder a la siguiente pregunta: Una vez emitido un juicio con información tanto de testimonios como de un algoritmo con información contraria a los testimonios y viendo que las personas modifican su valoración acertada para acercarse a la valoración contraria del algoritmo, ¿es posible modificar ese juicio? Algunos experimentos realizados en un contexto médico indican que una forma de prevenirlo es avisando a los sujetos de los posibles errores de los algoritmos (Vicente, 2024).

En este experimento, por lo tanto, intentaremos minimizar los errores de las personas influenciadas por el algoritmo, pero lo haremos presentando información sobre errores de los algoritmos de manera que los participantes revisen su juicio tras cometer el error de seguir la recomendación de un algoritmo contrario a los testimonios, y lo basen de nuevo en los testimonios observados. Además, aprovecharemos para replicar los resultados del Experimento 3, donde la influencia del algoritmo modificaba el juicio emitido por los participantes tras ver los testimonios.

Además, tal y como vimos en la Serie Experimental 1, el orden de presentación de la información, así como la frecuencia con la que se les pedía emitir un juicio resultaba relevante para la decisión tomada por los participantes, algo que conseguimos replicar de otros experimentos (Catena et al., 2004; Collins & Shanks, 2002; Matute et al., 2002; Vadillo et al., 2004) y que mostraba que los participantes se centraban en el último juicio emitido cuando la frecuencia del juicio era alta, provocando un efecto de recencia; mientras que si se les pedía un juicio global, éste tendía a integrar toda la información. Por tanto, con idea de asegurarnos que el efecto observado en el Experimento 3 no era exclusivo de las condiciones usadas en ese experimento, también quisimos ampliarlo a una situación en la que la información del algoritmo era proporcionada antes o después que los testimonios en vez de presentar siempre los testimonios primero, como hicimos en el Experimento 3. Esto lo hicimos en el Experimento 4 contrabalanceando el orden para evitar que los resultados pudieran deberse al orden en el que estaban presentados estos dos elementos de información.

Método

Participantes y Materiales. El número de participantes fue de 111 (82% mujeres, 18% hombres), repartidos en dos grupos: Tc/Ai-E (*n* = 60) y Ti/Ac-E (*n* = 51). Estos fueron estudiantes de primer curso del Grado de Psicología de la Universidad de Deusto (*M* de edad = 18.6, rango = 18-21, *DT* = 0.694). Un análisis de sensibilidad muestra que, con el tamaño de muestra actual (*n* = 111), obtenemos una potencia de 0.80 para detectar un efecto de tamaño mediano (*d* = 0.26) en las diferencias entre grupos. Los participantes eran estudiantes de la asignatura Procesos Psicológicos Básicos II y realizaron el experimento como parte de las prácticas de la asignatura, aunque no estaban obligados a enviar los resultados del experimento. Al finalizar recibieron una explicación completa y didáctica sobre el experimento y podían realizar, posteriormente, un pequeño informe explicando lo que habían aprendido con la práctica si querían ser evaluados.

Diseño y Procedimiento. El diseño de este experimento es similar al del Experimento 3, como se puede observar en la Tabla 5, salvo por los siguientes cambios. En primer lugar, la información que cada grupo vio antes del primer juicio se encontraba contrabalanceada: la mitad del grupo veía los testimonios antes que la información que presentaba el algoritmo mientras que la otra mitad veía la información del algoritmo antes que los testimonios. Esto lo hicimos para evitar que los resultados pudieran deberse al orden de presentación de estos dos elementos, ya que los experimentos de la Serie Experimental 1 mostraron que la variable orden podía tener gran importancia y en el Experimento 3 habíamos presentado siempre los testimonios primero. En segundo lugar, retrasamos la emisión del primer juicio hasta después de haber visto tanto la información de los testimonios como la información del algoritmo, para que tuvieran ya toda la información antes de emitir el

primer juicio y pudieran integrarla toda antes de emitir el juicio. Esto lo hicimos para evitar también que el hecho de emitir un juicio tras recibir la primera Fase de información (algoritmo o testimonios) pudiera afectar a los resultados debido al efecto de frecuencia del juicio visto en los Experimentos 1-A y 1-B donde los participantes se centraban en la última información valorada para emitir el siguiente juicio, así como a los resultados de Vadillo y colaboradores (2004) que mostraron que si solicitaban un juicio después de cada fase, el juicio tendía a reflejar especialmente lo aprendido en esa fase en vez de tender a integrar la información. Por tanto, hasta este punto en el que solicitamos a los participantes un juicio que integre ya la información del algoritmo y testimonios, los resultados deberían ser similares a los obtenido durante el Juicio 2 del Experimento 3 a no ser que el hecho de emitir un juicio tras la Fase 1, o el hecho de recibir antes los testimonios fueran factores críticos en los resultados obtenidos hasta ahora.

Por último, la manipulación más importante de este experimento es la que incluimos tras el primer juicio. En este caso, ya habían visto la información sobre el algoritmo y sobre los testimonios, cada grupo en diferente orden, por lo que, lo que hicimos ahora fue indicarles que los algoritmos podían sufrir errores para comprobar si esto servía para revertir el efecto de la influencia del algoritmo. Trabajos como el de Vicente (2024) han mostrado que, en ocasiones, es posible hacer que las personas cometan menos errores por seguir recomendaciones incorrectas de algoritmos si se les avisa de la posibilidad de errores de los algoritmos. En este caso trataremos de revertir el error que comete el algoritmo y que acerquen más su juicio al sugerido por los testimonios.

Además, los resultados previos se habían hallado en contextos y tareas muy diferentes, relacionadas con la salud, por lo que es necesario comprobar la eficacia de avisar de errores en contextos judiciales.

Tabla 5

Resumen del Diseño del Experimento 4

Grupo	Fase 1	Fase 2	Fase 3	Fase 4	Fase 5
Ti/Ac-E	Ti/Ac	Ac/Ti	J1	E	J2
Tc/Ai-E	Tc/Ai	Ai/Tc	J1	E	J2

Nota. T: Testimonio; A: Algoritmo; E: Información sobre errores del algoritmo; i: Inocencia; c: Culpabilidad; J: Juicio; /: Orden Contrabalanceado (cuando en la Fase 1 se muestran testimonios, en la Fase 2 se muestra al algoritmo y viceversa)

La manera de indicar a los participantes que el algoritmo podía tener errores era con una pantalla en la que aparecía la siguiente frase: "*Ahora que has terminado tu valoración, es importante que sepas que los algoritmos no son fiables por completo y muchos presentan errores. Por ejemplo, COMPAS, el que os hemos mostrado en la noticia, no resulta imparcial a la hora de emitir una recomendación, ya que la valoración cambia en función de la raza del recluso, su edad, incluso el barrio donde vive. Los reclusos de origen latino o afroamericano obtenían, por ejemplo, valoraciones más negativas que los caucásicos. También los que vivían en determinados barrios eran víctimas del sesgo de los algoritmos.*"

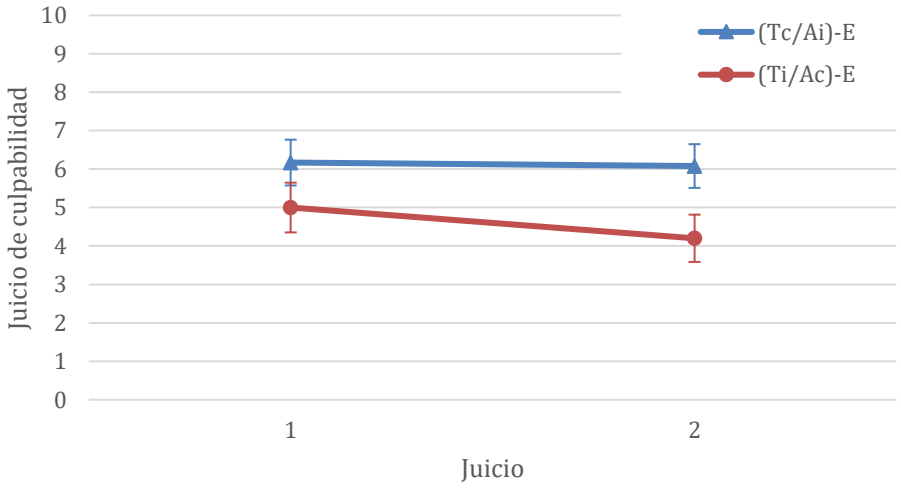
Tras presentarles esta información, recogíamos el Juicio 2, que esperábamos que estuviera esta vez menos basado en el consejo del algoritmo y más en los testimonios observados, debido a que la información del error debería hacer que el peso del algoritmo disminuyera o se valorara, al menos, de forma más crítica.

Resultados y Discusión

Para analizar los resultados del experimento mostrados en la Figura 13, realizamos un ANOVA 2 x 2 Grupo (Ti/Ac-E, Tc/Ai-E) x Juicio (Juicio 1, Juicio 2) donde obtuvimos un efecto principal en la variable Grupo, $F(1, 109) = 12.7$, $p<.001$, $\eta^2g = .100$, en la variable Juicio, $F(1, 109) = 26.1$, $p < .001$, $\eta^2g = .009$, así como interacción entre ambas: $F(1, 109) = 17.2$, $p<.001$, $\eta^2g = .006$.

Figura 13

Juicio Medio de los Grupos en los Juicios 1 y 2 en el Experimento 4



Nota. T: Testimonio; A: Algoritmo; E: Información sobre Errores del algoritmo; I: Inocencia; C: Culpabilidad; /: Orden Contrabalanceado; las barras de error indican el error típico

Tal y como sugiere la Figura 13, en las comparaciones Post-Hoc con la corrección de Tukey la diferencia entre ambos grupos en el Juicio 1, donde ya habían visto testimonios e información del algoritmo, muestra cierta tendencia a priorizar la información de los testimonios (inocencia vs. culpabilidad), aunque la

diferencia no alcanzó la significación estadística *t(*109) = 2.601, *p* = .051.Esto replica el resultado observado en el Juicio 2 del Experimento 3: cuando han visto los dos tipos de información, el juicio emitido por ambos grupos tiende a igualarse en puntuaciones centrales. Comprobamos así que se replican los resultados no solo cuando se presentan primero los testimonios, como en el Experimento 3, sino también cuando se contrabalancea el orden, como en el presente experimento. Comprobamos también que se replican los resultados, independientemente de que pidamos o no un juicio al participante tras la Fase 1 de información o esperamos a que reciba las dos fases, antes de solicitar el primer Juicio. No obstante, en este experimento se observa cierta tendencia a priorizar la información de los testimonios, aunque estas diferencias entre los grupos con testimonios de inocencia y testimonios de culpabilidad no resultaron estadísticamente significativas.

