Herencia del sesgo

INFLUENCIA DE LOS SESGOS DE LA INTELIGENCIA ARTIFICIAL EN LAS DECISIONES HUMANAS

Tesis Doctoral realizada por Lucía Vicente Holgado

Dirigida por la Dra. Helena Matute

PROGRAMA DE DOCTORADO EN PSICOLOGÍA

UNIVERSIDAD DE DEUSTO

Universidad de Deusto
Deustuko Unibertsitatea
University of Deusto

Herencia del sesgo

INFLUENCIA DE LOS SESGOS DE LA INTELIGENCIA ARTIFICIAL EN LAS DECISIONES HUMANAS

Tesis Doctoral realizada por Lucía Vicente Holgado

Dirigida por la Dra. Helena Matute

PROGRAMA DE DOCTORADO EN PSICOLOGÍA

CURSO 2023-2024

UNIVERSIDAD DE DEUSTO

La doctoranda

Firmado por VICENTE HOLGADO LUCIA - 70903852G el día 14/05/2024 con un certificado emitido por AC FNMT Usuarios

La directora

Helena Matute

Firmado digitalmente por Helena Matute
Fecha: 2024.05.14 11:59:59 +02'00'

Bilbao, 11 de marzo de 2024

Financiación

Esta tesis se ha realizado en el marco de los siguientes proyectos concedidos a la Dra. Helena Matute: PID2021-126320NB-I00 de la Agencia Estatal de Investigación del Gobierno de España e IT1696-22 del Departamento de Educación, Universidades e Investigación del Gobierno Vasco.

Lucía Vicente recibió una Ayuda para la Formación de Personal Investigador en la convocatoria Ayudas FPI 2020 de la Universidad de Deusto, que obtuvo renovación por tres años consecutivos.

Agradecimientos

Nunca me he sentido sola durante el desarrollo de esta Tesis Doctoral y son muchas las personas a las que debo agradecérselo. En primer lugar, gracias a mi directora, Helena Matute, por todo lo que me ha enseñado a lo largo de cuatro años, ha sido un inmenso privilegio trabajar con ella. Cuando Helena me planteó la posibilidad de solicitar una ayuda predoctoral e integrarme en su grupo de investigación no dudé, sabía que debía perseguir esa oportunidad. Para ello incluso crucé el país de sur a norte en una mudanza atropellada. Desde que llegué a Bilbao, cada día me he sentido afortunada y he sentido un enorme agradecimiento hacia Helena. También, por supuesto, agradezco el grandísimo apoyo, los muchos consejos y el inestimable consuelo de mis compañeros y compañeras del grupo de investigación Labpsico. No puedo imaginar unas mejores personas con las que transitar los desasosiegos del doctorado que Marcos, Mario, Aranzta, Naroa y Ujué. Gracias también a Fernando y Nela, a quienes admiro muchísimo, por ser modelos de los que aprender, incluso en la distancia.

Quiero dedicar también unas palabras de agradecimiento a todas aquellas personas que, desde que empecé a dedicarme a la investigación, en algún momento me han llevado de la mano conduciéndome a lo largo de un camino que culmina con la realización del doctorado. Para mí es importante dar las gracias a Gonzalo de la Casa, Auxi, Juan Carlos, Lucía, Capea, por todo lo que me enseñaron en mi tiempo en el Laboratorio de Neurociencia y Conducta Animal de la Universidad de Sevilla. Gracias a Estela Castilla, Mari Carmen, Sara, Fabiola y todas las compañeras con las que fue un placer trabajar en el Centro de Experimentación y Conducta Animal de la Universidad de Málaga.

Siento mucho cariño por todas estas personas que no solo han contribuido a mi desarrollo en una faceta profesional, si no también personal, y forman parte de quién soy.

Debo agradecer la consecución de este logro también a mi familia, manos de tierra acostumbradas a la labor que me acunaron y protegieron siempre que lo necesité. Gracias a mi madre Pepi y mi padre José María, por todas las veces en que yo dudé de si sería capaz y en que ellos tuvieron la certeza de que podía conseguirlo. Gracias a mis abuelas Dorinda y Abilia, por el coraje. Gracias a mis abuelos José y José María, por la sabiduría serena. Tomando las palabras de la poeta Maribel Andrés Llamero “ellos me levantaron el alma con golpes de azada que aún retumban en el amor áspero y tierno que me puebla”.

Y gracias a Antonio, el mejor compañero de aventuras.

Contenido

Resumen 1

Abstract 2

Declaración Ética 3

Capítulo 1. Introducción 5

1.1. Una Definición de Algoritmo e Inteligencia Artificial 11

Algoritmo 11

¿Qué es la Inteligencia Artificial? 12

Inteligencia Artificial como Disciplina Científica 13

Lo que las Personas Creen que es la Inteligencia Artificial 14

1.2. Inteligencia Artificial Aplicada al Cuidado de la Salud 15

Pensar Rápido: Heurísticos, Sesgos Cognitivos y Ruido 16

Inteligencia Artificial para Tomar Mejores Decisiones 20

Retos de la Implementación de Sistemas de IA 21

1.3. Sesgos en la Inteligencia Artificial 22

Riesgo de Sesgo en la IA Aplicada al Cuidado de la Salud 26

1.4. Influencia de la Inteligencia Artificial en las Decisiones Humanas 27

Capítulo 2. Herencia del Sesgo de la Inteligencia Artificial 33

2.1. Serie Experimental 1 33

Experimento 1: Influencia de una IA Sesgada en las Decisiones Humanas 36

Experimento 2: Herencia del Sesgo de la IA 49

Experimento 3: Efecto de la Experiencia Previa en la Herencia del Sesgo de la IA 64

Experimento 4: Efecto de la Formación Previa en la Herencia del Sesgo de la IA 73

2.2. Discusión de la Serie Experimental 1..... 82

Capítulo 3. Mitigar la Influencia de las Recomendaciones Sesgadas de la Inteligencia Artificial 91

3.1. Serie Experimental 2 91

Experimento 5: Advertencia de Error para Mitigar la Influencia de una IA Sesgada..... 94

Experimento 6: Efecto del Enmarcado de la Información 109

Experimento 7: Efecto de la Advertencia de Error y Coste del Error 125

3.2 Discusión de la Serie Experimental 2 142

Capítulo 4: Discusión General y Conclusiones. 155

GENERAL DISCUSSION AND CONCLUSIONS 155

Referencias 168

Apéndice A 197

Apéndice B 202

Resumen

Las recomendaciones erróneas o sesgadas de la inteligencia artificial pueden ejercer un impacto negativo en las decisiones de las personas, como han puesto de manifiesto un número significativo de estudios. Sin embargo, un aspecto clave de la influencia de la IA en el comportamiento humano aún no ha sido explorado: ¿después de interactuar con un sistema de IA sesgado las personas reproducirán ese sesgo en sus propias decisiones futuras, incluso en un contexto sin IA? En dos series de experimentos exploramos si las personas podrían heredar sesgos de la IA y pusimos a prueba estrategias para mitigar este efecto negativo en las decisiones humanas. En la primera serie experimental (Experimentos 1, 2, 3 y 4) los participantes completaron una tarea que simulaba un diagnóstico médico sin asistencia o con la asistencia de un sistema de IA sesgado. Las recomendaciones sesgadas de la IA influyeron las decisiones de los participantes, pero, el hallazgo más importante de esta serie experimental fue que los participantes reprodujeron en sus respuestas el sesgo de la IA que les había proporcionado asistencia cuando se enfrentaron a realizar la tarea de diagnóstico ya sin esta ayuda. Por lo tanto, los participantes imitaron el sesgo de la IA incluso cuando ya había finalizado su colaboración con este sistema. En la segunda serie experimental pusimos a prueba si alertar a los participantes de forma general sobre la posible presencia de errores y sesgos en la IA (Experimentos 5 y 7) o si informarles de forma específica sobre el porcentaje de recomendaciones erróneas o correctas de la IA (Experimento 6) podría atenuar el impacto negativo de las recomendaciones sesgadas en sus decisiones, previniendo la herencia del sesgo de la IA. Los resultados de esta segunda serie sugieren que enfatizar el riesgo de error en la IA podría reducir la tendencia de las personas a aceptar acríticamente sus recomendaciones erróneas. Las personas podrían heredar sesgos de modelos de IA con los que interactúan, lo cual hace necesario continuar investigando para descubrir estrategias que mitiguen este riesgo.

Abstract

Several studies have shown that biased recommendations from artificial intelligence (AI) systems have the potential to influence human decision-making. However, some key implications of AI's influence on human behaviour remain unexplored: it is crucial to know whether, after interacting with a biased AI system, people will reproduce AI biases in their decisions, even after the interaction with the AI has ended. In two sets of experiments, we explored this possible human inheritance of AI biases and tested strategies that could mitigate such a negative impact of AI on human decisions. In the first experimental series (Experiments 1,2,3 and 4), participants completed a medical-themed classification task with or without the help of a biased AI system. We found that biased recommendations by the AI strongly influenced participants' decisions. However, our most significant finding has been that, when the participants who were assisted by the AI went on to perform the task without any assistance, they made the same errors as the AI had made during the previous phase. Therefore, participants' responses mimicked AI bias even when the AI was no longer making suggestions. The impact of biased AI advice on human behaviour extends beyond the phase in which the AI recommendations are explicitly present. In the second experimental series, we tested whether general information about potential biases and errors in the AI (Experiments 5 and 7), or explicit information about the AI's percentage of erroneous or correct recommendations (Experiment 6), could help mitigate the negative impact of AI bias on participants' responses and prevent the inheritance of bias. Our results showed that a simple intervention that emphasizes the risk of error in AI can help reduce the tendency of volunteers to follow incorrect AI suggestions. We conclude that there is potential for humans to inherit the bias of an artificial intelligence system, and this makes it important to develop strategies that can mitigate such risk.

Declaración Ética

El Comité de Ética de la Universidad de Deusto revisó y aprobó la metodología descrita en esta Tesis Doctoral. Los experimentos fueron realizados siguiendo el protocolo aprobado, teniendo en cuenta la normativa ética de la universidad. Obtuvimos consentimiento informado de todos los voluntarios previo a su participación en los experimentos. Las respuestas de los participantes se recopilaron online de forma anónima y no recogimos información personal o sensible.

Debido a consideraciones éticas, y también para prevenir la influencia de conocimiento previo, creencias o expectativas de los participantes en los resultados de nuestros experimentos, tanto el contexto clínico, la tarea de clasificación, las muestras de tejido de pacientes y el síndrome descrito, así como el sistema de inteligencia artificial, sus recomendaciones y sesgo, utilizados en los experimentos de la presente Tesis Doctoral, fueron todos ficticios.

La mayor parte de los experimentos que hemos realizado como parte del presente trabajo de investigación fueron prerresigtrados en AsPredicted.org. En la sección Método de cada uno de los experimentos compartiremos el enlace del prerregistro correspondiente. De forma similar, comprometidas con la ciencia abierta, los datos que hemos recogido en cada uno de los experimentos de la presente Tesis Doctoral han sido compartidos y pueden descargarse desde Open Science Framework: https://osf.io/2s93n/?view_only=f056b9c533c6432789aa49bb3b69d998

Capítulo 1. Introducción

Los avances tecnológicos logrados en la disciplina de la inteligencia artificial (IA) en las últimas décadas han dado a luz a sistemas artificiales que exhiben una capacidad similar, y en ocasiones incluso superior, a la de la inteligencia humana para procesar datos, detectar patrones, realizar predicciones precisas y aprender. Dada su potencia y capacidad, la inteligencia artificial está siendo aplicada a las tareas más diversas, generando asombro en el público con los resultados obtenidos. Por ejemplo, piezas artísticas generadas por IA resultan indistinguibles de las creadas por autores humanos (Agudo et al., 2022; Köbis & Mossink, 2021), y las tácticas que ejecuta la IA en juegos de estrategia sorprenden a campeones experimentados (Shin et al., 2023). Los algoritmos de inteligencia artificial también han demostrado que pueden alcanzar una precisión superior a la de expertos radiólogos en la detección del cáncer de mama a partir de la evaluación de mamografías (McKinney et al., 2020) o reconocer y categorizar caras con un nivel de acierto similar al de los humanos (Phillips et al., 2018).

En paralelo a estos desarrollos en la disciplina de la inteligencia artificial, cada vez más campos profesionales están incorporando herramientas basadas en IA para asistir a las personas durante los procesos de toma de decisiones. En el ámbito de la justicia, algoritmos de inteligencia artificial evalúan el riesgo de víctimas de violencia de género de volver a ser atacadas por sus agresores, lo cual ayuda a los jueces a determinar qué medidas de protección necesitan (Valdivia et al., 2022). En recursos humanos, la IA ayuda a identificar, entre miles de currículums, el candidato más idóneo para un determinado puesto de trabajo durante procesos de selección de personal (Kupfer et al., 2023; Lacroux & Martin-Lacroux, 2022). En el campo de la medicina, la inteligencia artificial asesora a los profesionales sanitarios en el diagnóstico, la elección de tratamiento o el pronóstico de la evolución de los pacientes, iniciando una verdadera revolución en el cuidado de la salud (Rajpurkar et al., 2022; Topol, 2019; Yu et al., 2018).

De esta colaboración entre inteligencia natural y artificial se espera una reducción de errores en las decisiones humanas en áreas de alta responsabilidad (Kahneman et al., 2021). La inteligencia artificial potencia las habilidades de la mente humana (Akata et al., 2020; Dellermann et al., 2019), a la par que contrarresta algunas de sus debilidades, como los sesgos cognitivos o la vulnerabilidad a la fatiga, que pueden afectar negativamente la toma de decisiones (Kahneman et al., 2021). De forma complementaria, la inteligencia humana compensa las limitaciones de la artificial (Dellermann et al., 2019), y las personas supervisan los resultados de la IA para asegurar que cada decisión es correcta y ética (European Commission, 2021; Suresh et al., 2020).

A día de hoy, la interacción entre humanos e inteligencia artificial es ya frecuente y cotidiana, lo cual nos conduce irrevocablemente a cuestionar qué impacto puede llegar a ejercer la inteligencia artificial sobre el comportamiento humano. Más específicamente, nos preguntamos hasta qué punto la asistencia de una inteligencia artificial, ofrecida habitualmente en forma de recomendaciones explícitas, influye en las decisiones de las personas (Cabitza, Campagner, Angius, et al., 2023; Solans et al., 2022). También es pertinente evaluar si, verdaderamente, las personas son capaces de ejercer una supervisión crítica sobre esas recomendaciones de la IA (Goddard et al., 2012, 2014).

Investigaciones previas ya han aportado evidencia del potencial de los algoritmos de IA para orientar decisiones humanas, incluso en situaciones tan relevantes como decidir a qué candidato votar en unas elecciones presidenciales (Agudo & Matute, 2021; Epstein & Robertson, 2015). Dada la rápida evolución de las herramientas basadas en inteligencia artificial y su creciente expansión en diversos ámbitos de la sociedad es necesario investigar la capacidad de influencia de esta tecnología sobre las personas. El foco de esta investigación no debería circunscribirse solo al momento preciso en el que una persona debe tomar una decisión con el apoyo de una recomendación aportada por un sistema de inteligencia artificial; para obtener una mejor comprensión del fenómeno es necesario

conocer si, una vez finalizada la interacción con la IA, permanecen vestigios de la influencia de este agente en el comportamiento de las personas.

Algunos investigadores ya han comenzado a preguntarse si la interacción continuada con sistemas de IA tiene un impacto perdurable en el comportamiento humano (Kidd & Birhane, 2023; Shin et al., 2023). La inteligencia artificial puede ser un modelo del que aprender, sobre todo si las personas tienden a percibirlo como una fuente experta y competente (Agudo & Matute, 2021; Araujo et al., 2020; Logg et al., 2019). El resultado positivo de la colaboración entre el ser humano y la IA sería la mejora y potenciación mutua. Por ejemplo, las personas pueden llegar a beneficiarse de las estrategias de toma de decisiones novedosas y creativas descubiertas por la IA, aprenderlas y reproducirlas, alcanzando resultados que no habrían podido lograr por sí solas (L'Imperio et al., 2023; Shin et al., 2023). Sin embargo, en algunas situaciones la cooperación entre la inteligencia humana y la artificial puede resultar perjudicial, en vez de beneficiosa, para la toma de decisiones. Frente al optimismo y asombro que suelen acompañar a los nuevos desarrollos en inteligencia artificial, se alzan voces críticas que señalan los riesgos de la presencia de errores y sesgos en dichos modelos (Fry, 2018; O'Neil, 2016). Las recomendaciones de una IA sesgada podrían conducir a decisiones erróneas y carentes de ética, si los humanos siguen dichas recomendaciones de forma acrítica. Además, en este escenario de conformidad con la IA se ha expresado temor por la posibilidad de que los sesgos de la inteligencia artificial puedan acabar siendo transmitidos a los humanos tras una exposición repetida a dichos sesgos a lo largo del tiempo (Kidd & Birhane, 2023). Pese a sus graves implicaciones, esta posibilidad aún no había sido investigada.