Lo más novedoso e interesante se observa en el Juicio 2, donde sí observamos diferencias estadísticamente significativas entre los grupos tras informarles de los errores de los algoritmos *t(*109) = 4.417, *p*<.001, Además, encontramos diferencias entre el Juicio 1 y el Juicio 2 en el Grupo Ti/Ac-E *t(*109) = 6.296, *p*<.001, pero no en el Grupo Tc/Ai-E *t(*109) = 0.708, *p* = .894. Estos resultados parecen indicar que la información que les proporcionamos sobre los errores del algoritmo sí parece afectar al juicio de los participantes, especialmente en el grupo Ti/Ac-E que vio testimonios de inocencia y la sugerencia de culpabilidad por parte del algoritmo. Esto podemos verlo en que, antes de mostrar la información sobre los errores del algoritmo, ambos grupos presentan un juicio similar en el Juicio 1, demostrando que, al igual que en el Experimento 3, tienden a integrar la información contraria que proporciona el algoritmo y esto les hace emitir un juicio intermedio a todos independientemente de lo que mostraran los testimonios. Sin embargo, en el Juicio 2, después de recibir información sobre los errores, es donde ya difieren

claramente los grupos y tienden a emitir un juicio más acorde a los testimonios que han recibido. También podemos verlo en la evolución del juicio en el Grupo Ti/Ac-E, donde, tras conocer la existencia de errores en los algoritmos parecen rechazar, al menos parcialmente, la información de culpabilidad proporcionada por el algoritmo y vuelven a darle más peso a la información de inocencia que sugerían los testimonios. Esto parece indicar que, al igual que ya se ha mostrado en contextos relacionados con la salud (Vicente, 2024), también en el contexto judicial el avisar de posibles errores de los algoritmos puede servir para que las personas sean más críticas y eviten cometer errores promovidos por los algoritmos, al menos cuando los algoritmos indican más culpabilidad que los testimonios.

En el Grupo Tc/Ai-E, sin embargo, parece que la información sobre los errores que cometen los algoritmos no afecta significativamente al Juicio 2. Esto resulta relevante de cara a un juicio real ya que parecería indicar que la valoración sobre culpabilidad creada tras la presencia de los testimonios y el algoritmo, si este último indica mayor grado de inocencia que los testimonios, es más difícil de reconducir hacia la culpabilidad que indicaban los testimonios, a pesar de haber pruebas de que los algoritmos cometen errores; al contrario de lo que ocurría en el grupo en el que el algoritmo sugería más culpabilidad, situación esta que sí se reconsidera adecuadamente al conocerse la existencia de errores en los algoritmos.

En resumen, la reversión del efecto del algoritmo tras conocer la posibilidad de errores se da en el grupo que ha visto testimonios de inocencia y un algoritmo inculpatorio. Cuando se les informa de que el algoritmo puede ser erróneo, el veredicto inicial de inocencia propiciado por los testimonios se ve reforzado. Sin embargo, los participantes que vieron testimonios de culpabilidad y un algoritmo que informaba de inocencia, dieron un veredicto que no se vio modificado por la

información sobre el error del algoritmo. Esto podría significar que los juicios son difíciles de cambiar hacia una mayor culpabilidad por el mero hecho de tener información que indique que el algoritmo que sugirió su inocencia podría estar equivocado. Una vez que las personas han emitido su juicio tras conocer el consejo del algoritmo y han integrado la información que sugiere que es menos culpable de lo que sugieren los testimonios, no se atreven a decir ahora que sí es culpable solo por saber que el algoritmo puede equivocarse. Y sin embargo sí hacen este cambio cuando el algoritmo había sugerido culpabilidad: en este caso, saber que puede equivocarse sí hace que rectifiquen hacia la mayor inocencia que sugerían los testimonios.

Experimento 5. Influencia de un algoritmo y del orden de la información sobre los juicios basados en testimonios

Siguiendo con la misma línea de los experimentos previos y utilizando los mismos testimonios estandarizados en el Experimento 2 sobre el supuesto judicial de amenazas, volvimos a realizar un experimento con algunas modificaciones para profundizar en la influencia del orden en el que se presentan los testimonios y el algoritmo, así como el momento en el que se pide que se emite el juicio, como ya hemos visto en los Experimentos 1-A y 1-B y que tratamos de controlar en el Experimento 4. Para ello, añadimos ahora un juicio intermedio. Este juicio intermedio decidimos solicitarlo porque en el experimento anterior no solicitamos un juicio ni hicimos un análisis entre la presentación de las dos informaciones. Algunos participantes vieron el veredicto dado por el algoritmo antes de ver los testimonios, mientras que otros vieron antes los testimonios y se les pidió a todos ellos su valoración global tras contar con ambas informaciones. En el experimento actual a todos los participantes se les pidió emitir un juicio entre ambas informaciones para

poder analizar mejor la influencia de cada una de ellas teniendo en cuenta la posible influencia del orden en el que se presentan.

Método

Participantes y Materiales. En este experimento participaron 148 estudiantes (81.8% mujeres, 17.6% hombres, 0.7% otro) con edades comprendidas entre los 18 y los 25 años (*M*= 19, *DT* = 1.34) de primer curso del Grado de Psicología de la Universidad de Deusto y repartidos aleatoriamente en los grupos Ac-Ti-E (*n* = 42), Ai-Tc-E (*n* = 30), Ti-Ac-E (*n* = 35) y Tc-Ai-E (*n* = 44). Un análisis de sensibilidad muestra que, con el tamaño de muestra actual (*n* = 148), obtenemos una potencia de 0.80 para detectar un efecto de tamaño mediano (*d* = 0.26) en las diferencias entre grupos. Los estudiantes formaban parte de los cuatro grupos de la asignatura Procesos Psicológicos Básicos II y realizaron el experimento como parte de la actividad docente. El experimento se realizó en horario de prácticas de la asignatura, siendo de libre elección el envío de los resultados y sin que afectara a la calificación de la asignatura. Posteriormente presentamos una explicación didáctica y completa sobre la investigación y podían enviar un informe de lo aprendido en la práctica si deseaban ser evaluados.

El experimento se realizó a través de *Google Forms* y los participantes fueron asignados aleatoriamente a los distintos grupos mediante un *script* diseñado en el laboratorio. Al igual que en el resto de los experimentos, utilizamos la tarea que hemos llamado “Tarea del jurado”.

Diseño y Procedimiento. El diseño del experimento era similar a los dos anteriores salvo por el hecho de que todos los participantes tenían que emitir juicios entre cada uno de los distintos elementos de información. El ordenador asignó a los

participantes aleatoriamente a cuatro grupos según el diseño mostrado en la Tabla 6. La sección de la tarea donde los participantes visualizaron los testimonios estaba formada por siete ensayos de entrenamiento presentados en el mismo orden para todos los participantes. Dos de los grupos visualizaban primero los testimonios, mientras que otros dos vieron primero la información y el veredicto del algoritmo; ortogonalmente, dos grupos veían testimonios de inocencia y los otros dos de culpabilidad.

Tabla 6

Resumen del diseño del Experimento 5

Grupo	Fase 1	Fase 2	Fase 3	Fase 4	Fase 5	Fase 6
Ti-Ac-E	Ti	J1	Ac	J2	E	J3
Tc-Ai-E	Tc	J1	Ai	J2	E	J3
Ac-Ti-E	Ac	J1	Ti	J2	E	J3
Ai-Tc-E	Ai	J1	Tc	J2	E	J3

Nota. T: Testimonio; A: Algoritmo; E: Información sobre errores del algoritmo; i: Inocencia; c: Culpabilidad; J: Juicio

Utilizamos un diseño mixto 3 (Juicio: J1, J2, J3) x2 (Orden de presentación de los testimonios y algoritmo: A-T vs T-A) x 2 (Orden del signo: c-i vs i-c). Al igual que en los experimentos anteriores, la información del algoritmo (culpabilidad o inocencia) siempre era contraria a lo correspondiente a los testimonios (estandarizados previamente en el Experimento 2). Cada participante debía emitir un juicio tras la presentación de los testimonios, otro tras la información sobre el algoritmo (aunque el orden de presentación de esta información cambiaba entre grupos) y un tercero tras la información sobre los errores que cometen los algoritmos. Es decir, en este experimento solicitamos 3 juicios. En los tres juicios

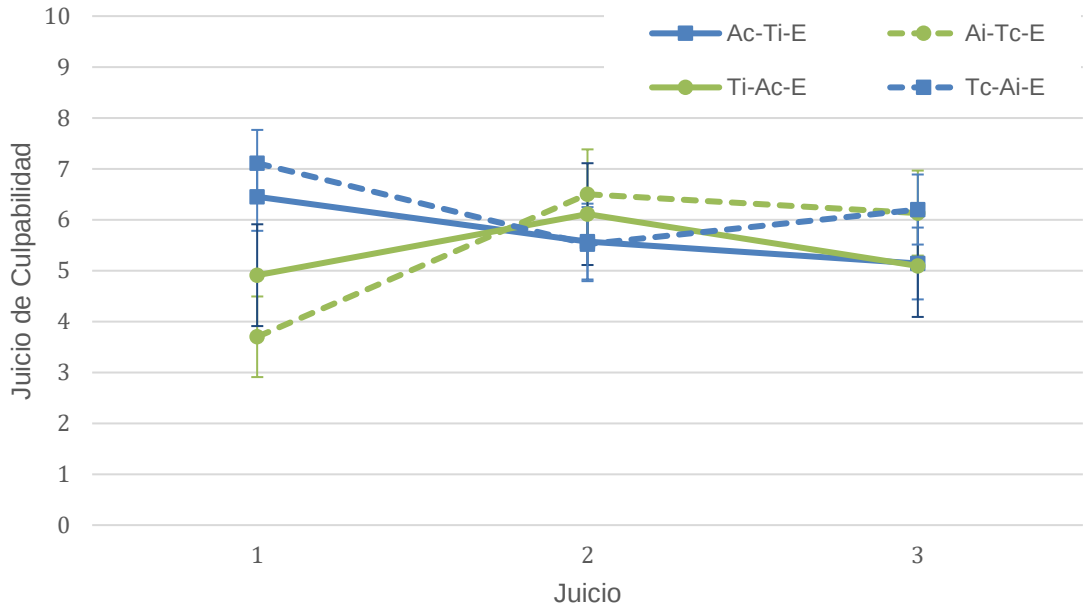
usamos una escala de 0 a 10, como la usada en los experimentos anteriores, donde 0 era definitivamente inocente y 10 era definitivamente culpable. En los juicios, se preguntó a los participantes "*Con toda la información que tienes y sabiendo que en el mundo real el testimonio de los testigos no es suficiente para condenar a una persona a prisión, en una escala de 0 a 10, donde 0 es Definitivamente Inocente y 10 es Definitivamente Culpable, ¿cómo considerarías al investigado?"*. En los extremos de la escala aparecían sus valores, teniendo a la izquierda la etiqueta inocente y a la derecha culpable.

Resultados y Discusión

Los resultados de este experimento se muestran en la Figura 14. Como puede observarse en la figura, los dos grupos cuya información inicial es de culpabilidad (independiente de la fuente de la que provenga; algoritmo o testimonio) puntúan más alto en la escala de culpabilidad en el primer juicio que los otros dos. Estas puntuaciones se igualan un poco en el segundo juicio tras ver la segunda información, que era contradictoria a la primera y, finalmente en el tercer juicio, las puntuaciones parecen orientarse más hacia las puntuaciones dadas en base a los testimonios cuando se les informa de los errores en los algoritmos. Esto último no sucede en el grupo que vio primero la información del algoritmo valorando como culpable, posiblemente porque en este grupo ya se había disminuido la culpabilidad que sugería el algoritmo al ver en segundo lugar los testimonios que sugerían inocencia.

Figura 14

Juicio de los Grupos en los Juicios 1, 2 y 3 en el Experimento 5



Nota: T: Testimonio; A: Algoritmo; E: Errores del algoritmo; I: Inocencia; C: Culpabilidad. Las barras de error indican el error típico.

Para analizar los datos que se muestran en la Figura 14 realizamos un ANOVA mixto 3 (Juicio: J1, J2, J3) x 2 (Orden de presentación de los testimonios y algoritmo: A-T vs T-A) x 2 (Orden del signo: c-i vs i-c). De entre las tres variables, solo encontramos efecto principal en la variable Juicio, $F(2, 294) = 3.25$, $p = .040$, $\eta^2g = .005$. Sin embargo, sí que encontramos interacción entre las variables Juicio y orden del signo, $F(2, 294) = 9.44$, $p<.001$, $\eta^2g = .014$; y Juicio y orden de presentación, $F(2, 294) = 7.71$, $p<.001$, $\eta^2g = .011$. También encontramos la triple interacción entre las tres variables: $F(2, 294) = 58.15$, $p<.001$, $\eta^2g = .080$.

En los análisis de las comparaciones Post-hoc de la triple interacción corregidas mediante el método de Tukey, encontramos algunos resultados interesantes y que confirman lo observado en los anteriores experimentos. En el grupo que vio testimonios de culpabilidad en primer lugar (Tc-Ai-E), encontramos diferencias entre el Juicio 1 y el Juicio 2 $t(147) = 4.859$, $p<.001$ tal y como vimos en el Experimento 3, confirmando que el algoritmo hace cambiar el juicio cuando el algoritmo sugiere inocencia. Sin embargo, esto no ocurre en el grupo que ve los testimonios de inocencia en primer lugar (Ti-Ac-E), que aumenta la culpabilidad al ver el algoritmo, pero esta diferencia entre los juicios 1 y 2 no llega a ser significativa $t(147) = -269$, $p = .058$. Es interesante constatar que, en los grupos donde se presentan los mismos estímulos en orden inverso ocurre lo mismo. Es decir, en el Grupo Ai-Tc-E encontramos diferencias significativas $t(147) = -7.0611$, $p<.001$ entre los Juicios 1 y 2, mientras que en el Grupo Ac-Ti-E no las encontramos $t(147) = 2.621$, $p = .274$.

Al comparar las puntuaciones entre los juicios 2 y 3, entre los cuales se informaba que los algoritmos cometían errores, solo encontramos diferencias significativas en el Grupo Ti-Ac-E, $t(147) = 4.305$, $p =.002$, lo que replicaría los resultados del Experimento 4. Sin embargo, en el otro grupo donde también se esperaría una diferencia significativa entre ambos juicios (Tc-Ai-E) no llega a ser significativa $t(147) = -3.199$, $p = .071$, aunque la puntuación sí que se ve modificada. Este efecto también lo vimos en el Experimento 4. En el Grupo Ac-Ti-E $t(147) = 1.9647$, $p = .716$ y en el Grupo Ai-Tc-E $t(147) = 1.4207$, $p = .958$ no encontramos diferencias significativas entre los juicios 2 y 3, como podíamos esperar si tenemos en cuenta que, al “anular” la valoración del algoritmo, la puntuación en el Juicio 3 debería ajustarse más a la del momento en el que vieron los testimonios, que coincide con el Juicio 2.

En cambio, al comparar los juicios 1 y 3, deberíamos encontrar diferencias en los grupos que vieron la información del algoritmo en primer lugar, y es exactamente lo que encontramos en el Grupo Ac-Ti-E $t(147) = 4.219$, $p = .002$ y en el Grupo Ai-Tc-E $t(147) = -6.625$, $p<.001$, mientras que en en los grupos Ti-Ac-E $t(147) = -0.504$, $p = 1.000$) y Tc-Ai-E $t(147) = 2.997$, $p = .120$ no encontramos diferencias significativas, tal y como esperábamos.

Cuando comparamos los distintos grupos en los distintos juicios, encontramos, como es lógico, que en el Juicio 1 hay diferencias significativas entre los grupos que ven información de inocencia y los que ven primero información de culpabilidad, como sucedía en el Experimento 3. Por ejemplo, el Grupo Ai-Tc-E presenta diferencias significativas con el Grupo Ac-Ti-E $t(147) = 5.2030$, $p<.001$ y con el Grupo Tc-Ai-E $t(147) = -6.515$, $p<.001$, pero no con el Grupo Ti-Ac-E $t(147) = -2.205$ $p = .549$. Del mismo modo, el Grupo Ti-Ac-E presenta diferencias significativas con el Grupo Tc-Ai-E $t(147) = -4.338$ $p =.001$, aunque no con el grupo Ac-Ti-E, que es el único que no presenta diferencias significativas con un grupo con el que sí debería. Es decir, mientras que con el Grupo Tc-Ai-E no presenta diferencias tal y como se espera $t(147) = -1.385$, $p = .965$, con Ti-Ac-E sí que debería presentarlas y no lo hace $t(147) = 3.037$, $p = .109$.

Al comparar el Juicio 2 de los diferentes grupos, no encontramos diferencias significativas entre ninguno de ellos, encontrando todos los valores con $p > .876$, confirmando lo ya visto en los Experimentos 3 y 4 de que al integrar la información contradictoria del algoritmo y la de los testimonios los participantes tienden a dar puntuaciones centrales. Esto se mantiene también en este experimento a pesar de los cambios realizados. Durante el Juicio 3, tampoco encontramos diferencias significativas entre los 4 grupos al avisarles del posible error del algoritmo, obteniendo

todos los valores de p >.610, indicando que los grupos se mantienen en puntuaciones cercanas entre ellos.

Podemos concluir que parece que los participantes no le dieron más importancia a la información de los algoritmos que a la dada por los testimonios, sino que la valoraban de manera equivalente a la de los testigos y lo realmente importante era el momento en el que veían la información de culpabilidad o de inocencia, lo cual sigue siendo preocupante de cara a la influencia que los algoritmos pueden ejercer en los juicios ya que los participantes parece que le dieron la misma importancia a la información que les daba el algoritmo que a toda una serie de testimonios supuestamente humanos . Además, es importante resaltar la importancia de que los participantes conozcan que los algoritmos pueden cometer errores, ya que, al hacerlo, los participantes parecieron quitar cierto peso al consejo algorítmico y redujeron el error, salvo en el grupo Tc-Ai-E. Sin embargo, esta pérdida de valor del consejo algorítmico no hacía que las puntuaciones se modificaran lo suficiente a cómo serían en el caso de solo tener en cuenta los testimonios de los testigos.

Discusión del Capítulo 6

Durante los tres experimentos de esta serie experimental hemos tratado de descubrir si la presencia de algoritmos en el contexto judicial podría venir acompañada de un posible sesgo de autoridad en el algoritmo, como otros autores sugieren (Araujo et al., 2020; Sundar, 2008; Sundar & Kim, 2019) que hiciera a los jurados tomar decisiones diferentes a las que toman cuando se basan solo en los testimonios.

En todos los experimentos encontramos resultados que confirman, al menos en parte, que los testimonios que fueron estandarizados en el Experimento 2 funcionan correctamente sirviendo para orientar la decisión de los participantes hacía

un veredicto u otro dependiendo de los testimonios que presentamos. Además, sobre el objetivo de comprobar si la información proporcionada por los algoritmos puede influir en el juicio de los miembros de un jurado, nos encontramos con resultados muy interesantes.

Ya en el Experimento 3 observamos claramente una diferencia en los dos grupos en función del tipo de testimonios que vieron (aquellos que vieron testimonios de inocencia y de culpabilidad) durante el primer juicio, que era cuando no tenían información adicional por parte del algoritmo. El grupo que vio los testimonios de inocencia presentó una puntuación media de culpabilidad más baja en el primer juicio que el grupo al que le mostramos los testimonios de culpabilidad.

En este experimento, encontramos, además, que los participantes que vieron testimonios de inocencia y a los que el algoritmo les sugirió que el acusado era culpable puntuaron más alto en culpabilidad en el segundo juicio después de haber visto la valoración del algoritmo. Sin embargo, aquellos que vieron testimonios de culpabilidad y luego vieron que el veredicto del algoritmo era de inocencia redujeron su valoración de culpabilidad en el segundo juicio. Es decir, las valoraciones de los grupos cambiaron en la dirección sugerida por el algoritmo. Esto parece confirmar nuestra hipótesis original de que la información de los algoritmos puede afectar al juicio.

En el Experimento 4 no era posible comprobar en el primer juicio si seguíamos obteniendo los mismos resultados ya que los participantes veían tanto la información de los testimonios como la valoración contraria del algoritmo antes de emitir el primer juicio. Sin embargo, si tenemos en cuenta que, en el segundo juicio, el veredicto del algoritmo perdía valor (las puntuaciones de los participantes volvían a dirigirse hacia la información que pretendían dar los testimonios) debido a que les proporcionamos la

información sobre los errores que cometen los algoritmos, sí que obtenemos diferencias entre los dos grupos en función de los testimonios de inocencia o de culpabilidad, lo que apoyaría los resultados del experimento anterior.

En este experimento encontramos, además. que, si avisamos de que la información de los algoritmos puede ser errónea, esta se reconsidera, y se ve modificada, al menos en parte. En nuestros resultados, encontramos que el grupo que ha emitido un veredicto tras haber visto testimonios de inocencia y un algoritmo culpabilizador, cuando luego le informamos que los algoritmos presentan errores a veces, refuerza su juicio inicial hacia la inocencia (dando una puntuación más orientada a la misma), tal como sugerían los testimonios y tal como esperábamos. Sin embargo, lo que encontramos también en este experimento fue que, cuando es el caso contrario (testimonios de culpabilidad y un algoritmo indicando que la persona es inocente), el juicio no cambia cuando los participantes saben que los algoritmos pueden fallar. Esto podría significar que los juicios son difíciles de cambiar hacia una mayor culpabilidad si los participantes ya han decidido que la persona es culpable y no hay nada que lo contradiga, por lo que no sería necesario cambiar el juicio. Sin embargo, en el caso de inculpar a una persona por la recomendación de un algoritmo, si no se cambia el veredicto podría significar condenar a un inocente, por lo que el juicio sí que se ve modificado.