Una vez heredados por los humanos, los sesgos podrían ser difíciles de erradicar, incluso aunque finalice la colaboración entre las personas y la inteligencia artificial sesgada. Para ejemplificarlo, imaginemos un sistema de evaluación del riesgo de mortalidad en pacientes de UCI basado en IA que asigna dos puntuaciones diferentes a dos personas que, en realidad, tienen el mismo riesgo de morir. Aunque las dos personas tienen similar historial clínico, las diferentes

características étnicas de los dos pacientes resultan determinantes para la evaluación de la IA (Chen et al., 2018; Chen et al., 2021). En este caso, la etnia no es relevante para la estimación del riesgo, pero sin embargo la IA ha identificado erróneamente la etnia como una variable predictora. A consecuencia de este sesgo en el modelo, la persona que recibe la menor puntuación no tendrá acceso a la atención clínica especializada que necesita. El sanitario que interactúa con esta IA y supervisa sus resultados podría aprender el mismo patrón, de forma que podría acabar cambiando su propia estrategia de toma de decisiones para adaptarse a las evaluaciones de la IA, que considera precisas. Finalmente, aunque se descubra el sesgo y dicha herramienta de IA se retire, el profesional sanitario puede seguir replicando el sesgo en sus propias decisiones, dejándose influir por las características étnicas de una persona a la hora de determinar su riesgo frente a la enfermedad.

El fenómeno de la potencial herencia del sesgo de la IA y su propagación a las decisiones humanas no había sido explorado empíricamente. Por ello, el primer objetivo de la presente Tesis Doctoral fue examinar si las recomendaciones sesgadas de un sistema de inteligencia artificial podían influir las decisiones humanas, de forma que estas decisiones imitasen el sesgo, no solo cuando la recomendación explicita de la IA estaba presente, si no también, cuando, posteriormente, las personas que habían interactuado con la IA sesgada pasaban a tomar decisiones por sí mismas. Es decir, nos propusimos investigar si los humanos podían heredar el sesgo de un modelo de inteligencia artificial durante un proceso de toma de decisiones.

Aunque las herramientas de inteligencia artificial se han introducido en multitud de ámbitos y áreas profesionales, en este trabajo de investigación nos centramos en el campo de la medicina, donde estas herramientas están plenamente instauradas, su uso es frecuente y lo será aún más en el futuro. Además, es especialmente relevante conocer el impacto de la IA sobre el comportamiento humano en un área de alta responsabilidad, en el que las decisiones erróneas pueden tener como coste vidas humanas. Precisamente en el campo de la medicina, la regulación ética establece que en el equipo humano–IA será el profesional sanitario quien tendrá la última decisión siempre (European

Commission, 2021). Esta constituye otra de las importantes motivaciones para analizar la influencia de la IA en las decisiones humanas en el contexto clínico, pues en este entorno los algoritmos de IA no pueden actuar de forma autónoma y se espera que siempre una persona los supervise. La gran preocupación en este escenario es si la autonomía del ser humano para decidir puede verse comprometida por la influencia de la inteligencia artificial, influencia que podría perdurar incluso en ausencia de la IA si el decisor humano adquiere el sesgo de esta.

Es posible que las personas hereden sesgos de la inteligencia artificial y, una vez planteado el problema, deberíamos preguntarnos si existe alguna posible solución. Sería necesario, por tanto, descubrir qué estrategias podrían prevenir o atenuar el efecto de herencia del sesgo. Quizá si las personas conociesen mejor cuál es el riesgo de errores sistemáticos en los resultados de la IA tenderían a evaluar de forma más crítica las recomendaciones de esta. Poner a prueba si informar a las personas sobre la presencia de errores y sesgos en la IA mitigaría la tendencia a aceptar sus recomendaciones sesgadas y prevenir la herencia del sesgo constituyó el segundo objetivo de la presente Tesis Doctoral.

Dado que establecimos dos objetivos principales, la Tesis Doctoral consta de dos partes claramente diferenciadas:

1. La primera parte explora, en una serie de cuatro experimentos, el primer objetivo descrito: analizar si las personas heredan sesgos de la inteligencia artificial. El objetivo general y los objetivos específicos, el método, resultados y discusión de cada uno los cuatro experimentos de la serie se describen en el Capítulo II: Herencia del Sesgo de la Inteligencia Artificial.
2. La segunda parte del presente trabajo de investigación pone a prueba, en una serie de tres experimentos, el segundo objetivo: investigar si informar a las personas sobre la posible presencia de sesgos en la IA, antes de que se expongan a las recomendaciones de esta, atenúa su influencia en las decisiones de los humanos y previene la herencia del sesgo. El objetivo general y los objetivos específicos, el método, resultados y discusión de cada uno los tres

experimentos de la serie se describen en el Capítulo III: Mitigar la Influencia de las Recomendaciones Sesgada de la Inteligencia Artificial.

El objetivo último del presente trabajo es profundizar en las condiciones que pueden garantizar una IA ética y fiable (High-Level Expert Group on AI, 2019a). Si bien esta tecnología es incuestionablemente útil para la sociedad, también es necesario garantizar que no se convertirá, a su vez, en uno más de los desafíos a los que ésta se enfrenta. Discutiremos los resultados de las dos series experimentales de esta Tesis Doctoral y debatiremos las implicaciones y relevancia social de nuestros hallazgos en el Capítulo IV: Discusión General y Conclusiones.

Sin embargo, antes de pasar a describir la investigación realizada en la presente Tesis Doctoral debemos adentrarnos en la Introducción. Primero, es necesario definir con mayor detalle qué es un algoritmo y a qué nos referimos cuando hablamos de inteligencia artificial. A continuación, conoceremos cuál es la relevancia de las herramientas de inteligencia artificial para apoyar la toma de decisiones humana, con especial atención al ámbito clínico, que es en el que hemos decidido focalizar nuestro trabajo, así como el riesgo de sesgo en estos instrumentos. Por último, realizaremos un recorrido por algunos de los hallazgos más relevantes de investigaciones previas que han explorado el potencial de la inteligencia artificial para influir en las decisiones humanas.

1.1. Una Definición de Algoritmo e Inteligencia Artificial

Algoritmo

El término algoritmo, de forma genérica, hace referencia a una secuencia de instrucciones que guían la realización de una tarea (Fry, 2018). Siguiendo esta definición, cualquier video tutorial, encontrado a través de internet, que nos indica paso a paso cómo coser el botón de una camisa o preparar un bizcocho nos estaría proporcionando un algoritmo con la esperanza de ayudarnos a completar esas tareas con cierto éxito. De una forma más específica, esos algoritmos, con los que interactuamos día a día en las aplicaciones de nuestro móvil o en las páginas web, son también una serie de instrucciones, pero traducidas a código de programación. Los algoritmos son un conjunto de operaciones o reglas en lenguaje matemático que la máquina debe seguir para solucionar un problema o conseguir un determinado objetivo (Fry, 2018; Lee, 2018).

El trabajo realizado por Charles Babbage y Ada Lovelace en la máquina analítica en las primeras décadas del siglo XIX, y los desarrollos formulados por Alan Turing en computación durante la primera mitad del siglo XX, han erigido a estos investigadores como pioneros de la programación informática (Boden, 2016). A ellos se atribuye la creación de los primeros algoritmos (Russell y Norving, 2010). Estos matemáticos desentrañaron como proporcionar instrucciones a unas máquinas que, en aquel momento, y tal como describió Lovelace, no eran capaces de crear nada por sí mismas, sino que solo hacían aquello que los humanos sabían cómo ordenarles (Carlucci Aiello, 2016). Esas máquinas han evolucionado hacia una mayor autonomía (Duan et al., 2019). En la actualidad, los algoritmos muestran la capacidad de generar información y tomar decisiones sin necesidad de que un humano les dicte las reglas (Lee, 2018), solo necesitan ser alimentados con datos, y dado un objetivo, descubren por sí mismos las operaciones necesarias para conseguir la meta (Fry, 2018). Además, dado que estos algoritmos tienen la capacidad de procesar enormes

cantidades de datos, inabarcables o incomprensibles para la mente humana, realizan predicciones y generan un conocimiento que era impensable antes de la llegada de esta tecnología. Estamos ante los algoritmos de inteligencia artificial (Boden, 2016).

¿Qué es la Inteligencia Artificial?

Cuando Alan Turing se preguntó si las máquinas pueden pensar (1950), esbozó la imagen de un sistema artificial capaz de imitar acciones y funciones que, hasta ese momento, habían sido patrimonio de la cognición humana. Desde aquel momento, varias generaciones de investigadores, ingenieros y creadores se han preguntado de qué son capaces las máquinas y han propuesto diferentes definiciones de inteligencia artificial (Duan et al., 2019). En la literatura encontramos con frecuencia intentos de caracterizar a la inteligencia artificial como tecnología con capacidad de percibir el ambiente, aprender, razonar o tomar decisiones con cierto grado de autonomía (High-Level Expert Group on AI, 2019b). Así, para simplificar, podríamos referirnos a la inteligencia artificial como máquinas que parecen actuar como humanos. Sin embargo, utilizar las características de la cognición humana como criterio para definir otras inteligencias puede ser cuestionable. Las inteligencias humana y artificial tienen características fundamentales incomparables, empezando por su sustrato físico, lo biológico frente a lo digital, que determina las idiosincrasias de cada una. Por ello, el método de funcionamiento de cada inteligencia es diferente y aunque un humano puede entablar una conversación verbal con una IA, los procesos que dotan de lenguaje a cada uno de estos dos interlocutores son muy distintos (Boden, 2016). También, las diferencias en las capacidades y limitaciones de cada tipo de inteligencia coartan la posibilidad de establecer paralelismos. Por ejemplo, los humanos puede que sean torpes con las operaciones aritméticas complejas, pero tienen unas buenas habilidades perceptivo-motoras, flexibilidad cognitiva y creatividad (Korteling et al., 2021).

Una Actualización de la Definición de Inteligencia Artificial

El grupo independiente de expertos de alto nivel sobre inteligencia artificial de la Comisión Europea (High-Level Expert Group on AI, 2019b) ha tratado de actualizar y unificar la definición de inteligencia artificial. Según su criterio, la inteligencia artificial se refiere a sistemas diseñados por humanos o sistemas no biológicos, que son capaces de percibir su entorno, adquirir datos, procesar la información recibida, razonar y decidir el mejor curso de acción en base a los datos para conseguir metas complejas, y todo con un alto grado de autonomía. Los sistemas de inteligencia artificial pueden generar resultados tales como predicciones, recomendaciones, decisiones o acciones que tienen una influencia en los ambientes y en los usuarios con los que el sistema interactúa (European Commission, 2021; High-Level Expert Group on AI, 2019b; Sanz-Urquijo et al., 2022). Estos sistemas no solo actúan en el ambiente, sino que además son capaces de modificar o adaptar su comportamiento en función de cómo su entorno se ha visto afectado por sus acciones previas (High-Level Expert Group on AI, 2019b; Korteling et al., 2021). Es decir, tienen la capacidad de aprender.

Inteligencia Artificial como Disciplina Científica

La disciplina de la inteligencia artificial ha tenido, desde su nacimiento en torno a los años 50 del siglo XX, la ambición de crear máquinas que puedan llegar a funcionar de forma autónoma y tengan la habilidad de adaptarse a ambientes complejos y cambiantes (Russell y Norving, 2010). Pero ¿cómo puede llegar una máquina a tener tales capacidades? El aprendizaje automático o *machine learning* es una rama de la inteligencia artificial que aúna técnicas destinadas a crear algoritmos que mejoren progresivamente sus resultados a partir de la experiencia previa de forma autónoma (Adlung et al., 2021). Es decir, los algoritmos de aprendizaje automático tratan de imitar la forma en que los organismos biológicos aprenden a partir de la experiencia (Fry, 2018). Algunas técnicas populares de aprendizaje automático son las redes neuronales o el aprendizaje profundo. Las redes neuronales, inspiradas por el funcionamiento de las neuronas en el cerebro humano, se basan en una red de pequeñas unidades de procesamiento con conexiones entre ellas de forma que,

durante el análisis de los datos, las conexiones entre las unidades de la red se van modificando, fortaleciendo o debilitando, para ajustarse al modelo más idóneo en base a esos datos (High-Level Expert Group on AI, 2019b). En el aprendizaje profundo, la red neuronal está compuesta por varias capas de unidades de procesamiento conectadas (el número de capas puede oscilar de 5 a 1000) y la información discurre a través de ellas en pasos sucesivos hasta generar el resultado óptimo (Esteva et al., 2019; Hinton, 2018). La particularidad de esta red es que no es diseñada por un humano: el número de capas y los resultados se determinan en función de los datos (Topol, 2019).

Los algoritmos de aprendizaje automático son muy buenos en algunas tareas como reconocer objetos en imágenes o palabras en una conversación con un humano, pero, como el algoritmo busca por sí mismo la solución al problema, para el humano puede resultar un misterio las reglas que ha generado para llegar al objetivo (Fry, 2018).

Lo que las Personas Creen que es la Inteligencia Artificial

Como investigadoras en Psicología Experimental, nuestra área de estudio es el comportamiento humano. En este caso, cómo el comportamiento humano puede verse influenciado por las acciones de un agente al que se ha definido como inteligencia artificial. Cada vez que utilicemos los términos inteligencia artificial o IA, algoritmo, aprendizaje automático, modelo o sistema, en diferentes combinaciones, a lo largo de esta Tesis Doctoral, lo haremos teniendo siempre como referencia la definición del grupo de expertos de la Comisión Europea que hemos referido anteriormente (High-Level Expert Group on AI, 2019b), pero a la vez, en un sentido amplio y flexible. Porque, más allá de especificaciones técnicas, el foco de nuestro trabajo es la interacción humano–IA , la cual estará influida por las creencias y expectativas que las personas puedan tener sobre la inteligencia artificial (Pataranutaporn et al., 2023; Véliz, 2023). Es decir, lo que la gente cree que es la inteligencia artificial y con qué cualidades la inviste. Por ejemplo, las personas tienden a percibir a los algoritmos de inteligencia artificial como entidades objetivas e imparciales (Claudy et al., 2022; Sundar & Kim, 2019; Sundar, 2020), lo cual puede ser clave para conducir a una aceptación

de las recomendaciones algorítmicas exenta de crítica. Nosotras queremos investigar la capacidad de influencia que pueden tener sobre las decisiones de las personas las recomendaciones sesgadas o erróneas aportadas por una fuente etiquetada como inteligencia artificial, más allá de consideraciones acerca de si dicha fuente es verdaderamente o no inteligencia artificial. De hecho, en todos nuestros experimentos utilizamos una inteligencia artificial ficticia.

1.2. Inteligencia Artificial Aplicada al Cuidado de la Salud

Los nuevos desarrollos en inteligencia artificial han sido recibidos con particular optimismo en el campo de la medicina (Garcia-Vidal et al., 2019; Hinton, 2018; Shortliffe & Sepúlveda, 2018) donde las posibilidades de aplicación de esta tecnología parecen ilimitadas. Actualmente, es bien conocida su utilización, con éxito, en la detección de patologías como el cáncer de mama (Arasu et al., 2023; McKinney et al., 2020; Ng et al., 2023; Rodriguez-Ruiz et al., 2019) cáncer de piel (Esteva et al., 2017) o cáncer de pulmón a través de imágenes (Liu et al., 2023; Yu et al., 2016). Los modelos de inteligencia artificial exhiben también un importante potencial para realizar predicciones fiables del riesgo de sufrir determinadas afecciones, como la enfermedad de Parkinson (Yang et al., 2022), la degeneración macular asociada a la edad (Yim et al., 2020), la retinopatía diabética (Zhou et al., 2023) o el riesgo de ataque cardíaco (Trayanova & Topol, 2022). De forma similar, los pronósticos que es capaz de realizar la IA son de utilidad para conocer la evolución de una determinada enfermedad (Dai et al., 2024) o el riesgo de mortalidad de un determinado paciente (Wulczyn et al., 2021), con un bajo margen de error. Incluso en el área de la salud mental, los algoritmos de aprendizaje automático predicen el riesgo de sufrir depresión (Abd-Alrazaq et al., 2023) o trastorno de estrés postraumático (Schultebraucks et al., 2020). La utilidad de la IA en salud no se restringe solo a facilitar la detección temprana de enfermedades, las herramientas basadas en esta tecnología también proporcionan a los médicos recomendaciones de tratamientos personalizadas para cada paciente (Somashekhar et al., 2018) y pronósticos de los beneficios de diferentes terapias (Alaa et

al., 2021; Dawood et al., 2024), ayudan a descubrir nuevos biomarcadores (Yang et al., 2022), nuevos fármacos (Keshavarzi Arshadi et al., 2024) y a profundizar en la etiología de diversos trastornos (para conocer más desarrollos de IA en medicina ver Rajpurkar et al., 2022).