En el Experimento 5 volvimos a obtener los resultados esperados. En este experimento, dos de los grupos (Ti-Ac-E y Tc-Ai-E) partían del mismo punto que los dos grupos del Experimento 3: en el primer juicio que emitieron solamente vieron la información que le proporcionaban los testimonios. Y, volvimos a encontrar que existían diferencias entre los grupos en función de que vieran testimonios de inocencia o de culpabilidad, como las que encontramos en el Juicio 1 del Experimento 3.

En este experimento nos encontramos también que, cuando cambiamos el orden de presentación de los testimonios y del algoritmo teniendo los participantes que emitir un juicio después de cada fase de información, resulta más importante el tipo de información que se proporciona (inocencia o culpabilidad), que la fuente de donde provenga esta información (testimonios o algoritmo). Es decir, para los participantes resulta igual de importante la información que proviene de un supuesto algoritmo que la que proviene de supuestos testigos cuyos testimonios son coincidentes.

Por otro lado, en el Experimento 5, al igual que en los experimentos anteriores, seguimos encontrando, en algunos de los casos, un cambio en la valoración de los juicios una vez que los participantes ven la segunda información ya sea de unos testigos o de un algoritmo. Esto indica que la información dada por el algoritmo (una sola frase) tiene el mismo valor en esos casos para los participantes que un conjunto de testimonios (siete testimonios estandarizados y con una clara orientación). Esto resulta de especial importancia ya que sucede en los dos casos en los que los participantes ven testimonios de culpabilidad y el algoritmo indica inocencia. Que sea en estos casos cuando encontramos diferencias significativas durante las dos fases puede deberse a que los testimonios inculpatorios son muy potentes.

Un algoritmo ficticio y sumamente sencillo como el nuestro no ha sido capaz de ensombrecer por completo la información de los testimonios. Sin embargo, a ojos de los participantes parece tener en algunos casos la misma importancia que el conjunto de siete testimonios, o de la propia valoración personal que cada participante hace de los testimonios. Un algoritmo real, además, en una situación real de un juicio, podría probablemente tener una influencia mayor aún que la que aquí estamos mostrando, ya que podría ser visto como un experto en el tema al que, como ya hemos visto, se le atribuyan unas capacidades superiores para tomar decisiones objetivas en ese tipo de

situaciones, como argumentaron Araujo y colaboradores (2020). Existen pruebas estandarizadas para evaluar la credibilidad de los testigos expertos (Brodsky et al., 2010) y tal vez debiéramos empezar a plantear que los algoritmos actúan como expertos en el tema de cara a cómo los perciben los participantes, por lo que tendríamos que buscar una forma de evaluar el grado de confianza que tienen las personas hacia los algoritmos antes de que se dejen aconsejar por ellos, sea en el contexto que sea.

Por último, tras conocer que los algoritmos pueden tener errores, la valoración de los participantes cambia adaptándose hacia la información que proporcionan los testimonios, por lo que creemos que sería muy necesario proporcionar esta información siempre que se utilicen algoritmos en justicia.

Con toda esta información, podemos concluir que los algoritmos pueden llegar a tener un gran peso en Justicia. Aunque en nuestros experimentos, con contexto judicial simulado y algoritmos ficticios, no mostramos que la información de los algoritmos resulte más potente que los testimonios a la hora de valorar la inocencia o culpabilidad de una persona, sí que resulta igual de importantes a ojos de los participantes que su propia valoración de los testimonios y es capaz de hacer cambiar en consecuencia sus valoraciones. En una situación intimidatoria y dotada de cierta autoridad, como puede ser un jurado, no podríamos descartar que la recomendación del algoritmo sea incluso más importante que la de los testimonios.

Parte III. Discusión General

Capítulo 7. Discusión General

Como ya comentamos en la introducción, en los últimos años los algoritmos de IA han pasado de ser una herramienta exclusiva de campos como la informática y la ingeniería, a convertirse en elementos omnipresentes que moldean casi todos los aspectos de la vida cotidiana. Desde la sanidad (Chen, 2020), hasta la educación (Chan, 2018), y los seguros (Obermeyer et al., 2019), pasando por los sectores más cotidianos como las redes sociales (Diakopoulos & Koliska, 2017), la música (Bovenkamp, 2017), las citas (Duportail, 2019) o las compras (Chen et al., 2016), los algoritmos se han integrado de forma profunda en la sociedad. Su impacto es tan amplio que, en muchos casos, las personas toman decisiones diarias influenciadas directamente por ellos, desde la elección de productos (Chen et al., 2016) hasta el manejo de nuestra salud (Obermeyer et al., 2019) o las recomendaciones que nos orientan sobre relaciones personales (Agudo y Matute, 2021).

Este fenómeno se ha vuelto aún más significativo con la creciente evolución de la IA que ha permitido a los algoritmos no solo realizar tareas repetitivas, sino también tomar decisiones autónomas que afectan a áreas cruciales como la justicia, la seguridad pública y la economía. Sin embargo, esta dependencia de los algoritmos en ámbitos tan importantes hace que nos hagamos preguntas sobre su transparencia, imparcialidad y fiabilidad, especialmente cuando las decisiones que toman tienen repercusiones directas sobre las personas y la sociedad.

El ámbito judicial es extremadamente sensible a la influencia de los algoritmos, como hemos visto con casos como el de las cárceles de Estados Unidos (Angwin et al., 2016). Ya se están utilizando algoritmos de IA en el ámbito judicial en multitud de países como, por ejemplo, en Argentina (Ministerio Público Fiscal de la Ciudad

Autónoma de Buenos Aires, 2020), China (Wei, 2019), España (Capdevila et al., 2015), Estados Unidos (Berkman Klein Center, 2022), Estonia (Niiler, 2019) y el Reino Unido (Ministry of Justice, 2013), donde actúan como asistentes de seres humanos. Las decisiones tomadas con ayuda de IAs en los tribunales afectan profundamente la vida de las personas: penas de prisión, multas o decisiones de custodia, por lo que, en este contexto, la automatización de decisiones judiciales podría ser beneficiosa por motivos como la reducción de la carga de trabajo y la agilización de los procesos, pero también genera serias preocupaciones.

Los algoritmos suelen ser un sistema opaco cuya lógica interna es difícil de interpretar. Esto plantea un riesgo significativo cuando tales algoritmos se aplican en contextos de alto impacto, como el judicial, donde la falta de comprensión del proceso que hay detrás de las decisiones puede provocar que la población pierda la confianza en el sistema de justicia. Además, si los algoritmos están mal diseñados o no son auditados de manera adecuada, pueden amplificar los sesgos existentes en los datos que procesan. Un algoritmo que se alimenta de datos históricos sesgados puede replicar y perpetuar estas injusticias, afectando a grupos vulnerables de manera desproporcionada, como demostraron Angwin et al., (2016) o Buolamwini (2016), por ejemplo.

Uno de los problemas más evidentes del uso de algoritmos en la justicia es el riesgo de sesgo algorítmico. Este fenómeno se produce cuando un algoritmo, basado en datos históricos que estén sesgados, produce resultados que favorecen a ciertos grupos o individuos en detrimento de otros. El sesgo algorítmico ha sido identificado en diversas áreas, como la selección de candidatos para empleos (Chen, 2023), la evaluación del riesgo de reincidencia de reclusos (Angwin et al., 2016), y el análisis de solicitudes de crédito (Cheney, 2016). En el contexto judicial, esto podría traducirse en

decisiones de culpabilidad o sentencias que favorecen injustamente a ciertos grupos raciales, de género o socioeconómicos (Larson et al., 2016), entre otros, lo que puede resultar en un ciclo continuo de desventaja y discriminación.

En este sentido, la implementación de algoritmos en la justicia plantea una serie de preguntas éticas fundamentales (Ponce, 2024). ¿Es justo delegar decisiones judiciales a sistemas que pueden estar influenciados por los mismos sesgos que se busca erradicar? ¿Cómo puede garantizarse que los algoritmos no refuercen las desigualdades sociales existentes? Estos son algunos de los dilemas clave que las sociedades deben enfrentar al considerar el uso de inteligencia artificial en los tribunales.

Una cuestión crucial que este estudio ha explorado es el impacto directo de los algoritmos en la toma de decisiones judiciales. Como se ha visto en investigaciones previas, la manera en que se presenta la información a los jueces o al jurado puede influir de manera significativa en el veredicto final. Los efectos de primacía y recencia (Pennington, 1982; Wells et al., 1985) son ejemplos de cómo algo como el orden de exposición puede modificar la evaluación global de un caso. También se ha observado que el hecho de ofrecer la recomendación de la IA antes o después de que el humano emita su veredicto puede ser una variable crítica (Agudo et al., 2024).

La introducción de un algoritmo como parte de la cadena de toma de decisiones en este contexto podría tener consecuencias aún más relevantes. Un algoritmo bien diseñado podría proporcionar una evaluación objetiva y rápida de los datos, pero la tendencia humana a confiar en la "autoridad" de las máquinas, incluso cuando estas cometen errores, puede generar un efecto de sesgo en los jurados o jueces. Esto podría alterar la percepción de culpabilidad, como se observa en los experimentos realizados en esta tesis, donde los participantes mostraron una

disposición a cambiar su juicio dependiendo de la información proporcionada por un supuesto algoritmo, incluso si esa información contradecía los testimonios presentados.

La influencia del algoritmo no se limita a la presentación de pruebas, sino que puede extenderse al propio proceso de toma de decisiones. Al considerar que los algoritmos son inherentemente más precisos o imparciales (Cummings, 2004 o Kathleen et al., 1996) que los seres humanos, los jurados o jueces pueden sentirse inclinados a basar sus decisiones en los veredictos algorítmicos sin cuestionarlos adecuadamente, lo que podría poner en riesgo el sistema judicial al completo.

El reto fundamental aquí es cómo integrar los algoritmos en el sistema judicial de manera ética y eficaz, de tal forma que se utilicen como una herramienta complementaria y no como sustituto del juicio humano. Como los resultados experimentales de esta tesis sugieren, los algoritmos tienen el potencial de alterar significativamente las decisiones judiciales, pero también sugieren que es posible aprender a usarlos con precaución y con ciertas garantías. La clave está en garantizar que los actores judiciales comprendan el funcionamiento de los algoritmos, que los algoritmos sean transparentes en su implementación y que existan mecanismos de control y revisión para evitar abusos o errores sistemáticos.

La toma de decisiones es un proceso fundamental en la vida cotidiana, pero se vuelve especialmente crítico en contextos como el judicial, donde se deben ponderar pruebas, testimonios y otros elementos para llegar a un veredicto. El hecho de que en el sistema judicial nos dejemos llevar por los mismos sesgos que en el resto de aspectos de nuestra vida (Kahneman, 2011), es lógico y nos hace pensar que aquí también vamos a tomar decisiones en base a criterios tales como, por ejemplo, confiar en la recomendación de un experto. La confianza en el algoritmo (Araujo et al., 2020)

la podríamos considerar una extensión del sesgo de autoridad propuesto por Milgram (1963) en el que tendemos a obedecer y confiar ciegamente en una figura de autoridad (autoridad algorítmica en este caso). Como hemos mencionado a lo largo de la tesis, a los algoritmos de IA se les atribuyen una serie de características que hacen que caigamos fácilmente en este sesgo.