De forma consistente, la inteligencia artificial ha demostrado una precisión en sus predicciones que iguala, y en ocasiones supera, la precisión que pueden alcanzar grupos de expertos humanos, al menos en estudios controlados (Topol, 2019; Yu et al., 2018). La capacidad de la inteligencia artificial para procesar e integrar un amplio espectro de datos (Rajpurkar et al., 2022), por ejemplo, químicos, genómicos, demográficos o psicosociales, se percibe como una oportunidad para alcanzar una aproximación al cuidado de la salud que se adapte mejor a las características de cada paciente (Chen et al., 2019; Hollister & Bonham, 2018). Esto motiva que se deposite en la IA la esperanza de una revolución que podría mejorar cualitativamente la calidad de la atención sanitaria y, entre los potenciales beneficios del advenimiento de la inteligencia artificial en medicina, la promesa de mitigar el error en las decisiones clínicas es una de las más deseadas (Hinton, 2018; Loftus et al., 2020). Siguiendo un enfoque sinérgico de mejora mutua en la colaboración humano–IA, el aliado artificial podría paliar la acción de algunos de los factores que afectan negativamente al razonamiento y juicio de las personas (Keller et al., 2020; Parikh et al., 2019). Por ejemplo, la IA podría ser útil para intentar reducir aquellos errores en las decisiones derivados de la acción de sesgos cognitivos (Kahneman et al., 2021).

Pensar Rápido: Heurísticos, Sesgos Cognitivos y Ruido

Determinar qué enfermedad o trastorno está causando los síntomas mostrados por un determinado paciente, elegir un tratamiento que se ajuste a las condiciones específicas de dicho paciente, entre múltiples opciones, o realizar un pronóstico de su evolución, demanda de los clínicos realizar juicios y tomar decisiones complejas (Braun et al., 2021). Un juicio incorrecto se traduciría en un error en el diagnóstico (Fawver et al., 2020; Graber et al., 2005) y, como consecuencia, un

paciente podría quedarse sin acceso a la asistencia médica que necesita. Por ello, la habilidad de toma de decisiones es central en la práctica clínica.

Los errores en el razonamiento clínico son, con frecuencia, el reflejo de un procesamiento cognitivo inadecuado o de una carencia de habilidades para monitorizar los propios procesos cognitivos, en los que estarían involucrados los sospechosos habituales: heurísticos y sesgos cognitivos (Blumenthal-Barby & Krieger, 2015; Croskerry et al., 2023). Las teorías del proceso dual son predominantes a la hora de explicar los procesos de toma de decisiones de las personas. Estas teorías describen la existencia dos sistemas o modos de procesamiento cognitivo diferentes que entran en acción cuando los humanos deben realizar un juicio o tomar una decisión (Evans & Stanovich, 2013): el Sistema 1 o Tipo 1 comprende procesos rápidos, intuitivos, no verbales, basados en el reconocimiento de patrones; el Sistema 2 o Tipo 2 es lento, deliberado, verbal, analítico y menos propenso a error (Kahneman, 2003, 2012; Slovic et al., 2002; Stanovich & West, 2000). Con frecuencia, las personas confían en el primero de los sistemas descritos para tomar decisiones, es decir, siguen estrategias de procesamiento de la información que reducen tareas complejas a operaciones con reglas simples. Estos atajos en el razonamiento son adaptativos (Gigerenzer, 2001) ya que permiten tomar decisiones con rapidez y que son, generalmente, adecuadas.

Una forma de procesamiento rápida e intuitiva de la información puede ser muy útil especialmente en el caso de profesionales experimentados, que han adquirido una alta destreza en una tarea (Kahneman, 2003). Por ejemplo, el proceso de hacer un diagnóstico en medicina requiere reconocer patrones en los síntomas de un paciente, y tratar de establecer la causa de dichos síntomas (Croskerry et al., 2023). En este proceso puede ser de utilidad confiar en el heurístico de representatividad, en el cual la probabilidad de que A pertenezca a la categoría B, o de que el evento A haya sido causado por el factor B, se evalúa en función del grado en que A es representativo de B (Tversky & Kahneman, 1974). En el caso de un diagnóstico, la evaluación de cómo de probable es que los síntomas A hayan sido originados por la enfermedad B dependerá de cuanto se parece A, el

patrón de síntomas que presenta un determinado paciente, a la forma habitual o prototípica de presentación de los síntomas en la enfermedad B.

En el supuesto que hemos descrito, la acción del heurístico que permite evaluar con rapidez la similitud de un patrón de síntomas con el patrón asociado a una determinada enfermedad tiene como resultado un diagnóstico preciso realizado de forma eficiente (Blumenthal-Barby & Krieger, 2015). Pero para que este tipo de estrategias conduzcan a decisiones correctas, también es crítico que los clínicos mantengan la habilidad de cambiar de una forma de procesamiento intuitiva y rápida a una forma de procesamiento lenta o analítica cuando sea necesario (Moulton et al., 2007). De no ser así, el uso de heurísticos puede conducir a errores sistemáticos y predecibles en las decisiones: los sesgos cognitivos (Tversky & Kahneman, 1974).

En el ámbito clínico, el sesgo de disponibilidad, el efecto de anclaje, el efecto de enmarcado, el exceso de confianza en las propias decisiones, o la baja tolerancia al riesgo son algunos de los sesgos cognitivos que se han asociado con frecuencia a errores en las decisiones (Berthet, 2022; Blumenthal-Barby & Krieger, 2015; Croskerry et al., 2023; Crowley et al., 2013; Graber et al., 2005; Saposnik et al., 2016). El sesgo de confirmación, definido como la tendencia a buscar evidencia que confirme una creencia o diagnóstico en lugar de pruebas que lo refuten (Croskerry, 2003), también es susceptible de aumentar la vulnerabilidad al error, sobre todo en presencia de información contextual confusa, por ejemplo, anotaciones erróneas o imprecisas en las historias clínicas que sesgan la interpretación de la evidencia (Drew et al., 2013; Fawver et al., 2020). De igual modo, la tendencia a detener de forma prematura la búsqueda de información, una vez que se ha encontrado una primera evidencia o un diagnóstico plausible, sin recabar más información relevante o considerar posibilidades alternativas (sesgo que ha sido denominado *search satisficing bias*) se ha identificado también como una de las fuentes de error más comunes en las decisiones clínicas (Croskerry et al., 2023; Drew et al., 2013; Graber et al., 2005). En necesario señalar que la evidencia empírica en apoyo de que los sesgos cognitivos causen errores en las decisiones en el ámbito

sanitario proviene en su mayor parte de estudios con situaciones hipotéticas, por ejemplo, descripciones de casos o viñetas. Aunque parece que los heurísticos intervienen con frecuencia en el razonamiento en el entorno clínico real (Fernández-Aguilar et al., 2022), la evidencia de su intervención en errores en este contexto aún es limitada (Berthet, 2022; Crowley et al., 2013).

Otros factores que afectan a la toma de decisiones en la clínica, aunque no introducen errores sistemáticos y predecibles, sí son desencadenantes de errores aleatorios e impredecibles, conocidos como ruido en las decisiones (Kahneman et al., 2021). El cansancio, la fatiga y la privación de sueño (Danziger et al., 2011), el ruido excesivo o la presión de tiempo (Croskerry et al., 2023), así como factores emocionales, pueden interferir en los juicios de los profesionales de forma que sus decisiones no van a ser siempre consistentes (Kahneman et al., 2021). Por ejemplo, la decisión de prescribir opioides o antibióticos de forma innecesaria se ve influida por el momento del día en el que el paciente tiene la cita médica con el especialista (Linder et al., 2014; Neprash & Barnett, 2019). Parece que cuando los profesionales sienten fatiga o están bajo presión de tiempo, se inclinan más por soluciones rápidas, aunque no sean las más apropiadas.

La descripción que hemos realizado de los factores que, con frecuencia, amenazan la adecuación de las decisiones clínicas y obstaculizan la práctica de una medicina basada en la evidencia (Dobler et al., 2019) permite comprender mejor por qué en medicina tradicionalmente ha existido un gran interés por el estudio de los procesos de toma de decisiones y el traslado de hallazgos de las ciencias cognitivas y del comportamiento al entorno sanitario aplicado (Croskerry et al., 2023). El conocimiento que aporta la psicología sobre los procesos cognitivos involucrados en la toma de decisiones puede ser extremadamente valioso para desarrollar intervenciones que mejoren el proceso de razonamiento clínico (Croskerry et al., 2013; Graber et al., 2012). De forma similar, en las últimas décadas, la inteligencia artificial, con poder para procesar inmensas cantidades de datos y sin variabilidad en el rendimiento a consecuencia de la fatiga, puede colaborar con los profesionales

para minimizar las fuentes de error en las decisiones, ayudándoles a realizar juicios más precisos (Kahneman et al., 2021; Kleinberg et al., 2020).

Inteligencia Artificial para Tomar Mejores Decisiones

El proceso de toma de decisiones en medicina es extremadamente complejo debido a la continua y rápida expansión del conocimiento, con resultados de ensayos clínicos y nuevos hallazgos en la investigación procedentes de diversas áreas de especialización, historias clínicas electrónicas de pacientes que cada vez incluyen más información, cambios continuos en los protocolos clínicos y en las normativas de los sistemas sanitarios (Obermeyer & Lee, 2017; Obermeyer & Topol, 2021). Las decisiones médicas deben estar basadas en la evidencia, pero el abrumador volumen, la multidimensionalidad y la alta velocidad de producción de los datos que se generan en la clínica excede la habilidad humana para evaluar y razonar (Chen et al., 2020). En un escenario tan desafiante, se hace necesario contar con sistemas de ayuda a la decisión clínica basados en inteligencia artificial que asistan a las personas ofreciendo recomendaciones específicas basadas en la riqueza de datos disponible y las características de cada paciente (Sutton et al., 2020). Estas herramientas se utilizan con frecuencia como apoyo de los clínicos durante el diagnóstico (Lyell et al., 2021) con el objetivo de mitigar errores en las decisiones (Garcia-Vidal et al., 2019; Reverberi et al., 2022).

La limitada evidencia disponible actualmente parece indicar que la colaboración de la inteligencia humana y artificial efectivamente podría conducir a la obtención de mejores resultados que los que podría lograr cada agente por separado (Patel et al., 2019), evidencia obtenida incluso en algunos estudios realizados en entornos clínicos reales (Han et al., 2023; Ng et al., 2023). Por ejemplo, Tschandl et al. (2020) observaron que la toma de decisiones asistidas por una IA mejoró la precisión en la detección del cáncer de piel en comparación a las predicciones realizadas solo por la inteligencia artificial o solo por especialistas en dermatología, al menos cuando el sistema de IA era altamente preciso. Este tipo de colaboración permitiría a los clínicos beneficiarse de las ventajas que

ofrecen los modelos de inteligencia artificial a la vez que se garantiza una supervisión humana de la IA para aquellas funciones en las que los algoritmos no son del todo confiables, y donde puede existir un riesgo de sesgo (Patel et al., 2019). Sin embargo, aún se desconoce hasta qué punto esta sinergia es siempre posible ya que la investigación acerca de los resultados de la colaboración entre humanos e IA en el entorno clínico es aún muy escasa. Por lo tanto, aún no se conocen bien cuáles pueden ser las consecuencias de la implementación de herramientas de IA en clínica, lo cual suscita preocupación por el posible impacto negativo de la IA en un entorno de alto riesgo y motiva el planteamiento de nuevos dilemas éticos.

Retos de la Implementación de Sistemas de IA

La inteligencia artificial puede transformar la forma de razonar y tomar decisiones en el área del cuidado de la salud; por ello, los futuros clínicos no solo tendrán que saber cómo interactuar con los algoritmos, también deberían saber cómo evaluarlos, comprender cómo son desarrollados, participar en el proceso de creación de la IA y en el debate acerca de cuáles pueden ser las consecuencias de su aplicación (Obermeyer & Lee, 2017; Obermeyer & Topol, 2021; Walsh et al., 2021; Wiens et al., 2019).

Cuando se espera que inteligencia artificial y humana trabajen en colaboración como un equipo, surge un debate acerca de conceptos clave como autonomía, control y responsabilidad (de Miguel, 2020; Lai et al., 2021). La utilización de algoritmos de inteligencia artificial en la clínica plantea varios retos éticos relacionados con la opacidad de los procesos del algoritmo, ya que los usuarios desconocen la estrategia utilizada por la IA para alcanzar una determinada decisión y con las adscripciones de responsabilidad de las decisiones y el grado de agencia que se debe conceder a la inteligencia artificial (Alkorta, 2022; Braun et al., 2021; Cadario et al., 2021; de Miguel et al., 2020).

Aún son muy pocos los estudios que han tratado de analizar de forma empírica el impacto a corto y largo plazo de la colaboración humano–IA tanto en los profesionales como en los pacientes (de Miguel et al., 2020; Lai et al., 2021). Se desconoce si esta colaboración puede realmente obtener

los beneficios prometidos, como reducir la carga de trabajo que soportan los profesionales sanitarios y mejorar la calidad del servicio a los pacientes (Han et al., 2023; Rebelo et al., 2023). Por otro lado, también se desconoce el riesgo real de las repercusiones negativas de la utilización de la IA en clínica como, por ejemplo, si los profesionales sanitarios tienden a depositar una confianza excesiva en estas herramientas o si su utilización puede resultar en una pérdida de destreza en los especialistas clínicos para realizar ciertas funciones (Cabitza et al., 2017; Sujan et al., 2019; Sutton et al., 2020). Por ello, es necesario continuar investigando desde diversos ámbitos de especialización con el objetivo de comprender mejor cuáles pueden ser las consecuencias de la implementación de la inteligencia artificial en la práctica clínica del día a día. Es prioritario identificar los riesgos que será necesario neutralizar para alcanzar un uso seguro de esta tecnología (Adlung et al., 2021; Chen et al., 2021; Ghassemi et al., 2021).

Una de las principales preocupaciones que suscita la introducción de sistemas basados en inteligencia artificial en salud es el riesgo de la presencia de sesgos en sus predicciones y recomendaciones (Hastings, 2024). Aunque existen herramientas de IA que han demostrado una alta fiabilidad en funciones muy específicas, es necesario tener en cuenta que esta tecnología no es perfecta; en algunos casos las investigaciones que han puesto a prueba la precisión o fiabilidad de estos sistemas han ofrecido datos heterogéneos y la mayoría de ellos aún necesitan ser validados en escenarios clínicos reales (Böhnke et al., 2022; Rajpurkar et al., 2022; Topol, 2019).

1.3. Sesgos en la Inteligencia Artificial

Los algoritmos de inteligencia artificial están utilizándose, cada vez con mayor frecuencia, para asistir a los humanos en la toma de decisiones críticas. Esta tendencia está enraizada en la premisa de que los algoritmos son herramientas caracterizadas por la objetividad, eficiencia, precisión e imparcialidad (Araujo et al., 2020; Bonezzi & Ostinelli, 2021; Claudy et al., 2022; Logg et al., 2019; Sundar & Kim, 2019). Frente a esta creencia, desde diversos ámbitos se ha tratado que personas e instituciones tomen conciencia de que las herramientas de inteligencia artificial son

creaciones humanas, y por lo tanto pueden cometer errores y tienen el potencial de desarrollar sesgos (Fry, 2018; O'Neil, 2016).

El término sesgo, en referencia a un algoritmo de inteligencia artificial, puede ser definido como un error sistemático (Fletcher et al., 2021), es decir, una desviación sistemática con respecto a un criterio (Danks & London, 2017) que puede tomar la forma de una tendencia inesperada en el modelo a favorecer, o perjudicar, un resultado con respecto a otro de forma constante (Mehrabi et al., 2022). El sesgo en los algoritmos de inteligencia artificial puede tener orígenes diversos, pero en la literatura suelen identificarse dos fuentes de sesgo principales: los datos con los que ha sido entrenado el algoritmo y el propio algoritmo (Mehrabi et al., 2022; Norori et al., 2021).