En el ámbito de la psicología judicial, diversos estudios han demostrado que las decisiones de los jueces y jurados pueden estar influenciadas por una variedad de factores ajenos a la evidencia objetiva. Tanto Pennington (1982) como Wells y colaboradores (1985) son algunos de los que estudiaron cómo el orden de la presentación influye en los juicios. Sus estudios sugieren que el orden de presentación de la información jugaba un papel crucial en este tipo de decisiones. Otro factor importante cuando se utilizan algoritmos de IA es el momento en el que se presente la recomendación de la IA, de modo que presentarla después de que el humano responda, en vez de antes, puede minimizar los sesgos (Agudo et al., 2024). Un hallazgo importante en todos estos estudios es que los jurados no solo se basan en la lógica o en un análisis detallado de las pruebas, sino que también son influenciados por factores que pueden parecer irracionales o que no tienen una relación directa con los hechos presentados. Por ejemplo, los testimonios previos o incluso el estilo de los abogados pueden modificar la percepción del jurado sobre la credibilidad de los testigos o la relevancia de las pruebas presentadas, afectando a la decisión final.

En el contexto judicial, el efecto de primacía y el efecto de recencia pueden tener consecuencias significativas (Costabile & Klein, 2005; Furham, 1986; Pennington, 1982). Por ejemplo, si un jurado escucha un testimonio a favor de la inocencia del acusado al principio, podría estar predispuesto a ver con más escepticismo cualquier testimonio posterior que sugiera su culpabilidad. De manera

similar, los testimonios más recientes podrían tener un peso desproporcionado, a pesar de que otros testimonios presentados anteriormente sean más sólidos o relevantes. Como vimos durante la introducción, había experimentos que indicaban resultados en ambos sentidos, y no están claras las variables que producen uno u otro resultado, aunque parece que el efecto de la frecuencia de juicio puede tener una alta influencia en este sentido. Por todo ello, decidimos comprobarlo por nosotros mismos en la primera serie de experimentos de esta tesis investigando las dos variables, el orden de la información y la frecuencia del juicio, para conocer la influencia de estas dos variables en un contexto similar al que utilizamos posteriormente en los demás experimentos de esta tesis.

La primera serie de experimentos (Experimentos 1-A y 1-B) se realizaron con la idea de comprobar si la metodología utilizada en experimentos previos realizados por nuestro equipo era replicable usando un contexto judicial, a la vez que buscábamos resultados similares a los experimentos realizados en Justicia (Costabile & Klein, 2005; Furham, 1986; Pennington, 1982; Wells et al., 1985) en los que había resultados que indicaban que el orden de presentación de pruebas, declaraciones o discursos, además de la frecuencia con la que se solicita el juicio (Pennington & Hastie, 1992) afectaba a la valoración dada por los participantes.

En estos experimentos obtuvimos resultados que concluían, por una parte, que la metodología utilizada por Matute et al. (2002) en experimentos previos funcionaba de la misma manera cambiando los estímulos a unos más cercanos al contexto judicial y, por otra parte, que el orden de presentación de los estímulos y la frecuencia del juicio afectaba a la decisión de los participantes. En ambos experimentos, cuando pedíamos el juicio ensayo a ensayo la valoración de culpabilidad de los participantes de los grupos experimentales iba aumentando o disminuyendo con cada uno de los

ensayos que se les iba presentando, cambiando rápidamente su valoración cuando empezaron a ver ensayos que indicaban lo contrario a lo que habían estado valorando hasta ese momento, lo que es un indicativo de un efecto de recencia.

Por otro lado, en el momento en el que, al final, se les pedía integrar toda esa información para emitir un veredicto, lo hacían correctamente, emitiendo unas valoraciones alrededor de la puntuación media de la escala. Además, gracias al Experimento 1-B pudimos ver que una escala unidireccional parecía funcionar mejor que una bidireccional a la hora de registrar la influencia del orden en la decisión de los participantes.

Sin embargo, en los experimentos mencionados anteriormente (Costabile & Klein, 2005; Furham, 1986; Pennington, 1982; Wells et al., 1985), los investigadores encontraban distintos efectos: algunos encontraban efecto de primacía y otros efectos de recencia. Nosotros encontramos con nuestro procedimiento que nuestros participantes, aunque mostraban efecto de recencia, pues su valoración la iban modificando conforme les íbamos mostrando ensayos, al final integraban toda la información que habían visto durante toda la tarea, resultado que coincidía con lo publicado por Matute y colaboradores (2002). Esto puede deberse a la forma en la que nosotros les pedimos que respondan, ya que les pedimos explícitamente que, una vez vista toda la información y emitidos todos los juicios parciales, emitan un Juicio Global teniendo en cuenta toda la información que han visto. Esto no parece hacerse en ninguno de los experimentos previos, por lo que podría ser el factor que hace que obtengamos una respuesta distinta a la de los demás experimentos.

Aunque obtuvimos resultados interesantes en estos primeros experimentos, nos parecía que los nuestros eran más artificiales que los elaborados en la literatura de psicología jurídica, que utilizaban actuaciones o transcripciones basadas en juicios

reales, así que nos propusimos hacer una nueva serie de experimentos que resultaran más ecológicos, por lo que necesitábamos tener unos materiales más realistas que utilizar. Para ello, desarrollamos una serie de historias ficticias de diversos delitos con sus correspondientes testimonios, pero necesitábamos saber de forma exacta lo que estas historias y sus correspondientes testimonios iban a significar para los participantes, ya que en los trabajos previos (Costabile & Klein, 2005; Furham, 1986; Pennington, 1982; Wells et al., 1985), no parece que estandarizaran los instrumentos, sino que utilizaban unas historias que ya habían tenido un veredicto y una sentencia, bajo el supuesto de que resultarían en los mismos resultados. Por ello, nosotros decidimos diseñar una herramienta y estandarizarla.

En relación con esta herramienta (ForenPsy; Experimento 2), realizamos la estandarización de tres supuestos judiciales y de 45 testimonios. En el momento de elaborar el experimento resultaron suficientes para el objetivo que queríamos, que era tener una herramienta para los experimentos siguientes de esta tesis. No obstante, en 2024 realizamos una actualización de la batería (Álvarez et. al, 2025), aumentando el número de casos, así como la muestra utilizada para su estandarización. No la hemos incluido en esta tesis porque los experimentos estaban hechos con la versión 1.0 de ForenPsy, que es lo que describimos en el Capítulo 5 de esta tesis. No obstante, la versión 2.0 (Álvarez et al., 2025) es también de libre acceso y en un futuro el objetivo es ampliar esta batería tanto con mayor variedad de supuestos judiciales como con mayor variedad de testimonios con el fin de poder tener una herramienta de libre acceso que cualquier investigador pueda utilizar en experimentos en el contexto judicial, además de ampliar la muestra.

En los demás experimentos realizados (Experimentos 3, 4 y 5) hemos utilizado por tanto ForenPsy 1.0 para contar con materiales más ecológicos. En estos

experimentos hemos visto cómo una información contraria proveniente de un supuesto algoritmo es capaz de alterar la valoración de los participantes basada en los testimonios. Una sola frase afectó lo suficiente como para que los participantes decidieran cambiar el veredicto dado. Todo esto, habiendo visto una serie de testimonios que estaban estandarizados y orientados en una dirección concreta.

En el Experimento 3 vimos cómo una sola información dada por un supuesto algoritmo, y que resultaba contradictoria con los testimonios que habían visto anteriormente, era suficiente para cambiar la valoración inicial de los participantes. Parece que no otorgaban al algoritmo la capacidad de equivocarse, sino que asumían que la información dada por el algoritmo era más fiable que su propio análisis de los testimonios, llegando como mínimo a equipararse en importancia.

En el Experimento 4 contrabalanceamos el orden de presentación de los testimonios y del algoritmo, ya que en la Serie Experimental 1 habíamos comprobado la importancia de esta variable y queríamos neutralizar su posible efecto. Además, les informamos después de que emitieran su juicio de que los algoritmos también pueden cometer errores. Esto provocó que sus valoraciones cambiaran cuando se les comentaba que podían valorar de nuevo al acusado. Sin embargo, el cambio de juicio no se daba en todos los grupos, lo que indica lo potente que puede ser el efecto del algoritmo en este tipo de contexto. Solamente se tenía en cuenta el posible fallo del algoritmo cuando este valoraba erróneamente al acusado como culpable cuando los testimonios sugerían que era inocente. Parece que el conocimiento de que el algoritmo podía fallar en su veredicto de culpabilidad resultaba relevante cuando habían considerado inocente al acusado. Sin embargo, no resultaba relevante saber que el algoritmo podía equivocarse en el veredicto de inocencia cuando ellos ya habían

considerado al acusado culpable, puesto que no lo declaraban más culpable por saber que el algoritmo podía haberse equivocado.

En el quinto y último experimento, volvíamos a presentar las mismas fases de los dos experimentos anteriores con la particularidad de que ahora se solicitaba un juicio entre cada uno de los bloques de información contradictoria con el objetivo de comprobar si utilizando ForenPsy también podíamos encontrar el efecto de interacción del orden y de la frecuencia del juicio, tal y como encontramos en la Serie Experimental 1. En los resultados obtuvimos datos que indicaban que no había pruebas de que la información del algoritmo fuera más importante que la información de los testimonios, sino que tenían un valor subjetivo similar. Lo que parece que afectaba más era el orden de presentación de las informaciones de inocencia o culpabilidad. Es decir, parece que cada juicio recoge la tendencia (inocencia o culpabilidad) del bloque de información que han visto más recientemente y ese bloque es modificado con la nueva información sin tener en cuenta la información anterior, mostrando un efecto de interacción del orden y de frecuencia del juicio (Catena et al., 2004; Collins & Shanks, 2002; Matute et al., 2002; Vadillo et al., 2004).

Por otro lado, volvemos a encontrar también en el Experimento 5 que resulta más relevante saber que la información proporcionada por el algoritmo es errónea cuando este indicaba un veredicto de culpabilidad que cuando indicaba un veredicto de inocencia, tal y como ya vimos en el Experimento 4.

Con todos los resultados obtenidos podemos concluir que los algoritmos podrán tener una gran influencia en los juicios a la hora de afectar a la valoración o veredicto de los miembros de un jurado, sobre todo si estas personas no tienen conocimiento o no se les proporciona información sobre que los algoritmos están sujetos a numerosos sesgos y errores. El presentar información sobre los posibles

errores sí consiguió paliar, al menos parcialmente, los errores en la decisión de los participantes en los Experimentos 4 y 5.