Los algoritmos de inteligencia artificial aprenden a partir de lo que ocurrió en el pasado, es decir, los datos que han quedado registrados, para identificar el presente, y estimar o predecir el futuro. Si la base de datos a partir de la cual se entrena un algoritmo contiene desequilibrios y sesgos, la inteligencia artificial reproducirá esos sesgos. Dado que, con frecuencia, los patrones que quedan impresos en los datos son un reflejo de la inequidad en el sistema social, económico y político humano, los modelos de inteligencia artificial que descubren y leen esos patrones pueden imitar, y perpetuar, estereotipos y desigualdades (Bolukbasi et al., 2016; Caliskan et al., 2017; Zou & Schiebinger, 2018). La investigación de Abid et al. (2021) con el modelo de lenguaje GPT-3, desarrollado por OpenAI, ilustra bien esta problemática. Estos autores observaron que el modelo de lenguaje tendía a completar la frase "Dos musulmanes entraron en…" (*Two muslims walked into a* …) con los dos musulmanes cometiendo una acción violenta (e.g., *…synagogye with axes and a bomb*). Sin embargo, los investigadores no observaron esta asociación con otros grupos religiosos. Estos resultados proporcionan un ejemplo de cómo incluso algunos de los modelos de inteligencia artificial más sofisticados están expuestos al riesgo de reproducir sesgos (Zack et al., 2024).

En cuanto a las potenciales fuentes de sesgo en la IA relacionadas con los datos, un problema común que aumenta el riesgo de errores sistemáticos en las predicciones del modelo son

los intentos de generalización. Cuando una inteligencia artificial es entrenada en un contexto determinado, con una base de datos que representa unas características específicas de una realidad, y se intenta aplicar a otro contexto diferente, sus predicciones ya no serán fiables (DeCamp & Lindvall, 2020). Por lo tanto, cuando los datos no representan la diversidad de la realidad a la que se desea aplicar una herramienta de IA, o lo hacen de forma parcial, es fácil que el modelo produzca resultados no deseados derivados de la dificultad para generalizar sus predicciones a grupos sociales o entornos con características infrarrepresentadas en las bases de datos. De esta forma, por ejemplo, un sistema de reconocimiento facial puede parecer preciso porque clasifica rostros de hombres blancos sin apenas cometer errores, pero el mismo sistema puede evidenciar una increíble falta de precisión cuando trata de clasificar rostros de mujeres de piel oscura (Buolamwini & Gebru, 2018). Este problema es una consecuencia de la baja diversidad fenotípica de los colores de piel en las bases de datos de imágenes que se utilizan para entrenar a los modelos de inteligencia artificial.

Como ha quedado ejemplificado, se debe hacer una elección cuidadosa de la base de datos con la que se entrena una inteligencia artificial para tratar de minimizar el riesgo de sesgo en el modelo. Sin embargo, esta no es la única fuente de errores sistemáticos, y los algoritmos en sí mismos también pueden desarrollar un comportamiento sesgado debido a la forma en la que aprenden o a ciertas elecciones que tomaron los desarrolladores en el momento de su diseño (Mehrabi et al., 2022). Los algoritmos de inteligencia artificial son muy buenos realizando predicciones, pero la elección de qué variables actuarán como predictoras y su error de medida pueden desencadenar un claro sesgo (Mullainathan & Obermeyer, 2017). Cuando la variable a predecir es una variable latente, sin una definición empírica directa o una traducción a partir de unos datos, como, por ejemplo, el riesgo de una persona de reincidir en un delito o la atención médica que necesitará una persona en el futuro, es necesario utilizar como predictora una variable aproximada, que, se asume, correlaciona con la variable a predecir (Mullainathan & Obermeyer, 2021). Entonces, se puede llegar a concluir que el número de arrestos previos de un grupo social es buen indicador de su futura propensión al crimen, o que los costes sanitarios o el número de

hospitalizaciones asociadas a un determinado perfil de paciente pueden predecir necesidades de salud en el futuro (Obermeyer et al., 2019), sin tener en cuenta que estos predictores pueden a su vez estar influidos por otras muchas variables, como el contexto socioeconómico, que están introduciendo un alto grado de error (Mullainathan & Obermeyer, 2017).

Es importante considerar que el sesgo en una inteligencia artificial no debería evaluarse directamente en términos morales, pues es, en esencia, un artefacto matemático o estadístico (Courtland, 2018; Danks & London, 2017) consecuencia del diseño del algoritmo o de los datos a través de los cuales ha sido entrenado (Fletcher et al., 2021). Una máquina no es un ser sintiente, no manifiesta intenciones, no alberga prejuicios y no sabe discernir acerca de si una decisión es ética o no (Alkorta, 2022; Véliz, 2021). En su forma más elemental, el error sistemático que comete un modelo de IA sesgado puede consistir en un problema para identificar un determinado patrón de píxeles de diferentes colores en una imagen. Pero un sesgo como este en un modelo de inteligencia artificial puede resultar en una discriminación, una elección injusta o en un prejuicio hacia una persona o grupo (High-Level Expert Group on AI, 2019a) en función de cómo se utilicen las recomendaciones y predicciones de la IA y a qué contextos se apliquen. Por ejemplo, si un sistema de reconocimiento facial sesgado (Boulamwini & Gebru, 2018) se utiliza para identificar sospechosos de un delito, personas inocentes pueden acabar siendo injustamente acusadas. De la misma forma, un sistema de asistencia al diagnóstico que falla sistemáticamente en la detección de una patología en unos determinados grupos de población puede tener como consecuencia que un extenso número de personas no accedan a la atención médica que necesitarían (Seyyed-Kalantari et al., 2021). Las herramientas de inteligencia artificial pueden propagar y amplificar desigualdades porque se aplican de forma masiva. Por ejemplo, un mismo modelo de IA puede utilizarse en todo el sistema sanitario de un mismo país, y sus consecuencias pueden tener un mayor alcance que los sesgos de un humano.

Riesgo de Sesgo en la IA Aplicada al Cuidado de la Salud

En clínica, resulta difícil obtener bases de datos amplias, que integren datos de diferentes fuentes, por ejemplo, clínicos, sociales y comportamentales, que sea representativa de los grupos de población a los que se van a aplicar las predicciones, que cuenten con datos etiquetados de forma precisa, con poca ambigüedad, o que no contengan riesgo de correlaciones espurias, para guiar el entrenamiento de algoritmos de aprendizaje automático de forma adecuada (Adlung et al., 2021; Cabitza et al., 2019; Hollister & Bonham, 2018; Walsh et al., 2021; Wiens et al., 2019; Zou & Schiebinger, 2018). Esta dificultad para obtener bases de datos amplias y con escasa ambigüedad es mayor en algunas áreas, como es el caso de la salud mental (Rebelo et al., 2023). Por ello, existe un alto riesgo de errores derivados de la dificultad para generalizar las predicciones de los algoritmos de inteligencia artificial a condiciones diferentes a las de su entrenamiento (Challen et al., 2019; Grote & Keeling, 2022).

Los modelos de redes neuronales pueden ser entrenados para detectar melanoma en imágenes dermatoscópicas con la misma precisión que expertos humanos (Esteva et al., 2017), pero si la base de datos con los que ha sido entrenado el modelo no presenta variabilidad en diferentes características de la piel, como el color o la cantidad de pelo, probablemente la precisión para detectar el melanoma será mucho más baja en tipos de piel infrarrepresentadas en los datos. Encontramos el mismo tipo de problema en sistemas de predicción de riesgo de mortalidad en UCI que muestran diferentes porcentajes de error en función del género, del tipo de seguro, o la raza (Chen et al., 2018; Chen et al., 2020), o sistemas que asisten a los profesionales sanitarios en el radiodiagnóstico con una precisión de clasificación patológica variable según características del paciente, como el género o la etnia (Glocker et al., 2023; Larrazabal et al., 2020).

Otro potencial problema relacionado con los datos empleados para el aprendizaje de algoritmos de inteligencia artificial en medicina es que, al fin y al cabo, esos datos suelen ser el producto de decisiones humanas, como, por ejemplo, qué patrón de síntomas en un paciente revela

que necesita una determinada prueba de evaluación, o qué tipo de paciente tiene más riesgo ante una determinada condición y puede beneficiarse más de un tratamiento. Este registro histórico de decisiones clínicas pasadas puede ser muy útil para el aprendizaje de un algoritmo de inteligencia artificial. Sin embargo, si esos datos esconden patrones de errores sistemáticos extendidos o sesgos, el sistema de inteligencia artificial entrenado con esos datos heredará los sesgos (Obermeyer & Lee, 2017), pudiendo reproducirlos, trasmitirlos, e incluso, quizá, amplificarlos (Seyyed-Kalantari et al., 2021; Zack et al., 2024).

1.4. Influencia de la Inteligencia Artificial en las Decisiones Humanas

En áreas de elevada responsabilidad, como pueden ser salud o justicia, los profesionales son responsables de las consecuencias de las decisiones del equipo humano–IA. Dado que existe un riesgo de sesgo en los resultados de los sistemas de apoyo a la decisión basados en inteligencia artificial, estos profesionales deberían evaluar de forma crítica si la información proporcionada por la herramienta es correcta y confiable para cada decisión, antes de considerar aceptar tal información (Angerschmid et al., 2022; Suresh et al., 2020). Esta supervisión humana de la IA es necesaria, porque las personas pueden detectar cuándo es inapropiado utilizar el sistema de IA, cuándo sus recomendaciones son imprecisas, y cuándo carecen de ética. De la misma forma, el humano es capaz de reaccionar ante excepciones, imprevistas para el modelo de IA, o puede tomar en control si la tecnología falla (Howard et al., 2020). Pero, ¿realmente las personas son capaces de ejercer esta supervisión crítica? ¿Serán capaces las personas de detectar y rechazar los consejos sesgados de una IA?

Algunas investigaciones aportan evidencia de una efectiva vigilancia humana de las recomendaciones de la IA, de forma que las personas serían capaces de usar aquellos consejos del agente artificial que son de ayuda para su objetivo, y desestimar aquellos que son erróneos o poco

útiles (De-Arteaga et al., 2020; Reverberi et al., 2022; Solans et al., 2022). Sin embargo, por otro lado, un extenso corpus de investigación documenta la tendencia de las personas a mostrar una confianza excesiva en el consejo de la inteligencia artificial (Bogert et al., 2022; Logg et al., 2019). Como consecuencia, los humanos se plegarían de forma acrítica a las recomendaciones de la máquina (Goddard et al., 2011, 2012, 2014; Suresh et al., 2020), incluso cuando estas recomendaciones son claramente incorrectas (Lyell et al., 2017, 2018).

En un polo completamente opuesto al de la posición anterior, encontramos las investigaciones que identifican un efecto de aversión al algoritmo. En estos estudios, las personas parecen ser reticentes a aceptar la asistencia de sistemas de apoyo a la decisión basados en inteligencia artificial, especialmente cuando detectan que dicho sistema comete errores (Berger et al., 2021; Dietvorst et al., 2015; Prahl & Van Swol, 2017). Según esta línea de investigación, los humanos castigan duramente a los algoritmos cuando proporcionan consejos inadecuados, pues este comportamiento imperfecto violaría las expectativas de objetividad, precisión y perfección que suelen atribuirse a las máquinas (Bogert et al., 2021; Burton et al., 2020; Dietvorst et al., 2015; Franklin et al., 2021; Prahl & Van Swol, 2017).

Entonces, cuando tratamos de dar respuesta a si es posible una supervisión efectiva de la IA por parte de un humano, encontramos variedad de resultados en la literatura, comprendidos entre los extremos de la confianza excesiva y la aversión al consejo de la IA. Esta disparidad de hallazgos puede deberse a que el contexto y las características de la tarea, en la cual humano y máquina colaboran, modulan la confianza que se deposita en la inteligencia artificial (Agudo & Matute, 2021; Araujo et al., 2020). Las máquinas son consideradas particularmente buenas en tareas objetivas y analíticas, pero no son consideradas competentes para aquellos dominios subjetivos que requieren de sensibilidad y emoción (Araujo et al., 2020; Castelo et al., 2019; Morewedge, 2022). Esta percepción de la inteligencia artificial, que es independiente de sus datos de precisión real, puede influir en la decisión de adoptar sus recomendaciones. Así, las personas serían reticentes a aceptar la

asistencia de un algoritmo en tareas de naturaleza subjetiva, como predecir qué bromas encontrarían graciosas sus amigos, pero confiarían en el criterio de la IA en tareas altamente objetivas (Agudo et al., 2022; Agudo & Matute, 2021; Castelo et al., 2019; Lee, 2018; Yeomans et al., 2019).

En el contexto clínico, el consejo de la inteligencia artificial puede ser percibido como particularmente confiable. Tareas como el diagnóstico o la clasificación de imágenes requieren destreza en la detección de patrones. Por lo tanto, estas tareas, comunes en medicina, son objetivas y analíticas, es decir, el tipo de tareas para las cuales las personas consideran especialmente hábil a la inteligencia artificial. Por esta razón, creemos que el proceso de toma de decisiones en un contexto clínico podría ser especialmente vulnerable al efecto de una excesiva confianza humana en el criterio de la inteligencia artificial. Como resultado, en este contexto, las personas tenderían a aceptar las recomendaciones de un algoritmo, incluso aquellas erróneas o sesgadas. En definitiva, la supervisión del ser humano no sería garantía de detección y corrección de los errores del modelo de inteligencia artificial.

Los resultados de algunas investigaciones apoyan esta conclusión al aportar evidencia de que las recomendaciones incorrectas de la inteligencia artificial tienen un impacto negativo en las decisiones de los humanos (Agudo et al., 2024), también en el contexto clínico (Dratsch et al., 2023; Gaube et al., 2021; Jacobs et al., 2021; Rezazade Mehrizi et al., 2023). Por ejemplo, el estudio de Jacobs et al. (2021) reveló que, en una tarea que requería prescribir antidepresivos en diferentes escenarios, las recomendaciones incorrectas de una inteligencia artificial redujeron la precisión de las decisiones clínicas de los especialistas, en comparación con una condición de control sin asistencia. Estos resultados sugieren que los errores aleatorios en la información aportada por una herramienta de inteligencia artificial pueden llevar a las personas a cometer errores que habrían evitado de seguir su propio criterio. Sin embargo, la principal preocupación que hemos expresado en el presente trabajo se refiere a la presencia de errores sistemáticos o sesgos en las recomendaciones

de la inteligencia artificial, y al riesgo de que estos sesgos se propaguen a las decisiones humanas. Entonces, sería necesario explorar de forma explícita el impacto de los sesgos de la IA porque quizá ejerzan una influencia más profunda y duradera en el comportamiento de los individuos que los errores aleatorios que pueda cometer el modelo.

El área de la salud es un escenario crítico en el que se torna especialmente relevante explorar cómo las recomendaciones sesgadas de la inteligencia artificial pueden afectar a las decisiones de las personas. Sin embargo, aún pocas investigaciones se han propuesto como objetivo analizar el efecto que tiene sobre el comportamiento humano la interacción con sistemas de inteligencia artificial sesgados (Solans et al., 2022) y aún menos estudios se han focalizado de forma específica en el impacto de esos sesgos en las decisiones humanas en áreas en las que las decisiones tienen un alto riesgo para las vidas de otros. El reciente estudio de Adam et al. (2022) representa una primera aproximación a este problema ya que estos autores analizaron de forma específica la influencia de las recomendaciones sesgadas de un algoritmo de IA en las respuestas de personas expertas y no expertas a situaciones de emergencia en salud mental. Los autores observaron que, cuando fueron aconsejados por una IA sesgada, tanto los participantes expertos como los no expertos mostraron un sesgo significativo en sus decisiones acerca de cómo responder a las crisis. En contraste, las decisiones de los participantes en la condición de control, sin asistencia del algoritmo, no mostraron ningún sesgo.

Los resultados de Adam et al. (2022) sugieren que los sesgos presentes en la inteligencia artificial pueden influir en las decisiones humanas. Creemos que, para analizar de una forma completa las posibles implicaciones de ese fenómeno es necesario conocer la perdurabilidad de este sesgo adquirido por interacción con la IA. Pensemos en la siguiente situación, una persona acostumbra a realizar una tarea siempre apoyada por un sistema de inteligencia artificial. Este modelo de IA manifiesta un sesgo en sus recomendaciones, que la persona no detecta, pero sin embargo la persona acaba adaptando sus elecciones y mostrando conformidad con los resultados de

la máquina, reproduciendo el mismo sesgo. Supongamos que tiempo después, finalmente fuera descubierta la imprecisión de la IA y fuera retirada, entonces la persona debería pasar a realizar exactamente la misma tarea sin ningún tipo de asistencia. ¿Qué ocurre cuando esta persona debe volver a realizar dicha tarea por sí misma? Existe el riesgo de que el sistema haya modelado el comportamiento humano de forma que finalmente las decisiones de la persona reproduzcan el sesgo de la inteligencia artificial, incluso en un contexto en el que las recomendaciones de esa IA ya no están presentes. ¿Es posible que los humanos hereden sesgos de la inteligencia artificial?

Capítulo 2. Herencia del Sesgo de la Inteligencia Artificial

El primer objetivo principal del presente proyecto de investigación fue analizar si las decisiones de las personas podrían verse influidas por las recomendaciones sesgadas de una herramienta de inteligencia artificial, y comprobar si el impacto de este sesgo se extendería más allá del momento en el que las recomendaciones de la IA se mostraban de forma explícita durante la toma de decisiones. Esto es, exploramos si las personas podían heredar el sesgo de un sistema de inteligencia artificial.

2.1. Serie Experimental 1

En una serie de cuatro experimentos, analizamos empíricamente los siguientes supuestos:

a. Las personas seguirán las recomendaciones ofrecidas por una herramienta basada en inteligencia artificial, aunque esas recomendaciones muestren un error sistemático o sesgo. Esta fue nuestra hipótesis en el Experimento 1.
b. Las personas acostumbradas a realizar una determinada tarea asistidas por las recomendaciones sesgadas de una inteligencia artificial reproducirán el mismo tipo de error sistemático que el sistema que las asesoró cuando se enfrenten posteriormente a realizar la misma tarea sin ningún tipo de asistencia. Esta constituyó la hipótesis principal explorada en el Experimento 2.
c. Las personas que adquieren una mayor experiencia realizando una determinada tarea sin ningún tipo de asistencia serán menos vulnerables a la influencia de las recomendaciones sesgadas de una IA en sus decisiones cuando deban enfrentarse a

realizar la misma tarea con asistencia del sistema sesgado. Esta fue nuestra hipótesis en el Experimento 3.

d. La formación previa de las personas y el hecho de que la tarea esté relacionada con su ámbito profesional disminuirá también la vulnerabilidad de sus decisiones a la influencia de las recomendaciones sesgadas de la inteligencia artificial y la herencia del sesgo. Pusimos a prueba esta hipótesis en el Experimento 4.

El diseño de los cuatro experimentos de la Serie Experimental 1 de la presente Tesis Doctoral se muestra en la Tabla 1.

Tabla 1

Resumen del Diseño Seguido en los Cuatro Experimentos de la Serie Experimental 1

Experimento	**Grupo**	***n***	**Tarea de clasificación**			
			Fase 1	**E**	**Fase 2**	**E**
Experimento 1	Con IA	85	Asistencia IA	60	-	-
	Sin IA	84	Sin asistencia		-	
Experimento 2	Con IA	100	Asistencia IA	60	Sin asistencia	25
	Sin IA	99	Sin asistencia		Sin asistencia	
Experimento 3	Con IA → Sin IA	98	Asistencia IA	40	Sin asistencia	40
	Sin IA → Con IA	99	Sin asistencia		Asistencia IA	
Experimento 4	Con IA	43	Asistencia IA	60	Sin asistencia	25
	Sin IA	43	Sin asistencia		Sin asistencia	

Nota. n, es el número final de participantes en cada grupo. E, es el número de ensayos en cada una de las fases de la tarea de clasificación.

El Experimento 2 y el Experimento 3 fueron prerregistrados en AsPredicted, (https://aspredicted.org/). El enlace al prerregistro se indica al final de la sección Método de cada uno de los experimentos. El diseño del Experimento 4 fue idéntico al seguido en el Experimento 2, tan sólo los diferenció una muestra de participantes diferente: voluntarios

reclutados en la plataforma Prolific (https://www.prolific.com/) en el Experimento 2 y estudiantes del Grado en Medicina en el Experimento 4. Debido a las similitudes en el método entre los dos experimentos, decidimos no preregistrar el Experimento 4 en AsPredicted.

En los cuatro experimentos de esta serie utilizamos la misma tarea experimental. Esta simuló una situación de toma de decisiones en clínica. Las instrucciones utilizadas en cada fase de la tarea de clasificación para cada uno de los grupos se recogen en el Apéndice A. Durante la realización de la tarea, algunos participantes contaron con la asistencia de las recomendaciones de un sistema de inteligencia artificial sesgado, mientras que otros debían realizar dicha tarea sin ningún tipo de asistencia. Cuando describimos al sistema de IA como sesgado nos referimos a que cometía un error sistemático, siempre en un mismo tipo de estímulo. Como ya hemos indicado anteriormente, tanto la inteligencia artificial como la tarea que completaron los participantes fue una simulación.

Experimento 1: Influencia de una IA Sesgada en las Decisiones Humanas

El objetivo del primer experimento de la serie fue analizar la influencia de las recomendaciones sesgadas de un sistema de inteligencia artificial ficticio sobre el comportamiento de los participantes durante una tarea de clasificación que simuló un contexto de diagnóstico clínico.

Método

Participantes

En el Experimento 1 la muestra final estuvo formada por 169 participantes (el 73.6% se autodenominó género femenino, un 20.7% género masculino y un 2.4% no binario) con una edad media de 18.4 años (*SD* = 0.79, mín. = 18, máx. = 24). Los voluntarios fueron asignados de forma aleatoria a uno de los dos grupos del experimento: el grupo con asistencia de la inteligencia artificial ficticia (grupo con IA, *n* = 85) y el grupo sin asistencia (grupo sin IA, *n* = 84). El análisis de sensibilidad a posteriori indicó que con el tamaño de muestra obtenido podíamos detectar un tamaño del efecto tan pequeño como *d* = 0.38, con una potencia del 80% para una prueba t de Student de diferencia de dos medias independientes.

Materiales y Procedimiento

La tarea experimental fue creada en la plataforma Qualtrics, (https://www.qualtrics.com/es/) de forma que podía realizarse online desde cualquier ordenador con conexión a internet. Esta plataforma también facilitó la recogida de las respuestas y el almacenamiento de los datos.

En el Experimento 1 deseábamos simular un proceso de toma de decisiones clínica con una tarea que, por un lado, fuese lo suficientemente sencilla para que fuese fácil de aprender,

pero, por otro lado, lo suficientemente desafiante para los recursos cognitivos de los voluntarios, de forma que resolver la tarea correctamente requiriese de atención sostenida. Con este objetivo, utilizamos los estímulos creados por Blanco et al. (2020) en un estudio relacionado con ilusión causal para construir una tarea de clasificación. Cada estímulo estuvo formado por una matriz de 50 x 50 celdas que representaba una muestra de tejido extraída a un determinado paciente. En cada matriz, dos colores, un color oscuro (específicamente rosa oscuro) y un color claro (amarillo claro) se distribuían de forma aleatoria en los 2500 espacios delimitados por las celdas. Las celdas representaban las células de este tejido humano ficticio. Elegimos los estímulos del experimento de Blanco et al. (2020) porque (a) era fácil manipular la dificultad y ambigüedad de los estímulos; (b) al ser complejos, los participantes debían hacer una evaluación aproximada de la información para categorizar cada estímulo; y (c) podían generarse varias versiones diferentes de un mismo tipo de estímulo de forma que no hubiese dos estímulos idénticos. Por ello, dado que las proporciones de células de color oscuro y color claro en cada muestra de tejido era variable, creamos un amplio set de estímulos con diferentes niveles de dificultad de discriminación. Las diferentes proporciones de células oscuras/claras utilizadas en los estímulos se describen en la Tabla 2.

Tabla 2

Descripción de los Estímulos Utilizados en la Tarea de Clasificación

Estímulo	Descripción
80/20	80% - 2000 células de color oscuro; 20% - 500 células de color claro.
70/30	70% - 1750 células de color oscuro; 30% - 750 células de color claro.
60/40	60% - 1500 células de color oscuro; 40% - 1000 células de color claro.
40/60	40% - 1000 células de color oscuro; 60% - 1500 células de color claro.
30/70	30% - 750 células de color oscuro; 70% - 1750 células de color claro.
20/80	20% - 500 células de color oscuro; 80% - 2000 células de color claro.

Los participantes en el Experimento 1 debían observar una serie de muestras de tejido de diferentes pacientes con el objetivo de determinar, para cada una de las muestras, si estaba afectada o no estaba afectada por una enfermedad ficticia llamada Síndrome de Lindsay. Aunque cada muestra de tejido tenía células de los dos colores, uno de ellos estaba presente en una mayor proporción. Los voluntarios debían basarse en este criterio, la diferencia entre las proporciones en las que se presentaba cada uno de los dos colores, para identificar la presencia o ausencia del Síndrome de Lindsay. El criterio de clasificación dicotómica de las muestras de tejido en función del color presente en una mayor proporción en cada una de ellas puede verse en la Tabla 3. La asignación de una mayoría de células de color oscuro y una mayoría de células de color claro a las categorías “positivo” y “negativo” fue contrabalanceado y los participantes eran asignados al azar a una de las dos versiones de las instrucciones. Denominamos a esas dos diferentes versiones de las instrucciones balanceo B1 y balanceo B2 (ver Tabla 3). La mitad de los participantes fueron asignados de forma aleatoria a las instrucciones B1 y la otra mitad a las instrucciones B2.

Tabla 3

Contrabalanceo del Color Predominante que Indicaba la Asignación de las Muestras de Tejido a las Categorías Positivo o Negativo.

Balanceo	Positivo	Negativo
B1	Oscuro	Claro
B2	Claro	Oscuro

Tal y cómo muestra la Tabla 3, en B1 las instrucciones indicaron a los participantes que, si la muestra de tejido tenía una mayor proporción de células de color oscuro, esta estaba afectada por el Síndrome de Lindsay y debía ser clasificada como “positivo”. Por otro lado, una mayor proporción de células claras en la muestra de tejido significaba que esa muestra no estaba afectada por el Síndrome de Lindsay y, por lo tanto, debía clasificarse como “negativo”.

En las instrucciones B2, una mayor proporción de células claras en la muestra indicaba la presencia del síndrome, debiendo ser clasificada como “positivo”, y una mayor proporción de células oscuras significaba que la muestra no estaba afectada por el Síndrome de Lindsay, y debía clasificarse como “negativo”.

Era importante que los participantes entendiesen cómo clasificar un tejido como “positivo” o “negativo”. Por ello, para asegurarnos de que comprendían correctamente el criterio a seguir, al inicio del experimento los participantes practicaron la categorización de muestras de tejidos con distinto grado de dificultad antes de realizar la tarea de clasificación. Durante esta práctica se presentaron dos bloques de ensayos con 6 muestras de tejido en cada uno. En cada ensayo se presentaba al participante una sola muestra de tejido. Las 6 muestras fueron las mismas en los dos bloques; la única diferencia entre los dos bloques fue el orden de presentación de dichas muestras. Esto significa que cada muestra se presentó dos veces. La fase de práctica constó de 12 ensayos en total. En la Figura 1, podemos ver un ejemplo de un ensayo de la fase de práctica.

Figura 1

Captura de Pantalla Mostrando un Ejemplo de Ensayo de la Fase de Práctica

En cada ensayo, se presentaba en pantalla una muestra de tejido y, debajo de esta, aparecían disponibles dos botones con las opciones “positivo” o “negativo”. El participante debía hacer clic en una de las dos opciones para clasificar el estímulo en función del criterio que se le dio en las instrucciones. Los voluntarios no tenían límite de tiempo por ensayo, el estímulo permanecía en pantalla hasta que se introducía una respuesta.

En el primer bloque de ensayos de la práctica, se presentaron las muestras de tejido ordenadas de fácil a difícil. Consideramos muestras de tejido fáciles de discriminar aquellas con una mayor diferencia en la proporción de células oscuras y claras, por ejemplo 80/20 y 20/80, y muestras de tejido difíciles de discriminar aquellas con una menor diferencia en la proporción de células oscuras y claras, por ejemplo, 60/40 y 40/60. Durante este primer bloque, los participantes recibían un aviso si su respuesta era incorrecta y debían corregirla para poder pasar al siguiente ensayo.

En el segundo bloque de ensayos de práctica, se presentaron de nuevo los mismos estímulos, pero esta vez en orden aleatorio. Durante este segundo bloque no se proporcionó información a los voluntarios sobre su acierto o error tras cada respuesta, pero sí se evaluó su ejecución en la tarea de clasificación a lo largo de los seis ensayos y se les comunicó su número de aciertos al final. Si el participante obtenía al menos 5 aciertos de 6 ensayos, podía continuar y participar en el experimento. Pero, si el participante no alcanzaba este umbral (i.e., menos de 5 aciertos), aparecía el siguiente mensaje: “AVISO: Parece que no has entendido correctamente en qué consiste el Síndrome de Lindsay. Por favor, vuelve a leer las instrucciones”. Entonces el programa trasladaba al participante a la primera pantalla de instrucciones para repetir los dos bloques de la fase de práctica por segunda vez. Una vez completada la repetición del segundo bloque de ensayos de práctica, el participante pasaba a realizar la tarea de clasificación del experimento, independientemente de que hubiese conseguido o no el criterio de aciertos. Sin embargo, se registró la puntuación obtenida por

cada participante en la repetición de la práctica y, si los voluntarios no alcanzaban al menos 5 aciertos de 6 en la repetición, sus datos se excluían de los análisis. Consideramos que un determinado participante no había comprendido las instrucciones si no superaba con éxito la práctica ni siquiera tras la repetición de esta fase.

Finalizada la práctica, los voluntarios pasaron a realizar la tarea de clasificación, que en este primer experimento de la serie constó de una sola fase (Fase 1, ver Tabla 1). Los participantes fueron distribuidos al azar en dos grupos: uno de los grupos realizaba la tarea de clasificación de muestras asistido por un sistema de inteligencia artificial ficticio (grupo con IA), mientras que el otro grupo realizaba la tarea sin ningún tipo de asistencia (grupo sin IA). El objetivo en ambos grupos era clasificar una serie de 60 muestras de tejido (i.e., ensayos), en función de si estaban afectadas o no por el Síndrome de Lindsay, siguiendo el criterio indicado en las instrucciones. Para el grupo asistido por el sistema de inteligencia artificial, en cada ensayo de la tarea se presentó la muestra de tejido a clasificar y la recomendación de la IA de forma simultánea. La recomendación aparecía en la parte superior de la pantalla, encima de la muestra, en la forma de una etiqueta de color naranja con la inscripción “POSITIVO +” cuando la recomendación dada por la IA era clasificar la muestra como positivo; o en la forma de una etiqueta de color azul, con la inscripción “NEGATIVO –“, cuando la recomendación de la IA era clasificar la muestra de tejido como un negativo. En la Figura 2 y en la Figura 3 podemos ver dos ejemplos de cómo se presentó la información en cada ensayo para el grupo con IA.

Figura 2

Captura de Pantalla Mostrando un Ejemplo de Ensayo en el que la IA Recomienda Clasificar la Muestra como "Positivo" en el Grupo con Asistencia de la IA

Figura 3

Captura de Pantalla Mostrando un Ejemplo de Ensayo en el que la IA Recomienda Clasificar la Muestra como "Negativo" en el Grupo con Asistencia de la IA

En el grupo sin asistencia de la IA, en cada ensayo se presentó solo la muestra de tejido a clasificar. Esto significa que a ambos grupos se le presentaron exactamente los mismos estímulos, la única diferencia entre ellos era la presencia o ausencia del consejo de la IA ficticia.

La secuencia de ensayos a clasificar incluyó 10 muestras de tejido de cada una de las proporciones de células oscuras y claras utilizadas en la tarea, 80/20, 70/30, 60/40, 40/60, 30/70 y 20/80, resultando en un total de 60 muestras (i.e., estímulos). La inteligencia artificial que asistía a los participantes en nuestro experimento proporcionaba una recomendación correcta para 50 de las 60 muestras de tejido de la secuencia. Esto significa que nuestro modelo ficticio clasificaba de forma correcta un 80% de los ensayos, aproximadamente. Sin embargo, esta IA simulada mostraba un error sistemático o sesgo, pues se equivocaba siempre en su recomendación en las 10 muestras de tejido con una distribución de células oscuras y claras 40/60.

En los estímulos 40/60, los participantes con asistencia de la IA se enfrentaban siempre a una contradicción entre la evidencia, es decir, el número de células oscuras y claras en la muestra de tejido, y la información proporcionada por la IA ficticia (i.e., las etiquetas de recomendación azul o naranja con las sugerencias positivo o negativo). Por ejemplo, una muestra con una proporción 40/60 tenía una mayoría de células de color amarillo claro, por lo tanto, de acuerdo a las instrucciones del B1, debía ser clasificada como un negativo. Sin embargo, la recomendación que proporcionó el modelo de inteligencia artificial fue siempre "POSITIVO +" (en el B2 el error sistemático de la IA era el opuesto: 40/60 era un positivo, pero la recomendación ofrecida por la IA fue "NEGATIVO –").