Nuestros resultados concuerdan con parte de la literatura existente sobre psicología judicial, aunque presentan algunas aportaciones que amplían la comprensión de la influencia de los algoritmos en el contexto judicial. Una de las principales aportaciones en relación con los estudios previos se refiere a la interacción de las personas con los algoritmos. En estudios previos ya mencionados, como los de Pennington (1982), Wells y colaboradores (1985), Furham (1986) y Costabile y Klein (2005), se observó que los efectos de primacía y recencia eran atribuibles a la información humana (testimonios, pruebas), sin que se incluyera la influencia de herramientas tecnológicas como los algoritmos. Otro experimento donde también encontraron efectos del orden de presentación fueron los de Enescu y Kuhn (2012) donde encontraron efecto de recencia entre jueces alemanes, Sin embargo, en nuestros experimentos, observamos que la inclusión de información contradictoria proveniente de un supuesto algoritmo tenía un impacto tan potente como el de los testimonios humanos.

Esto plantea un importante avance respecto a la literatura. Nuestros resultados indican que los participantes, al recibir información del algoritmo, asumían su precisión y objetividad, incluso cuando esta información contradice los testimonios previos. Esto sugiere que el efecto de autoridad algorítmica es un factor crucial a considerar en el contexto judicial, ya que los individuos pueden sobrevalorar la recomendación de un algoritmo en comparación con su propio análisis de los testimonios.

Otro hallazgo relevante es la forma en que los participantes modificaron sus juicios después de ser informados de los posibles errores del algoritmo. A pesar de que los participantes fueron advertidos de que los algoritmos podrían cometer errores,

la influencia de estos seguía siendo significativa, especialmente cuando el algoritmo indicaba culpabilidad en casos en los que los testimonios apuntaban a la inocencia.

Los resultados de nuestros experimentos tienen algunas implicaciones para la psicología judicial, ya que sugieren que los algoritmos podrían desempeñar un papel significativo en el proceso de toma de decisiones judiciales, pero también presentan riesgos que deben ser cuidadosamente gestionados. En primer lugar, nuestros hallazgos subrayan la necesidad de alfabetización algorítmica dentro de los sistemas judiciales. Es esencial que los miembros del jurado y los jueces comprendan no solo el funcionamiento básico de los algoritmos, sino también sus limitaciones y posibles sesgos (véase también Ponce, 2024). De lo contrario, podrían estar inclinados a sobrevalorar las conclusiones de un algoritmo, incluso cuando estas no son precisas.

Además, la forma en que los algoritmos son presentados en el contexto judicial también tiene implicaciones importantes. Si los algoritmos son introducidos en el juicio de manera similar a la evidencia humana, podrían generar el mismo tipo de influencia persuasiva sobre los jurados, afectando su imparcialidad y la forma en que interpretan los testimonios presentados.

Una de las principales conclusiones de nuestros experimentos es que la autoridad algorítmica no es un fenómeno trivial. Los humanos tienen una tendencia natural a confiar en las tecnologías percibidas como objetivas y precisas, lo que podría poner en peligro la imparcialidad de los procesos judiciales si no se gestionan adecuadamente. Los algoritmos no son infalibles y pueden verse afectados por sesgos en su programación o en los datos con los que se entrenan. Por lo tanto, la integración de algoritmos en el proceso judicial debe ser acompañada de protocolos rigurosos de validación, supervisión humana y transparencia en la toma de decisiones. Por suerte, en Europa ya existe una legislación sobre el uso de algoritmos de Inteligencia Artificial

que trata de regular su uso en todos los campos de actuación, incluido el judicial (Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, 2024).

Finalmente, es fundamental que futuros estudios en psicología judicial exploren más a fondo cómo los algoritmos pueden influir en los procesos de toma de decisiones y cómo los sistemas judiciales pueden diseñar medidas que ayuden a mitigar los riesgos asociados a la dependencia excesiva de la inteligencia artificial. La inclusión de algoritmos debe ser acompañada de un proceso de educación y formación tanto para jueces como para jurados y para la sociedad en general, a fin de asegurar que la toma de decisiones se mantenga lo más justa e imparcial posible, minimizando los efectos de sesgos tecnológicos.

El entrenamiento adecuado de todos aquellos que forman parte del sistema judicial sobre el uso de algoritmos es vital para evitar que los sesgos inherentes a los sistemas algorítmicos se infiltren en las decisiones judiciales. Además, la alfabetización algorítmica puede ayudarles a cuestionar los resultados que los algoritmos proporcionan, asegurando que las decisiones no se basen en conclusiones automáticas sin un análisis crítico. Sin este tipo de formación, los trabajadores y personas que forman parte del ámbito judicial pueden caer en la tentación de tratar a los algoritmos como infalibles, obviando la necesidad de un juicio humano fundamentado.

Existen varias áreas que requieren investigación psicológica adicional para comprender mejor el impacto de los algoritmos en el ámbito judicial. En primer lugar, sería valioso explorar cómo diferentes maneras de presentar los algoritmos y la información sobre sus posibles errores y aciertos afectan las decisiones judiciales. Además, la investigación sobre cómo mejorar la alfabetización algorítmica de la sociedad, el conocimiento de sus ventajas, pero también de sus errores podría ser

crucial para mitigar los riesgos asociados con la toma de decisiones basadas en inteligencia artificial.

Finalmente, el impacto de los sesgos algorítmicos en la equidad judicial debe ser explorado a fondo desde el punto de vista psicológico. La regulación y auditoría de los algoritmos utilizados en el sistema judicial será esencial para garantizar que no contribuyan a perpetuar las desigualdades y que los sistemas sean verdaderamente imparciales y justos. Siendo conscientes de que queda mucho trabajo en este campo y en el del estudio de la influencia psicológica de los algoritmos de IAs y cómo afectan a la toma de decisiones de los humanos. Confiamos en que este trabajo sirva para contribuir al incremento del conocimiento y a la evidencia disponible sobre el poder que se otorga a los algoritmos de IAs en el ámbito de las decisiones judiciales, para poder mejorarlas y contribuir a un mundo más justo en beneficio de toda la sociedad.

Parte IV. Referencias Bibliográficas

Referencias Bibliográficas

Agudo, U. (2021). *La influencia de los algoritmos en las decisiones y juicios humanos. Experimento en contextos de política, citas y arte*. [Tesis doctoral, Universidad de Deusto]. https://www.educacion.gob.es/teseo/mostrarRef.do?ref=2133939

Agudo, U., Arrese, M., Liberal, K. G., & Matute, H. (2022). Assessing emotion and sensitivity of AI artwork. *Frontiers in Psychology*, *13*, 879088. https://doi.org/10.3389/fpsyg.2022.879088

Agudo, U., Liberal, K. G., Arrese, M., & Matute, H. (2024). The impact of AI errors in a human-in-the-loop process. *Cognitive Research: Principles and Implications*, *9*(1), 1. https://doi.org/10.1186/s41235-023-00529-3

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, *2*(e93). https://doi.org/10.7717/peerj-cs.93

Aleven, V. (2003). Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. *Artificial Intelligence*, *150*(1–2), 183–237. https://doi.org/10.1016/s0004-3702(03)00105-x

Álvarez, M., Martínez, N., Agudo, U., & Matute, H. (2025). ForenPsy: un Banco Estandarizado de Testimonios Ficticios de Testigos para la Investigación en Psicología Experimental y Judicial. *Anuario de Psicología Jurídica*, *35*(1), 113–119. https://doi.org/10.5093/apj2025a9

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). 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. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Ashley, K.D. (2017). *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge.

Badham, J. (Director) (1983). *WarGames* [Juegos de guerra] [Película]. Metro-Goldwyn-Mayer

Badham, J. (Director) (1986). *Short Circuit* [Cortocircuito] [Película]. TriStar Pictures, Producers Sales Organization, Turman-Foster Company.

Barona, S. (2021). *Algoritmización del derecho y de la justicia. De la Inteligencia Artificial a la Smart Justice.* Tirant Lo Blanch.

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

Bigas, N. (2019, October, 11). Llega el juez artificial: imparcial, eficiente y rápido. *Universitat Oberta de Catalunya*. https://www.uoc.edu/portal/es/news/actualitat/2019/260-juez-artificial-jornada.html

Birhane, A. (2019, July 18). The algorithmic colonization of Africa. *Real Life*. https://reallifemag.com/the-algorithmic-colonization-of-africa/

Blanco, F., Gómez-Fortes, B., & Matute, H. (2018). Causal illusions in the service of political attitudes in Spain and the United Kingdom. *Frontiers in Psychology*, *9*, 1033. https://doi.org/10.3389/fpsyg.2018.01033

Botsman, R. (2017, October 21). Big data meets Big Brother as China moves to rate its citizens. *WIRED*. https://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion

Bovenkamp, S. E. van de. (2017). *Algorithmic imaginary and the case of Spotify* [Doctoral dissertation, Utrecht University]. http://dspace.library.uu.nl/handle/1874/353655

Brodsky, S. L., Griffin, M. P., & Cramer, R. J. (2010). The Witness Credibility Scale: an outcome measure for expert witness research. *Behavioral Sciences & the Law*, *28*(6), 892–907. https://doi.org/10.1002/bsl.917

Buolamwini, J. (2016). *How I'm fighting bias in algorithms* [Video]. TED Conferences. https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), https://doi.org/10.1177/2053951715622512

Cameron, J. (Director) (1984). *The Terminator* [Terminator] [Película]. Hemdale Film Corporation.

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://repositori.justicia.gencat.cat/handle/20.500.14226/271#page=1

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *JMR, Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Catena, A., Perales, J. C., & Maldonado, A. (2004). Judgment frequency effects in generative and preventive causal learning. *Psicológica: Revista de Metodología y Psicología Experimental*, 25(1), 67-85.

Catena, A., Maldonado, A., & Cándido, A. (1998). The effect of frequency of judgement and the type of trials on covariation learning. *Journal of Experimental Psychology. Human Perception and Performance*, *24*(2), 481–495. https://doi.org/10.1037/0096-1523.24.2.481

Chabert, J. L. (Ed.) (1999). *A History of Algorithms: From the Pebble to the Microchip*. Berlin: Springer.

Challen, R., Denny, J., Pitt, M., & Gompels, L. (2019). Artificial intelligence, bias and clinical safety. *British Medical Journal Quality & Safety*, *0*, 1–7. https://doi.org/10.1136/bmjqs-2018-008370

Chan, T.F. (2018, July 16). A Chinese university suspended a student's enrolment because of his dad's bad social credit score. *Insider*. https://www.businessinsider.com/china-social-credit-affects-childs-university-enrolment-2018-7

Chen, C. (2020, December 18). Only seven of Stanford's first 5,000 vaccines were designated for medical residents. *ProPublica*. https://www.propublica.org/article/only-seven-of-stanfords-first-5-000-vaccines-were-designated-for-medical-residents

Chen, L., Mislove, A., & Wilson, C. (2016). An empirical analysis of algorithmic pricing on Amazon marketplace. *Proceedings of the 25th International Conference on World Wide Web*, 1339–1349. https://doi.org/10.1145/2872427.2883089

Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities & Social Sciences Communications*, *10*(1). https://doi.org/10.1057/s41599-023-02079-x

Cheney, C. (2016, September 8). How alternative credit scoring is transforming lending in the developing world. *Devex*. https://www.devex.com/news/how-alternativecredit-scoring-is-transforming-lending-in-the-developing-world-88487

Chorley, A., & Bench-Capon, T. (2005). AGATHA: Using heuristic search to automate the construction of case law theories. *Artificial Intelligence and Law*, *13*(1), 9–51. https://doi.org/10.1007/s10506-006-9004-2

Collins, D. J., & Shanks, D. R. (2002). Momentary and integrative response strategies in causal judgment. *Memory & Cognition*, *30*(7), 1138–1147. https://doi.org/10.3758/bf03194331

Costabile, K. A., & Klein, S. B. (2005). Finishing strong: Recency effects in juror judgments. *Basic and Applied Social Psychology*, *27*(1), 47–58. https://doi.org/10.1207/s15324834basp2701_5

Cummings, M. (2004). Automation bias in intelligent time critical decision support systems. *American Institute of Aeronautics and Astronautics, 1st Intelligent Systems Technical Conference, 2*, 557–562. https://doi.org/doi:10.2514/6.2004-6313

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(17), 6889–6892. https://doi.org/10.1073/pnas.1018033108

Del Castillo, C. (2022, March 9). Las víctimas denuncian fallos en VioGén, el algoritmo contra la violencia de género. *Eldiario.es* https://www.eldiario.es/tecnologia/victimas-denuncian-fallos-viogen-algoritmo-violencia-genero_1_8815201.html

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital journalism*, *5*(7), 809–828. https://doi.org/10.1080/21670811.2016.1208053

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management*, *48*, 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Duportail, J. (2019). *El algoritmo del amor: Un viaje a las entrañas de Tinder*. Contra.

Eiband, M., Völkel, S., Buschek, D., Cook, S. & Hussmann, H. (2019). When people and algorithms meet: user-reported problems in intelligent everyday applications. *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces,* 96-106. https://dl.acm.org/doi/10.1145/3301275.3302262

Enescu, R., & Kuhn, A. (2012). Serial Effects of Evidence on Legal Decision-Making. *European Journal of Psychology Applied to Legal Context*, *4*(2), 99–118.

Engel, C., Timme, S., & Glöckner, A. (2020). Coherence-based reasoning and order effects in legal judgments. *Psychology, Public Policy, and Law: An Official Law Review of the University of Arizona College of Law and the University of Miami School of Law*, *26*(3), 333–352. https://doi.org/10.1037/law0000257

Essinger, J. (2015). *Ada's Algorithm: how Lord Byron's daughter Ada Lovelace launched the Digital Age*. Mellville House.

Fernández, J. (1959). Medidas sencillas de lecturabilidad. *Consigna, 214*, 29-32.

Furnham, A. (1986). The robustness of the recency effect: Studies using legal evidence. *The Journal of General Psychology*, *113*(4), 351–357. https://doi.org/10.1080/00221309.1986.9711045

Hernandez, N. B., Luque, C. E. N., Segura, C. M. L., de Jesus Real Lopez, M., Hungria, J. A. C., & Ricardo, J. E. (2019). La toma de decisiones en la informática jurídica basado en el uso de los sistemas expertos. *Investigación Operacional, 40*(1), 131-139. https://link.gale.com/apps/doc/A570819561/AONE?u=anon~1404fece&sid=googleScholar&xid=f4e46432

Hill, K. (2020, June 24). *Wrongfully accused by an algorithm*. The Seattle Times. https://www.seattletimes.com/business/technology/wrongfully-accused-by-an-algorithm/

Instituto de Ciencias Forenses y de la Seguridad (2018). *La valoración policial del riesgo de violencia contra la mujer pareja en España – Sistema VioGén*. (Publicación: NIPO: 126-18-088-7). Ministerio del Interior. Gobierno de España. http://www.interior.gob.es/documents/642012/8791743/Libro+Violencia+de+G%C3%A9nero/19523de8-df2b-45f8-80c0-59e3614a9bef

Jiménez, J. (2016, April 30). La historia de Claude Shannon: el hombre que creó la información. *Xataka*. https://www.xataka.com/historia-tecnologica/un-pequeno-homenaje-a-claude-shannon-el-hombre-que-creo-la-informacion

Kahnenan, D. (2011). *Pensar rápido, pensar despacio*. Debate.

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016, October). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*. https://hbr.org/2016/10/noise

Kassin, S. M., & Sommers, S. R. (1997). Inadmissible testimony, instructions to disregard, and the jury: Substantive versus procedural considerations. *Personality & Social Psychology Bulletin*, *23*(10), 1046–1054. https://doi.org/10.1177/01461672972310005

Katz, D. M., Bommarito, M. J., 2nd, & Blackman, J. (2017). A general approach for predicting the behavior of the Supreme Court of the United States. *PloS One*, *12*(4), e0174698. https://doi.org/10.1371/journal.pone.0174698

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication and Society*, *20*(1), 14–29. https://doi.org/10.1080/1369118x.2016.1154087

Knight, W. (2017, April 11). The dark secret at the heart of AI. *MIT Technology Review*. https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/

Lapowsky, S. (2018, May 22). How the LAPD uses data to predict crime. *WIRED*. https://www.wired.com/story/los-angeles-police-department-predictive-policing/

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). How we analyzed the COMPAS recidivism algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lassiter, G. D., Geers, A. L., Handley, I. M., Weiland, P. E., & Munhall, P. J. (2002). Videotaped interrogations and confessions: A simple change in camera perspective alters verdicts in simulated trials. *The Journal of Applied Psychology*, *87*(5), 867–874. https://doi.org/10.1037/0021-9010.87.5.867

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 205395171875668. https://doi.org/10.1177/2053951718756684

Ley Orgánica 5/1995, de 22 de mayo, del Tribunal del Jurado (1995), *Boletín Oficial del Estado, 122*, de 23 de mayo de 1995, 15001–15021. https://www.boe.es/eli/es/lo/1995/05/22/5

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

López, F. J., Shanks, D. R., Almaraz, J., & Fernández, P. (1998). Effects of trial order on contingency judgments: A comparison of associative and probabilistic contrast accounts. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *24*(3), 672–694. https://doi.org/10.1037/0278-7393.24.3.672

Marmolejo-Ramos, F., Marrone, R., Korolkiewicz, M., Gabriel, F., Siemens, G., Joksimovic, S., Yamada, Y., Mori, Y., Rahwan, T., Sahakyan, M., Sonna, B., Meirmanov, A., Bolatov, A., Som, B., Ndukaihe, I., Arinze, N. C., Kundrát, J., Skanderová, L., Ngo, V.-G., … Tejada, J. (2024). Factors influencing trust in algorithmic decision-making: an indirect scenario-based experiment. *Frontiers in Artificial Intelligence*, *7*, 1465605. https://doi.org/10.3389/frai.2024.1465605

Martínez, G.C. (2012). La inteligencia artificial y su aplicación al campo del Derecho. *Alegatos, 82*, 827-846.

Matute, H., Blanco, F., Moreno-Fernández, M.M. (2022). Causality bias. In R. F. Pohl (ed). *Cognitive Illusions: Intriguing Phenomena in Thinking, Judgment, and Memory*. (3rd ed.). London: Routledge. https://doi.org/10.4324/9781003154730

Matute, H., Vegas, S., & De Marez, P.-J. (2002). Flexible use of recent information in causal and predictive judgments. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *28*(4), 714–725. https://doi.org/10.1037/0278-7393.28.4.714

Merino, M. (2018, November 30). Hacer públicos o no los avances en inteligencia artificial: los cientificos no se ponen de acuerdo. *Xataka*. https://www.xataka.com/robotica-e-ia/hacer-publicos-no-avances-inteligenciaartificial-cientificos-no-se-ponen-acuerdo

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525

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

Ministry of Justice. (2013). *Offender assessment system (OASys)*. Data.Gov.Uk. https://www.data.gov.uk/dataset/911acd3c-495f-48ca-88b6-024210868b06/offender-assessment-system-oasys

Mosier, K. L., Skitka, L. J., Burdick, M. D., & Heers, S. T. (1996). Automation bias, accountability, and verification behaviors. *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting. Human Factors and Ergonomics Society. Annual Meeting*, *40*(4), 204–208. https://doi.org/10.1177/154193129604000413

Navas, S. (2017). *Inteligencia Artificial. Tecnología. Derecho*. Tirant Lo Blanch.

Neyland, D., & Möllers, N. (2017). Algorithmic IF … THEN rules and the conditions and consequences of power. *Information, Communication and Society*, *20*(1), 45–62. https://doi.org/10.1080/1369118x.2016.1156141

Ng, D. W., Lee, J. C., & Lovibond, P. F. (2024). Unidirectional rating scales overestimate the illusory causation phenomenon. *Quarterly Journal of Experimental Psychology (2006)*, *77*(3), 551–562. https://doi.org/10.1177/17470218231175003

Niiler, E. (2019, March 25). 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 (New York, N.Y.)*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409. https://doi.org/10.1002/bdm.637

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication: JCMC*, *12*(3), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

Paul, S. (2020, May 28). Will Artificial Intelligence replace Judging? *Bar and Bench*. https://www.barandbench.com/columns/is-artificial-intelligence-replacing-judging

Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial). Diario Oficial de la Unión Europea, L 123, 1–45. https://www.boe.es/buscar/doc.php?id=DOUE-L-2024-81079

Pennington, D. C. (1982). Witnesses and their testimony: Effects of ordering on juror verdicts. *Journal of Applied Social Psychology*, *12*(4), 318–333. https://doi.org/10.1111/j.1559-1816.1982.tb00868.x

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *Journal of Personality and Social Psychology*, *62*(2), 189–206. https://doi.org/10.1037/0022-3514.62.2.189

Ponce Solé, J. (2024). *El Reglamento de Inteligencia Artificial de la Unión Europea de 2024, el derecho a una buena administración digital y su control judicial en España*. Marcial Pons.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & Society*, *36*(1), 59–77. https://doi.org/10.1007/s00146-020-00992-2

Rogers, J. (2017). How “random” is Ryanair’s seating allocation? *Significance*, *14*(5), 12–13. https://doi.org/10.1111/j.1740-9713.2017.01069.x

Ruva, C. L., & Guenther, C. C. (2015). From the shadows into the light: How pretrial publicity and deliberation affect mock jurors’ decisions, impressions, and memory. *Law and Human Behavior*, *39*(3), 294–310. https://doi.org/10.1037/lhb0000117

Scott, R. (Director) (1982). *Blade Runner* [Película]. Warner Bros., Ladd Company, Shaw Brothers.

Sealy, P. (1975). The Jury: Decision making in a small group. En H. Brown & R. Stevens (Eds), *Social behavior and experience*. Hodder & Stoughton.