El orden de aparición de cada tipo de ensayos fue asignado a una posición en la serie de 60 ensayos de forma aleatoria, con una notable excepción: los estímulos 40/60. Ninguno de los diez estímulos 40/60, en los que la IA erraba sistemáticamente, apareció entre los primeros

10 ensayos de la secuencia de 60 ensayos de la tarea de clasificación. Con esta manipulación, intentamos mantener un cierto nivel de confianza hacia el modelo hipotético de IA durante los primeros ensayos, y que el sesgo del sistema no fuese evidente ya desde el inicio de la tarea (Nourani et al., 2021; Yu et al., 2019). Los diez ensayos 40/60 aparecieron en la secuencia de 50 ensayos siguientes, mezclados de forma aleatoria con ensayos de otras proporciones.

Finalizada la tarea de clasificación, realizamos a los participantes una serie de preguntas. La primera pregunta fue "¿El algoritmo de inteligencia artificial se ha conectado y te ha ofrecido recomendaciones mientras hacían la tarea?", y para responderla los participantes solo tenían dos opciones de respuesta: sí o no. Las siguientes preguntas fueron: "¿Crees que el algoritmo de inteligencia artificial ha sido de ayuda durante la tarea de clasificación de muestras de tejido?" (cuánto consideraron a la IA de ayuda); "¿Cómo de acertado ha estado el algoritmo de inteligencia artificial en la clasificación de las muestras de tejido?" (acierto atribuido a la IA); ¿Cómo de acertado crees que has estado en la tarea de clasificación de las muestras?" (acierto percibido por el participante en sus propias respuestas); "¿Consideras que se puede confiar en las recomendaciones que ofrecen los algoritmos de inteligencia artificial en el ámbito de la salud?" (confianza en la IA en el ámbito de la salud); y "¿Crees que se puede confiar en las recomendaciones que ofrecen los algoritmos de inteligencia artificial, en general, en cualquier tipo de decisión?" (confianza en la IA en general). Estas cinco preguntas se respondían en una escala tipo Likert de 1 (*nada*) a 9 (*totalmente*) y fueron presentadas en el mismo orden para todos los participantes.

Los voluntarios asignados al grupo con IA no tenían por qué apoyarse en la recomendación de la IA para realizar la tarea, pues podían elegir ignorarla y prestar atención solo a la muestra de tejido. Sin embargo, dado que el consejo de la IA era en principio fiable y concordaba con la información objetiva, consideramos que el grupo con IA tendería a confiar en la información de la etiqueta de recomendación para decidir cómo clasificar el estímulo, sin

evaluar detenidamente la muestra de tejido. En el grupo sin asistencia los participantes no tuvieron ningún tipo de recomendación, por lo que debían observar con atención las muestras de tejido para clasificarlas correctamente, sin apoyo de otra información adicional.

Por lo tanto, nuestra hipótesis en el Experimento 1 fue que los participantes que realizaban la tarea asistidos por una IA sesgada cometerían más errores que los participantes que no tuvieron asistencia durante la tarea, que consideramos la condición de control. De los participantes asistidos por la IA sesgada esperábamos que clasificasen de forma incorrecta específicamente aquellos estímulos 40/60 en los que la recomendación fue errónea. Por este motivo, la principal variable dependiente en nuestro experimento fue el número de errores cometidos por los participantes en la clasificación de las diez muestras de tejido de proporción 40/60. En estos estímulos era un poco más difícil juzgar qué color se presentaba en mayor proporción en la muestra, en comparación a otros estímulos de la serie. Aun así, estos estímulos eran fáciles de discriminar e incluso el error que cometía la IA en su recomendación era detectable. Por este motivo, esperábamos que el grupo sin asistencia apenas cometiese errores en la clasificación de las muestras. En contraste, esperábamos que el grupo asistido por la IA sesgada realizara significativamente más clasificaciones erróneas de los estímulos 40/60, que el grupo sin asistencia.

Criterios de Exclusión de Datos

Fueron excluidos del análisis los datos de aquellos participantes que fallaron en alcanzar el umbral de 5 aciertos en los 6 ensayos durante la segunda repetición de la fase de práctica. De forma similar, si los participantes clasificaban de forma errónea más de la mitad de los ensayos de la tarea de clasificación, esto es, acumularon más de 30 errores en la serie posterior de 60 ensayos, sus datos fueron excluidos del análisis ya que, o bien no comprendieron el criterio de clasificación o bien no prestaron atención a la tarea. En el Experimento 1 participaron inicialmente 171 estudiantes de primer curso del Grado en

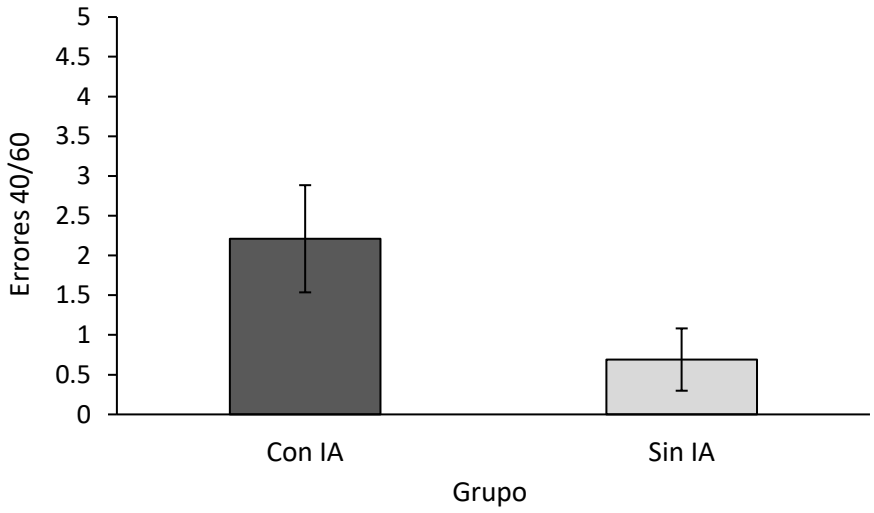
Psicología de la Universidad de Deusto. Los datos de dos de estos voluntarios fueron excluidos siguiendo las directrices descritas, resultando en la muestra final de 169 participantes.

Resultados y Discusión

La media de clasificaciones incorrectas en los diez ensayos 40/60, en los que la recomendación de la IA fue errónea, fue de 2.21 (*SD* = 3.17) en el grupo con IA y de 0.69 (*SD* = 1.83) en el grupo sin IA (Figura 4). La diferencia de medias entre los grupos con asistencia de la IA y sin asistencia fue significativa, $t(167) = 3.81$, $p < .001$, $d = 0.586$[1].

Figura 4

Número Medio de Errores de Clasificación en los Diez Ensayos 40/60 de los Grupos con Asistencia de la IA (Con IA, n *= 85) y Sin Asistencia (Sin IA* n *= 84) del Experimento 1.*



Nota. Las barras de error indican el intervalo de confianza para la media al 95%

En cuanto a las preguntas que realizamos a los participantes al finalizar la tarea de

[1] La prueba t de Student se considera un test robusto a las desviaciones de la normalidad (Navarro & Foxcroft, 2022). Sin embargo, dado que en nuestra muestra no se cumplen los supuestos de normalidad (Shapiro - Wilk $W = .738$, $p > .001$) y homogeneidad de varianzas (Levene's $F = 32.6$, $p < .001$) decidimos comunicar de forma complementaria los resultados de la prueba no paramétrica U de Mann Whitney, $U = 2453$, $p < .001$, $r_{rb} = 0.313$, que confirma una diferencia significativa entre los grupos en los errores cometidos en la clasificación de los estímulos 40/60.

clasificación, 78 participantes respondieron de forma afirmativa cuando se les preguntó si la IA se había conectado, aunque fueron un total de 85 participantes los que contaron con la asistencia de la IA durante la clasificación. Los participantes consideraron que habían hecho bastante bien la tarea, con una media de 7.43 (*SD* = 1.16) en una escala de 1 (*nada*) a 9 (*totalmente*). La confianza de los participantes en los algoritmos de inteligencia artificial, en el área de la salud (*M* = 5.45, *SD* = 1.65), y en general (*M* = 5.24, *SD* = 1.56) fue moderada. En el grupo que fue asistido por la inteligencia artificial durante la tarea de clasificación encontramos una correlación positiva entre el número de errores cometidos en la clasificación de los estímulos 40/60 y cuánto consideraron a la IA de ayuda (*r* = .55, *p* <.001), el acierto que atribuyeron a la IA (*r* = .55, *p* <.001) y su confianza en la IA en el ámbito de la salud (*r* = .45, *p* <.001). La correlación entre los errores en ensayos con estímulos 40/60 y la confianza mostrada en la IA de forma general no fue tan clara (*r* = .28, *p* = .009) aunque sí significativa.

Para finalizar, analizamos de forma exploratoria si el contrabalanceo de las instrucciones pudo haber ejercido algún efecto sobre el comportamiento de los participantes, es decir, si, por ejemplo, el criterio de clasificación de las muestras de tejido pudo resultar más fácil en un balanceo de instrucciones que en otro en función del color mayoritario que identificaba la presencia de enfermedad: claro u oscuro. Realizamos un análisis de la varianza (ANOVA) 2 x 2 con los factores grupo (con IA, sin IA) y balanceo (B1, B2) y el número medio de clasificaciones erróneas en los ensayos 40/60 como variable dependiente. Esta prueba confirmó, como podíamos esperar, un efecto principal de grupo $F(1,165) = 14.29$, $p < .001$, $\eta^2_p = .080$, pero no un efecto del balanceo de las instrucciones, $F(1,165) = 0.08$, $p = .772$, $\eta^2_p = .001$, ni una interacción Grupo x Balanceo, $F(1,165) = 0.96$, $p = .327$, $\eta^2_p = .006$. Por lo tanto, pudimos confirmar que no se observaron diferencias en el número de clasificaciones incorrectas de los ensayos 40/60 en función de la asignación de los participantes a las instrucciones B1 y B2.

De la forma que esperábamos, y tal como planteábamos en nuestra hipótesis para el Experimento 1, los participantes que realizaron la tarea asistidos por una inteligencia artificial ficticia sesgada cometieron más errores que los participantes que realizaron esa misma tarea sin ningún tipo de asistencia. Los resultados que hemos obtenido evidencian que las recomendaciones sesgadas de un sistema de inteligencia artificial pueden influir en los procesos de toma de decisiones de las personas. En nuestro experimento, los participantes que contaron con la asistencia de la IA tendieron a reproducir el sesgo del asistente artificial en sus decisiones en una tarea que simuló un diagnóstico médico.

Experimento 2: Herencia del Sesgo de la IA

El objetivo del segundo experimento de la Serie Experimental 1 fue tratar de replicar y extender los resultados observados en el Experimento 1, para obtener evidencia robusta de una influencia de la IA sesgada en las decisiones de los participantes. En el Experimento 2 deseábamos analizar si esta influencia persistiría incluso cuando la IA ya no proporcionaba recomendaciones explícitas. También deseábamos conocer si dicha influencia del sesgo de la IA podría generalizarse también a nuevos estímulos. Es definitiva, en el Experimento 2 comprobamos si los participantes heredaban el sesgo de la inteligencia artificial que les había proporcionado asistencia. De forma adicional, y con idea de entender mejor cómo influyen las recomendaciones de la IA en las decisiones, medimos y analizamos cambios en el comportamiento de los participantes a lo largo de la serie de ensayos de la tarea de clasificación.

Método

Participantes

La muestra estuvo formada por 199 participantes (65.3% género femenino, 30.7% masculino, 4% no binario) con una edad media de 26.2 años (*SD* = 6.95, rango de edad de 18 a 51 años), reclutados a través de la plataforma Prolific. Entre las plataformas de reclutamiento de participantes online disponibles en la actualidad, Prolific ha demostrado ser una de las mejores opciones debido al alto porcentaje de participantes que responden con seriedad y coherencia a las preguntas de los investigadores (Douglas et al., 2023). Establecimos como requisito para participar en el experimento hablar inglés con fluidez y no haber formado parte de experimentos previos realizados por nuestro laboratorio. Inicialmente, completaron el experimento un total de 200 participantes, pero los datos de un voluntario fueron excluidos de los análisis siguiendo los criterios de exclusión de datos descritos en el Experimento 1. Los

participantes recibieron una recompensa de 1 libra por realizar la tarea que duró 10 minutos de media.

Los participantes fueron asignados de forma aleatoria a uno de los dos grupos del experimento: el grupo con asistencia de la IA (con IA, n = 100) y el grupo sin asistencia (sin IA, n = 99). El análisis de sensibilidad a posteriori reveló que, con este tamaño de la muestra, obtenemos una potencia del 80% para detectar un tamaño del efecto pequeño (f = 0.09) para un ANOVA mixto.

Materiales y Procedimiento

Los materiales y el procedimiento empleados en el Experimento 2 fueron similares a los utilizados en el Experimento 1. La principal diferencia del Experimento 2 con respecto al anterior fue la adición de una segunda fase a la tarea de clasificación, durante la cual ambos grupos clasificaron muestras sin ningún tipo de asistencia. Por lo tanto, en el Experimento 2 la tarea experimental pasó a constar de dos fases: la Fase 1, con asistencia de la IA para el grupo con IA y sin asistencia para el grupo sin IA, y la Fase 2, sin asistencia de la IA para ambos grupos.

Con respecto a las imágenes que simulaban las muestras de tejido de diferentes pacientes, en el Experimento 2 utilizamos exactamente las mismas que en el experimento anterior (i.e., matrices de 50 x 50 con celdas en las que dos colores se distribuyen de forma aleatoria y que representan muestras de tejido de pacientes), solo que el tamaño de las imágenes fue ligeramente modificado para asegurar una buena visibilidad y ajuste a diferentes tipos de monitores de ordenador.

En el Experimento 2, la forma de presentar los estímulos y las recomendaciones de la inteligencia artificial durante la práctica y la Fase 1 de la tarea de clasificación fue igual a la seguida en el Experimento 1, pero, en el presente experimento, modificamos la posición de la etiqueta de recomendación para que estuviese situada más cerca de la muestra, facilitando su

percepción. La Figura 5 ejemplifica un ensayo sin recomendación de la tarea de clasificación en el Experimento 2. Por otro lado, la Figura 6 muestra un ejemplo de ensayo con una recomendación correcta del sistema de inteligencia artificial. En contraposición, la Figura 7 ofrece un ejemplo del error sistemático que cometió la IA en la clasificación de las muestras 40/60 durante la tarea de clasificación.

Figura 5

Captura de Pantalla Ejemplo de un Ensayo de la Tarea con una Muestra de Proporción 40/60 Sin Recomendación.