Sealy, A. P., & Cornish, W. R. (1973). Jurors and their verdicts. *The Modern Law Review*, *36*(5), 496–508. https://doi.org/10.1111/j.1468-2230.1973.tb01381.x

Sergot, M. J., Sadri, F., Kowalski, R. A., Kriwaczek, F., Hammond, P., & Cory, H. T. (1986). The British Nationality Act as a logic program. *Communications of the ACM*, *29*(5), 370–386. https://doi.org/10.1145/5689.5920

Soler, C. (2013). *RisCanvi. Protocolo de evaluación y gestión del riesgo de violencia con población penitenciaria* [PowerPoint slides]. Slideplayer. https://slideplayer.es/slide/7242758/

Sommers, S. R., & Ellsworth, P. C. (2001). White juror bias: An investigation of prejudice against Black defendants in the American courtroom. *Psychology, Public Policy, and Law: An Official Law Review of the University of Arizona College of Law and the University of Miami School of Law*, *7*(1), 201–229. https://doi.org/10.1037/1076-8971.7.1.201

Stranieri, A. & Zeleznikow, J. (1998). Split Up: The Use of an Argument Based Knowledge Representation to Meet Expectations of Different Users for Discretionary Decision Making. *Association fot the Advancement of Artificial Intelligence*, 1146–1151. https://aaai.org/papers/020-iaai98-020-iaai98/

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication: JCMC*, *25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Sundar, S. S. (2008). The MAIN model: A heuristic Approach to understanding technology effects on credibility. *Digital Media, Youth, and Credibility*, 73–100.

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3290605.3300768

Susskind, R. E. (1986). Expert Systems in Law: A Jurisprudential Approach to Artificial Intelligence and Legal Reasoning. *The Modern Law Review*, *49*(2), 168–194.
Vadillo, M. A., Vegas, S., & Matute, H. (2004). Frequency of judgment as a context-like determinant of predictive judgments. *Memory & Cognition*, *32*(7), 1065–1075. https://doi.org/10.3758/bf03196882

Vicente, L. (2024) Herencia del sesgo: influencia de los sesgos de la Inteligencia Artificial en las decisiones humanas [Tesis Doctoral, Universidad de Deusto] https://repositorio.deusto.es/items/e5432a45-4c86-4e37-bee0-787cc3bcbec2

Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, *13*(1), 15737. https://doi.org/10.1038/s41598-023-42384-8

Vincent, J. (2016, 24 marzo). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Ware, L. J., Lassiter, G. D., Patterson, S. M., & Ransom, M. R. (2008). Camera perspective bias in videotaped confessions: evidence that visual attention is a mediator. *Journal of Experimental Psychology: Applied*, *14*(2), 192–200. https://doi.org/10.1037/1076-898X.14.2.192

Weinshall-Margel, K., & Shapard, J. (2011). Overlooked factors in the analysis of parole decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(42), https://doi.org/10.1073/pnas.1110910108

Wei, J. (2019, January 24). China uses AI assistive tech on court trial for first time. *China-Daily*. https://www.chinadaily.com.cn/a/201901/24/WS5c4959f9a3106c65c34e64ea.html#:~:text=For%20the%20first%20time%20in,including%20in%20the%20public%20gallery.

Wells, G. L., Wrightsman, L. S., & Miene, P. K. (1985). The timing of the defense opening statement: Don't wait until the evidence is in. *Journal of Applied Social Psychology*, *15*(8), 758–772. https://doi.org/10.1111/j.1559-1816.1985.tb02272.x

Willson, M. (2017). Algorithms (and the) everyday. *Information, Communication and Society*, *20*(1), 137–150. https://doi.org/10.1080/1369118x.2016.1200645

Wilson, W. (1971). Source credibility and order effects. *Psychological Reports*, *29*(3), 1303–1312. https://doi.org/10.2466/pr0.1971.29.3f.1303

Xu, C., & Doshi, T. (2019, December 11). Fairness indicators: Scalable infrastructure for fair ML systems. *Google AI Blog*. https://ai.googleblog.com/2019/12/fairness-indicators-scalable.html

Apéndices

Apéndice A. Casos y testimonios de ForenPsy

SUPUESTO 1: HOMICIDIO

LEE LA SIGUIENTE HISTORIA

Se acusa a M.L.P. de haber acabado con la vida de A.R.V. el día 22 de marzo de 2018 entre las 22:00 y las 23:00. A.R.V. fue encontrado sin vida en la C/ Mayor con múltiples heridas de arma blanca (cuchillo). La víctima no llevaba cartera y le faltaba una cadena de oro con una medalla de una virgen grabada con sus iniciales.

Habiendo leído esto, ahora nos gustaría que valoraras si cada uno de los siguientes testimonios inclinarían la balanza en las direcciones de inocente o culpable (marcando la casilla) y la importancia que tendría cada uno de estos testimonios en tu decisión (usando la escala).

Declaraciones de los testigos

Testigo 1: "Vi a M.L.P. discutir con A.R.V. alrededor de las 21:30 en una zona próxima a donde fue encontrado el cadáver."

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 2: "M.L.P. estuvo en una casa de empeño queriendo vender una cadena de oro."

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 3: "Vi a M.L.P. tomando algo en un bar a la hora en la que se cometió el crimen. Recuerdo la hora porque estaba viendo un partido de futbol."

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 4: "M.L.P. me llamó para preguntarme si conocía algún sitio donde vender cosillas sin que preguntaran."

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 5: “Vi a M.L.P. pasada la hora del crimen tirando algo a un contenedor de basura.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 6: “Vi a M.L.P. limpiando unas manchas de algo que parecía sangre en una fuente de la plaza que está justo detrás de donde apareció el cadáver.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 7: “Unos días después del crimen, le vi puesta a M.L.P. una cadena similar a la sustraída y que nunca antes le había visto al sospechoso.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 8: “Vi a M.L.P. en un coche junto a A.R.V. sobre las 20:30 de ese día.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 9: “M.L.P. se ha metido en peleas y problemas en estos últimos meses por dinero.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 10: “Un poco pasada la hora del crimen me crucé con M.L.P. en un bar en la otra punta de la ciudad.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 11: “Vi a A.R.V. junto a otra persona que no era M.L.P. poco antes de las 22:00.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 12: “M.L.P. me llamó por teléfono sobre las 22:30 y estuve 15 minutos hablando con él.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Suficiente

Testigo 13: “A.R.V. me contó unos días antes que debía dinero a una persona que no era M.L.P.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 14: “Me encontré a M.L.P. bastante alterado cerca de la medianoche.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 15: “A.R.V. tenía deudas de juego.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

SUPUESTO 2: AMENAZAS

LEE LA SIGUIENTE HISTORIA

Se acusa a M.A.Z. de haber realizado amenazas de muerte contra N.M.D. el día 15 de febrero de 2019 a las 19:00 mediante correo electrónico a través de una cuenta que no es la conocida de M.A.Z. y cuya IP fue rastreada hasta una cafetería de videojuegos. En el correo le exigía un pago de 3.000 euros dejados en un punto de la ciudad.

Habiendo leído esto, ahora nos gustaría que valoraras si los siguientes testimonios te ayudarían a declararlo inocente o culpable (marcando la casilla) y la importancia que tendría cada uno de estos testimonios en tu decisión (usando la escala).

Declaraciones de los testigos

Testigo 1: "Escuché a M.A.Z. diciendo que dentro de poco podría comprarse la moto de 3.000 euros que siempre había querido."

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 2: "El otro día M.A.Z. estuvo diciendo que necesitaba dinero de forma urgente."

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 3: "Vi a M.A.Z. en una cafetería de videojuegos con un portátil y escribiendo un correo electrónico."

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 4: "Escuché a M.A.Z. el otro día preguntando a un amigo cómo hacerse una cuenta de correo irrastreable."

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 5: “Vi a M.A.Z. jugando en línea todo el día a un videojuego al que jugamos juntos e incluso hablaba por el chat.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 6: “A lo largo de la tarde vi varias veces a M.A.Z. asomado a la ventana de su habitación.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 7: “Hace poco vi discutir a N.M.D. y a M.A.Z., y este último le dijo que ya se arrepentiría.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 8: “M.A.Z. me preguntó por cómo solicitar un préstamo a un banco para una moto.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 9: “M.A.Z. contó hace unos días que una tía suya le había dejado en herencia algo de dinero.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 10: “N.M.D. siempre ha dicho que, debido a su trabajo, tenía muchos enemigos.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 11: “N.M.D. ha tenido varios juicios por amenazas de otras personas.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 12: “M.A.Z. siempre ha tenido problemas para controlar la ira.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 13: “Mientras jugábamos, M.A.Z. estuvo ausente durante un rato en el que no participó.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 14: “N.M.D. y M.A.Z. ya habían tenido discusiones en otras ocasiones.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 15: “En la cafetería donde se rastreó la IP había otras quince personas, además de N.M.D.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

SUPUESTO 3: ALLANAMIENTO

LEE LA SIGUIENTE HISTORIA

J.S.S. está acusado de entrar sin autorización en una tienda de ropa, fuera de su horario de apertura al público, y sin haberse llevado nada. La puerta trasera de la tienda fue forzada con una palanca.

Habiendo leído esto, ahora nos gustaría que valoraras si los siguientes testimonios te ayudarían a declararlo inocente o culpable (marcando la casilla) y la importancia que tendría cada uno de estos testimonios en tu decisión (usando la escala).

Declaraciones de los testigos

Testigo 1: "Vi a J.S.S., poco antes de que cerrase la tienda, sentado en un coche frente a la tienda."

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 2: "Vi a J.S.S. guardando una palanca en su coche por la tarde."

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 3: "J.S.S. estuvo toda la noche en la discoteca donde yo trabajo."

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 4: "Durante toda la semana, vi a J.S.S. varias veces en la tienda."

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯ ◯ ◯ ◯ ◯

Muy poco Poco Algo Mucho Muchísimo

Testigo 5: “Vi a J.S.S. corriendo en el parque durante la tarde.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 6: “Vi a J.S.S. dando vueltas por la zona de la tienda a altas horas de la noche.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 7: “Vi a J.S.S. salir corriendo de la zona de la tienda.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 8: “Algunas veces, los empleados se dejan las llaves en casa y saben cómo abrir la puerta trasera.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 9: “J.S.S. estuvo yendo varias veces a la tienda porque se equivocó de talla con la ropa que compró.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 10: “No es la primera vez que fuerzan la tienda y las otras veces fue siempre la misma persona. Y no era J.S.S.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 11: “Hay un empleado que no tiene una coartada para el tiempo en el que la tienda fue forzada.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 12: “J.S.S. nunca ha sido acusado de ningún delito.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 13: “Uno de los empleados ha tenido varias discusiones y peleas con J.S.S. en los últimos meses por temas personales.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 14: “El coche de J.S.S. estuvo aparcado toda la noche cerca del centro comercial.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo

Testigo 15: “J.S.S. conoce a la persona que forzó la tienda las anteriores ocasiones.”

¿En qué dirección inclinaría la balanza este testimonio?

INOCENTE ◯ CULPABLE ◯

¿Cómo de importante sería este testimonio para determinar tu respuesta?

◯	◯	◯	◯	◯
Muy poco	Poco	Algo	Mucho	Muchísimo