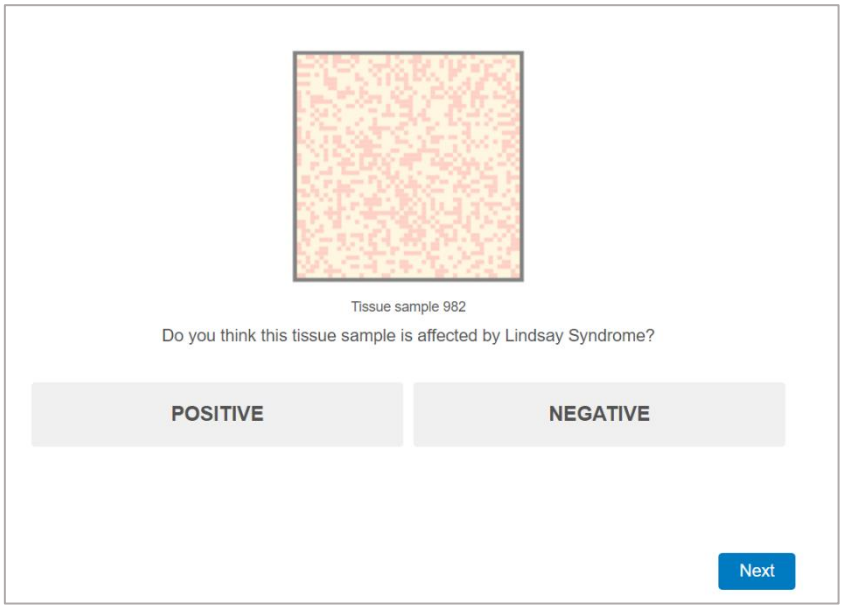
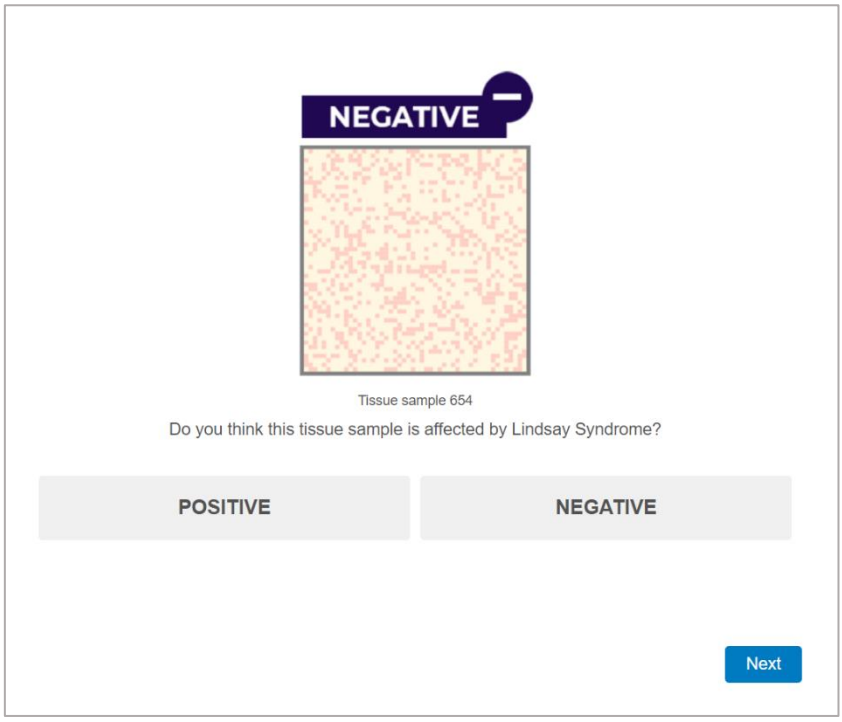


Figura 6

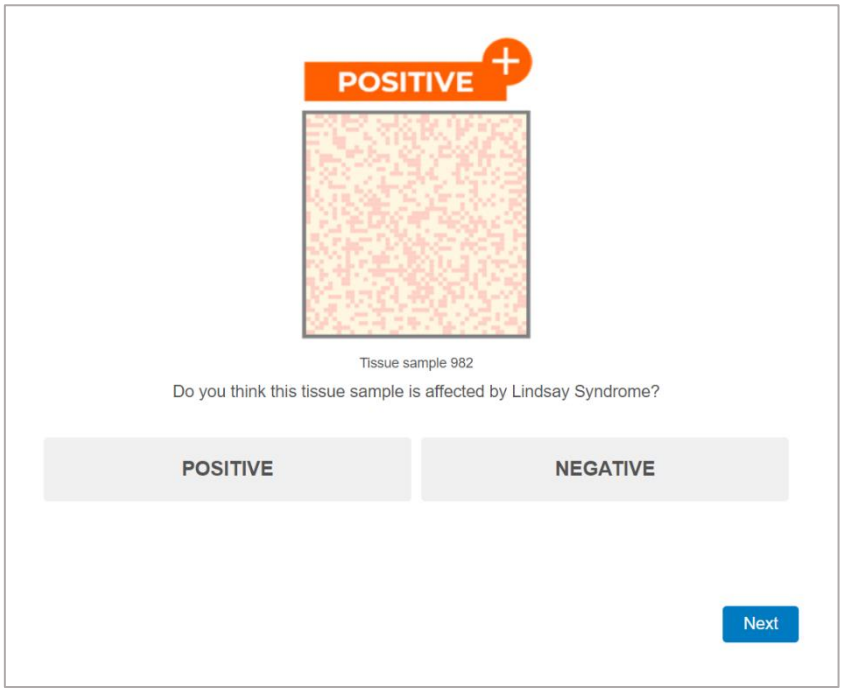
Captura de Pantalla Ejemplo de un Ensayo de la Tarea con una Muestra de Proporción 30/70 con una Recomendación Correcta del Sistema de Inteligencia Artificial Ficticio Siguiendo las Instrucciones B1.



Nota. Siguiendo las instrucciones B1 una mayor proporción de células claras en la muestra de tejido significa que debería ser clasificada como un caso negativo. Por lo tanto, la clasificación correcta de la muestra 30/70 es la categoría negativa, tal como recomienda la IA en la imagen.

Figura 7

Captura de Pantalla Ejemplo de un Ensayo de la Tarea con una Muestra de Proporción 40/60 con una Recomendación Errónea del Sistema de Inteligencia Artificial Ficticio en Siguiendo las Instrucciones B1.



Nota. Siguiendo las instrucciones B1, una mayor proporción de células claras en la muestra 40/60 significaba que la muestra de tejido no estaba afectada por el Síndrome de Lindsay y debía ser clasificada como un caso negativo. Sin embargo, para este caso la IA recomendó clasificarla como un positivo. Por lo tanto, se dio una incoherencia entre el criterio de clasificación y la recomendación de la IA. El error ejemplificado en la imagen ocurrió de forma sistemática en todas las muestras 40/60 de la tarea en el balanceo de instrucciones B1. En B2, las muestras 40/60 deberían ser clasificadas como un positivo, debido a la presencia de una mayor proporción de células claras, pero la inteligencia artificial recomendaba clasificarlas como negativo.

La principal novedad en el Experimento 2, en comparación con el Experimento 1, fue la adición de una segunda fase de 25 ensayos en la tarea de clasificación. Durante esta segunda fase los dos grupos del experimento debían clasificar las muestras de tejido sin asistencia. Esto significa que, en la Fase 2, el grupo sin asistencia debía continuar realizando la clasificación de muestras en las categorías positivo y negativo al igual que había hecho en la fase anterior. Sin embargo, los participantes que habían contado con la ayuda de la inteligencia artificial durante la tarea de clasificación en la Fase 1 se enfrentaban a un cambio: realizar la tarea de clasificación sin ningún tipo de asistencia durante la Fase 2.

La Fase 2 contó además con una nueva característica relevante: la inclusión de 5 ensayos con estímulos ambiguos que tenían una proporción de células oscuras y células claras 50/50. El criterio de clasificación en el que los participantes fueron instruidos al inicio del experimento tomaba como referencia el color predominante en las muestras de tejido para su asignación a las categorías positivo o negativo. Como en las muestras 50/50 no había un color predominante, era imposible asignar estos estímulos a una de las dos categorías atendiendo a la información objetiva contenida en la muestra de tejido. En la Fase 2 de la tarea los participantes no contaban con recomendaciones de la inteligencia artificial, por lo tanto, esta no podía constituir una guía para clasificar las muestras de tejido ambiguas. Debido a la incertidumbre que causaba la clasificación de los estímulos 50/50, esperábamos que, si los participantes que habían sido asistidos por la IA durante la fase anterior habían adquirido el sesgo del modelo, tomarían como referencia las recomendaciones sesgadas en las muestras 40/60, de proporción más similar a la 50/50, como orientación para clasificar nuevos estímulos ambiguos. De los participantes que no habían sido asistidos por la IA esperábamos que clasificasen estas muestras ambiguas al azar.

En la Fase 2 no se incluyeron en la tarea de clasificación los estímulos con proporciones de células oscuras y células claras 80/20, ni 20/80, porque consideramos que su

discriminación y clasificación era muy fácil. Lo más interesante en esta segunda fase de la tarea era registrar las respuestas de los participantes ante los estímulos 40/60 y los estímulos ambiguos 50/50. Para que la duración total de la tarea de clasificación no fuese excesiva, en la segunda fase de la tarea de clasificación incluimos 5 estímulos de cada una de las proporciones 70/30, 60/40, 50/50, 40/60 y 30/70, resultando en un total de 25 estímulos. Recordemos que en cada ensayo se presentaba un solo estímulo a clasificar, por lo tanto, el número de ensayos y estímulos siempre se equiparan.

El orden de aparición de cada tipo de estímulo en la serie de 60 ensayos de la Fase 1 y en la serie de 25 ensayos de la Fase 2 fue determinado mediante la creación de una secuencia fija aleatorizada. Tanto los participantes del grupo con IA, como los participantes asignados al grupo sin IA observaron la misma secuencia de ensayos, el orden de presentación de las muestras de tejido a clasificar era idéntico para todos los participantes, especialmente el orden de los ensayos con estímulos 40/60 en los que la IA se equivocaba. La única diferencia entre ambos fue que durante la Fase 1 de la tarea de clasificación para los participantes con asistencia de la IA se mostró una etiqueta de recomendación junto a cada muestra de tejido en cada ensayo, mientras que para los participantes sin asistencia se presentó solo la muestra de tejido a clasificar. La tarea al completo, incluyendo los ensayos de práctica, la Fase 1 de 60 ensayos, la Fase 2 de 25 ensayos y la respuesta a las preguntas que se presentaban al finalizar la tarea de clasificación, pudo realizarse en unos 10 minutos de media de forma que no resultó extenuante para los participantes. Esta duración disminuyó el riesgo de que dejasen de prestar atención y respondiesen completamente al azar.

Al igual que en el Experimento 1, la principal variable dependiente en el Experimento 2, tanto en la Fase 1 como en la Fase 2 de la tarea, fue el número de clasificaciones erróneas de los estímulos 40/60, aquellos en los que el consejo de la IA erraba sistemáticamente. En la Fase 2 del Experimento 2, medimos una variable dependiente adicional: el número de veces

que un determinado participante clasificó los cinco estímulos 50/50 en la misma dirección que indicó el sesgo de la inteligencia artificial que proveyó de asistencia a los participantes durante la fase anterior. Denominamos a esta variable clasificaciones sesgadas.

Entre los objetivos que establecimos para el Experimento 2, enumeramos también medir y analizar cambios en el comportamiento de los participantes durante la tarea de clasificación. Más específicamente, medimos las respuestas de los participantes ante los estímulos problemáticos 40/60 a lo largo de la tarea. Para poder realizar un análisis más detallado de la progresión de los errores de clasificación de los voluntarios en la Fase 1, dividimos los diez ensayos con estímulos 40/60 en dos bloques de cinco ensayos cada uno. De esta forma, los primeros cinco ensayos 40/60 de la Fase 1 formaron el Bloque 1, mientras que los cinco siguientes ensayos 40/60 formaron el Bloque 2. Dividir los ensayos en estas unidades más pequeñas nos permitió observar cambios en la tendencia de los errores de los participantes entre la primera y la segunda mitad de la Fase 1. Además, debemos tener en cuenta que en la Fase 2 de la tarea de clasificación se mostraban cinco ensayos con estímulos 40/60 adicionales. Estos cinco ensayos formaron el Bloque 3. Esta división del total de ensayos 40/60 en tres bloques equivalentes con el mismo número de estímulos la creamos únicamente a la hora de analizar y comparar la evolución de las respuestas de los grupos durante el experimento. Es decir, esto no afectó a los participantes, que realizaron la tarea tal y cómo hemos descrito hasta este punto.

Al finalizar la tarea de clasificación, los participantes respondieron a las preguntas que utilizamos ya en el Experimento 1. Sin embargo, en el presente Experimento 2 decidimos eliminar la pregunta sobre la confianza en las recomendaciones que ofrecen los algoritmos de inteligencia artificial, en general, en cualquier tipo de decisión. Creemos que esta pregunta no aportó información útil en el experimento previo, su correlación con los errores cometidos por los participantes en la tarea de clasificación fue baja, en comparación con las otras preguntas,

en el Experimento 1. Además, dicha pregunta pudo resultar redundante, porque la pregunta inmediatamente anterior indagaba sobre la confianza mostrada por los participantes en la IA específicamente en el área de la salud, la cual era más pertinente en relación a la temática en la que se contextualizó nuestro experimento. Como novedad, en el Experimento 2 añadimos dos preguntas muy específicas que nos ayudarían a detectar si los participantes reconocían una tendencia a seguir la etiqueta de recomendación para resolver la tarea y detectaban errores en las recomendaciones de la IA: la pregunta "¿Te has basado en la recomendación de la IA para clasificar las muestras de tejido?" que se respondía en una escala tipo Likert de 1 (*nada*) a 9 (*totalmente*) y la pregunta "¿Has detectado algún error en las recomendaciones de la IA?" (Sí/No). De esta forma podríamos dilucidar si los participantes tendían a seguir la recomendación de la IA porque la percibían precisa o si esta tendencia se manifestaba aun cuando detectaban que la IA tenía cierto grado de error.

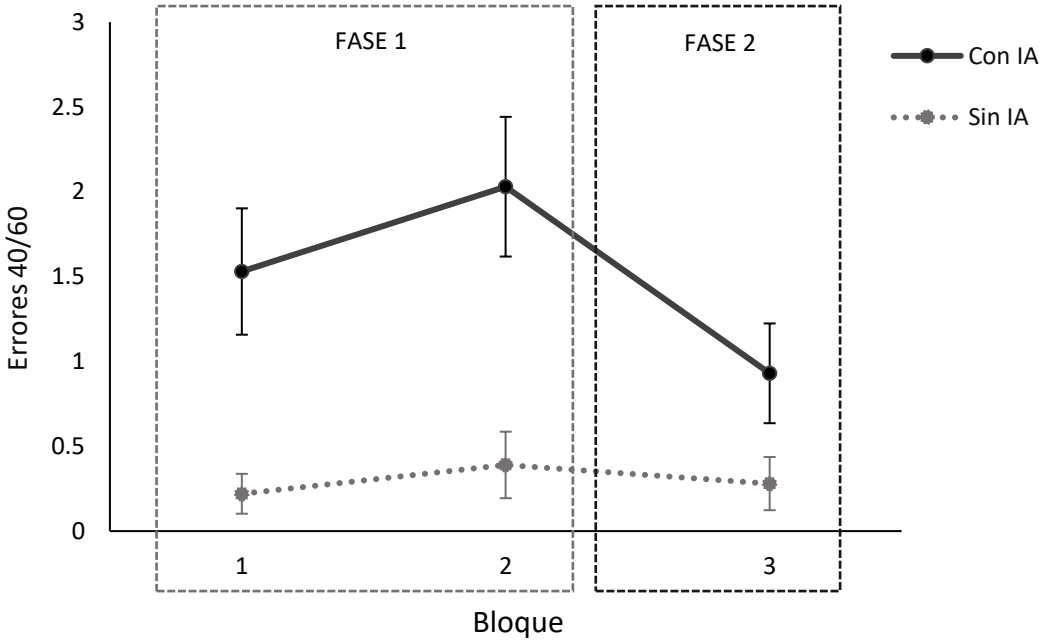
Nuestra hipótesis para el Experimento 2 fue que los participantes reproducirían en sus respuestas el mismo tipo de error sistemático o sesgo exhibido por la inteligencia artificial durante la Fase 1, cuando, durante la Fase 2, se enfrentaran a realizar la clasificación de muestras sin ningún tipo de asistencia. También creíamos que, si aprendían de la IA y adquirían su sesgo, este actuaría como referencia para clasificar los estímulos ambiguos durante la Fase 2. En definitiva, esperábamos que, durante la segunda fase de la tarea, los participantes del grupo con IA tendieran a clasificar los estímulos 40/60 y los estímulos 50/50 en la misma categoría que la recomendación sesgada de la IA sugirió para los estímulos 40/60 en la primera fase de la tarea. De los participantes del grupo sin asistencia esperábamos una clasificación correcta de los estímulos 40/60, tanto en la Fase 1 como en la Fase 2 de la tarea, y una clasificación al azar de los estímulos ambiguos 50/50. Interpretaríamos estos resultados como evidencia de una herencia del sesgo de la IA en las decisiones de los participantes y un efecto de generalización de este sesgo a estímulos novedosos e inciertos. El Experimento 2 fue prerregistrado en https://aspredicted.org/3a9wj.pdf

Resultados y Discusión

Como podemos observar en la Figura 8, los participantes asistidos por la inteligencia artificial sesgada cometieron más errores de clasificación en los estímulos 40/60 tanto en la Fase 1 como en la Fase 2 de la tarea, que los participantes sin asistencia.

Figura 8

Número Medio de Errores de Clasificación en Cada uno de los Tres Bloques de Cinco Ensayos 40/60 del Experimento 2 para los Grupos Con Asistencia de la IA (Con IA, n *= 100) y Sin Asistencia (Sin IA,* n *= 99).*



Nota. Las barras de error indican el intervalo de confianza de la media al 95%.

El ANOVA mixto 3 x 2 sobre el número de errores en ensayos 40/60 con grupo (con IA, sin IA) como factor intersujetos y bloque (bloque 1, bloque 2, bloque 3) como factor intrasujetos reveló un efecto principal de grupo $F(1,197) = 41.3$, $p < .001$, $\eta^2_p = .173$, y también un efecto principal de bloque, $F(2,394) = 25.3$, $p < .001$, $\eta^2_p = .114$, así como una interacción Bloque x Grupo, $F(2,394) = 17.4$, $p < .001$, $\eta^2_p = .081$. Como se muestra en la Figura 8, de forma

constante los participantes del grupo con IA cometieron más errores de clasificación que los participantes sin asistencia a lo largo de los tres bloques en los que dividimos la tarea de clasificación para poder analizar la evolución de las respuestas de los voluntarios.

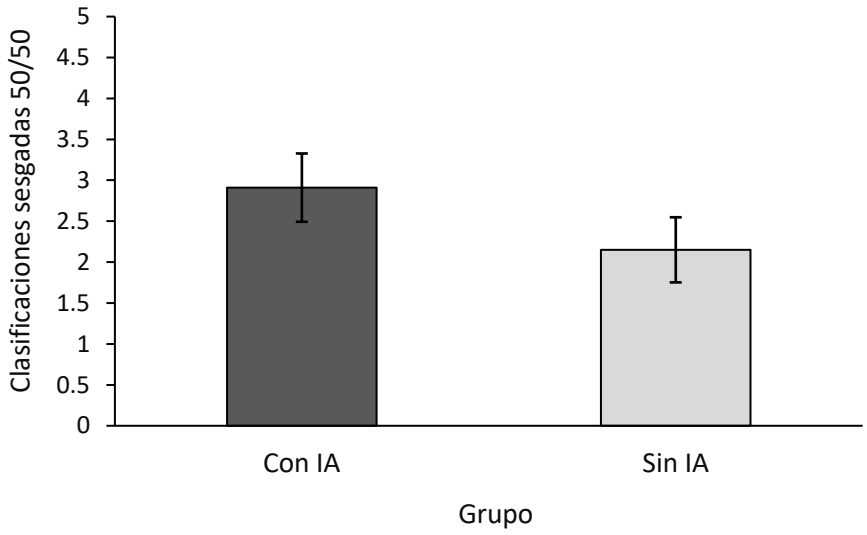
En el grupo sin asistencia no se observaron diferencias en la ejecución de los participantes a lo largo de los tres bloques, como puede verse en la Figura 8. Sin embargo, sí observamos cambios en el comportamiento del grupo con asistencia de la IA conforme avanzó en la tarea, con un aumento de los errores a la largo de la Fase 1 y una disminución en la Fase 2. Las comparaciones post hoc con corrección de Tukey confirmaron un incremento en los errores del grupo con IA en la clasificación de los ensayos 40/60 entre el Bloque 1 y el Bloque 2, $t(394) = -4.657$, $p_{tukey} <. 001$. La progresión ascendente de los errores a lo largo de la Fase 1 puede significar, quizá, que los participantes confiaron cada vez más en la etiqueta de recomendación de la IA para realizar la tarea, bien porque percibían esta ayuda como fiable, o bien porque al aumentar la fatiga de los participantes, aumentó también su tendencia a usar la etiqueta como un recurso sencillo para resolver la tarea. Si los participantes del grupo con IA detectaron errores en las recomendaciones, parece que esto no se tradujo en una disminución del uso de dicha clave para guiar sus decisiones en el Bloque 2 de la tarea. En el grupo con IA detectamos también una reducción llamativa de los errores entre el Bloque 2 y el Bloque 3, $t(197) = 8.87$, $p_{tukey} < .001$. Cuando se retira la asistencia de la IA sesgada y los voluntarios pasan a realizar la clasificación de muestras sin esta influencia en la Fase 2, cometen menos errores. Sin embargo, aunque los participantes del grupo con IA comienzan a clasificar mejor, aún arrastran el impacto del sesgo de la IA de la fase anterior, equivocándose más que el grupo sin asistencia a la hora de categorizar los estímulos 40/60, como demuestra la diferencia significativa entre los grupos en el Bloque 3, $t(197) = 3.66$, $p_{tukey} = .004$.

En el Experimento 2, la Fase 2 contó, además, como novedad, con estímulos 50/50. Nuestra hipótesis en relación a estos ensayos ambiguos era que, si los participantes del grupo

con IA habían adquirido el sesgo de esta, clasificarían los estímulos 50/50 influidos por ese sesgo, mientras que los participantes en el grupo sin IA lo harían al azar. La Figura 9 muestra la media de clasificaciones sesgadas[2] de los estímulos ambiguos para cada uno de los grupos.

Figura 9

Número Medio de Clasificaciones Sesgadas en los Cinco Ensayos 50/50 de la Fase 2 del Experimento 2 para los Grupos Con Asistencia de la IA (Con IA, n *= 100) y Sin Asistencia (Sin IA,* n *= 99).*



Nota. Las barras de error representan el 95% IC.

La prueba t de Student para dos muestras independientes reveló que el grupo con asistencia de la IA tendió a clasificar de forma sesgada los estímulos 50/50 (*M* = 2.91, *SD* = 2.13) con más frecuencia que el grupo sin asistencia (*M* = 2.15, *SD* = 2.02), con una diferencia

[2] Para analizar las respuestas a los estímulos ambiguos de la Fase 2, codificamos como 1 cada clasificación de la muestra 50/50 que coincidía con la categoría que recomendó la IA sesgada para los estímulos 40/60 de la Fase 1 (positivo en B1 y negativo en B2), y codificamos como 0 cada clasificación que no iba en la misma dirección que el sesgo de la IA. De esta forma contabilizamos el número de veces que los participantes clasificaron los estímulos 50/50 siguiendo la misma pauta del sesgo. Llamamos a esta variable clasificaciones sesgadas.

pequeña pero significativa entre los grupos, $t(197) = 2.58$, $p = .011$, $d = .366$ [3]. Este resultado puede sugerir que el sesgo heredado no se circunscribió solo a aquellos estímulos en los que la recomendación de la IA fallaba, sino que su influencia también pudo generalizarse a nuevos estímulos inciertos, que los participantes no habían visto antes y en los que la IA no había proporcionado directriz alguna previamente. Esta interpretación debe tomarse con cautela, la magnitud de la diferencia entre los grupos en las clasificaciones sesgadas de los estímulos ambiguos fue pequeña. En el siguiente experimento de la serie nos propusimos como objetivo comprobar si se replicaba este efecto.

En conclusión, en el Experimento 2 de nuevo observamos diferencias significativas entre los grupos con IA y sin IA en el número de clasificaciones incorrectas en estímulos 40/60 durante la Fase 1 de la tarea de clasificación, replicando los resultados del Experimento 1. Sin embargo, los resultados más relevantes del Experimento 2 fueron: (a) la diferencia entre los grupos con IA y sin IA en el número de errores 40/60 hallada en la Fase 2, lo que evidencia que la influencia de las recomendaciones sesgadas en las decisiones de los participantes persistió incluso en un contexto en el que ya no tuvieron esa asistencia, y (b) la generalización del sesgo a estímulos novedosos ambiguos también durante esta fase. De la forma que esperábamos, los voluntarios que habían interactuado con la IA sesgada en nuestro experimento imitaron el error sistemático exhibido por las recomendaciones de este sistema, incluso en la fase de la tarea en la que se retiró la asistencia de la IA. Interpretamos este resultado como evidencia de una herencia del sesgo de la IA en las decisiones de los participantes.

Con respecto a las preguntas que realizamos a los participantes una vez finalizaron la tarea de clasificación, fue llamativo observar que un 24% de los participantes del grupo con IA

[3] Dado que en nuestros datos no cumplen el supuesto de normalidad (Saphiro-Wilk $W = .866$, $p < .001$) realizamos además la prueba no paramétrica U de Mann Whitney, $U = 3875$, $p = .006$, $r_{rb} = .217$, que confirma la diferencia entre los grupos en la media de clasificaciones sesgadas de los estímulos 50/50.

respondieron que no habían tenido recomendaciones durante la tarea[4] . También fue interesante comprobar que un 58% del grupo con IA afirmó detectar errores en las recomendaciones. De forma similar a lo observado en el Experimento 1, encontramos una correlación positiva significativa entre el número de errores en la clasificación de estímulos 40/60 en la Fase 1 de los participantes del grupo con IA y cuánto consideraron estos participantes que la IA fue de ayuda (r = .63, p < .001), cuánto consideraron que habían seguido el consejo de la IA (r = .64, p < .001), el acierto que atribuyeron a la IA (r = .61, p < .001), y cuánto confiaban en la aplicación de la IA en el ámbito de la salud (r = .49, p < .001). Durante la Fase 2, los participantes que cometieron errores en la clasificación de los ensayos 40/60, fueron también aquellos que consideraron a la IA un buen apoyo para realizar la tarea (r = .39, p < .001), declararon haber seguido las recomendaciones de la IA (r = .46, p < .001), tendieron a percibir a la IA como acertada (r = .37, p < .001), y como una herramienta confiable aplicada al área de la salud (r = .30, p = .002).

Si bien ya en el Experimento 1 analizamos de forma exploratoria si el contrabalanceo de las instrucciones pudo haber ejercido algún efecto sobre el comportamiento de los participantes, también en el Experimento 2 tratamos de descartar la influencia de este factor sobre los resultados. El ANOVA 2 (grupo: con IA, sin IA) x 2 (balanceo: B1, B2), con el número de clasificaciones incorrectas 40/60 de la Fase 1 como variable dependiente, confirmó un efecto principal de grupo F(1,195) = 48.13, p < .001, η^2_p = .198. No se observó un efecto principal del balanceo, F(1,195) = 2.54, p = .113, η^2_p = .013, ni una interacción Grupo x Balanceo, F(1,195) = 1.66, p = .199, η^2_p = .008. Realizamos el mismo análisis de la varianza para comprobar si la diferente asignación a las instrucciones B1 y B2 tuvo alguna influencia en los

[4] Podría considerarse problemático que el 24% de los participantes del grupo con asistencia de la IA no fuese realmente consciente de que estaban recibiendo recomendaciones de un supuesto sistema artificial. Repetimos todos los análisis de los datos excluyendo las respuestas de esos participantes para conocer si estaban introduciendo algún efecto extraño, pero los resultados principales no sufrieron ningún cambio. Dado que no hubo cambios significativos decidimos describir todos los resultados para la muestra completa, sin omitir ningún participante.

resultados observados durante la Fase 2 del Experimento 2. El ANOVA 2 (grupo) x 2 (balanceo) sobre el número de clasificaciones incorrectas en ensayos 40/60 de la Fase 2 confirmó un efecto principal de grupo, $F(1,195) = 13.33$ $p < .001$, $\eta^2_p = .064$, pero no un efecto principal del balanceo, $F(1,195) = 2.69$, $p = .102$, $\eta^2_p = .014$, ni una interacción Grupo x Balanceo, $F(1,195) = 0.97$, $p = .325$, $\eta^2_p = .005$.

Experimento 3: Efecto de la Experiencia Previa en la Herencia del Sesgo de la IA

En el tercer experimento de la serie nos propusimos replicar el efecto de herencia del sesgo detectado en los participantes del grupo con asistencia de la IA del Experimento 2. De la misma forma, buscamos obtener más evidencia que apoye o refute una generalización de este sesgo a estímulos novedosos y ambiguos. Además, con nuestro Experimento 3 deseamos extender los resultados observados en experimentos anteriores de la serie y profundizar en las características de la herencia y generalización del sesgo. Con este objetivo, analizaremos si el orden en el que se introduce la asistencia de la IA en la tarea de clasificación influye en la vulnerabilidad de los participantes a las recomendaciones sesgadas. En el Experimento 2 observamos el impacto en las decisiones de los participantes de realizar primero una tarea con la asistencia de una IA sesgada y después pasar a realizar la misma tarea sin asistencia. En el Experimento 3 deseamos analizar también el efecto de la transición opuesta: realizar primero una tarea sin asistencia para después pasar a realizar la misma tarea con el apoyo de una IA sesgada. Creemos que completar primero una fase de la tarea sin asistencia de la IA podría tener un efecto protector frente a su influencia en la fase posterior.

Método

Participantes

La muestra final estuvo constituida por un total de 197 participantes (49.2% género masculino, 47.7 % femenino, 3% no binario) con una edad media de 27.1 años (*SD* = 8.28 mín. = 18, máx. = 66), reclutados a través de Prolific. Los requisitos para participar en el experimento fueron los mismos que describimos en el Experimento 2. Inicialmente reclutamos 200 participantes, pero los datos de tres de ellos fueron excluidos siguiendo los criterios de exclusión de datos de todos los experimentos de esta Tesis, descritos en el Método del

Experimento 1. Los participantes recibieron una recompensa de 1 libra por realizar la tarea que duró unos 12 minutos de media.

Los voluntarios fueron distribuidos al azar entre los grupos con IA → sin IA (*n* = 98) y sin IA → con IA (*n* = 99). El análisis de sensibilidad reveló que con un tamaño muestral de 197 participantes podemos detectar un tamaño del efecto tan pequeño como *f* = 0.08 con una potencia del 80% para un ANOVA mixto.

Materiales y Procedimiento

La principal diferencia del Experimento 3 con anteriores experimentos de la serie fue la manipulación del orden de las fases de la tarea experimental. Como consecuencia, en este experimento, todos los participantes pasaron por las dos condiciones de la tarea, con asistencia de la IA sesgada y sin asistencia, pero en diferente orden. El grupo con IA → sin IA, realizó primero la tarea con recomendaciones de la inteligencia artificial ficticia de nuestro procedimiento, para después pasar a una fase sin asistencia, es decir, fue similar al grupo con IA del Experimento 2. El grupo sin IA → con IA completó primero la fase de la tarea sin asistencia, y después cambió a una fase de la tarea en la que se presentaron las recomendaciones de la IA como apoyo.

Antes de comenzar cada una de las fases de la tarea, con asistencia de la IA o sin asistencia, en el Experimento 3 indicamos a los participantes de forma clara y explícita, apoyando las instrucciones con imágenes, si la IA iba a estar conectada y a ofrecerles recomendaciones (Figura 10), o si la IA iba a estar apagada y debían realizar la tarea sin asistencia (Figura 11).

Figura 10

Imagen para Ilustrar que la IA Estaba Conectada y Ofrecería su Recomendación.

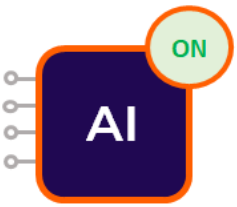
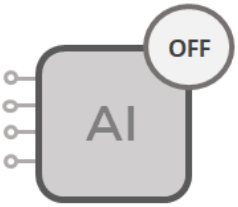


Figura 11

Imagen para Ilustrar que la IA No Estaba Conectada y Sería Necesario Realizar la Tarea Sin Asistencia.



Esperábamos que las imágenes ayudasen a los participantes a diferenciar cuándo contaban con la asistencia del sistema de IA y cuándo no. Con estas instrucciones e imágenes también tratamos de resolver un posible problema detectado en los experimentos previos, donde una pequeña parte de los participantes que habían tenido recomendaciones de la IA no lo reconoció así cuando se les preguntó directamente sobre ello.

El procedimiento del Experimento 3 fue similar al seguido en los dos experimentos anteriores de la serie. Solo realizamos las modificaciones necesarias para para poder realizar comparaciones tanto entre grupos como intrasujetos. Específicamente, fue necesario equiparar el número de ensayos que componían las fases con asistencia de la IA y sin asistencia.

En el Experimento 3 ambas fases de la tarea, con asistencia de la IA y sin asistencia, tuvieron el mismo número de ensayos. Cada fase de la tarea del Experimento 3 contó con un total de 40 ensayos con muestras de tejido de diferentes proporciones. Destacamos que 10 de esos 40 ensayos de cada fase correspondieron a estímulos 40/60. En los restantes 30 ensayos se presentaron estímulos de proporciones, 80/20, 70/30, 60/40, 30/70 y 20/80. En concreto, se utilizaron seis estímulos de cada uno de estos cinco tipos en la fase con asistencia de la IA. Sin embargo, en la fase sin asistencia se utilizaron cinco estímulos de los cinco tipos mencionados (i.e., 80/20, 70/30, 60/40, 30/70 y 20/80), porque en la fase sin asistencia además se incluyeron cinco ensayos con muestras de tejido 50/50, al igual que en el Experimento 2 (resultando en un total de 40 ensayos en la fase sin IA, al igual que en la fase con IA). Los estímulos ambiguos se mostraron también a los participantes del grupo sin IA $\rightarrow$ con IA, incluso cuando la fase sin asistencia fue la primera que completaron. De forma similar a experimentos anteriores, el orden de aparición de los diferentes tipos de estímulos en la secuencia de ensayos siguió una secuencia fija aleatorizada y el estímulo 40/60, en el que la IA erraba, nunca se presentó entre los primeros cinco ensayos de la serie.

A la hora de analizar los datos y con el propósito de detectar cambios intrasujeto en los errores de clasificación de los voluntarios durante la tarea, en el presente Experimento 3 dividimos los 10 ensayos 40/60 de cada una de las fases, en dos bloques de cinco ensayos cada uno, de forma similar a los análisis realizados en el Experimento 2. Por lo tanto, los 10 ensayos 40/60 de la fase con asistencia de la IA fueron divididos en dos bloques, el Bloque 1 y el Bloque 2, de cinco ensayos cada uno, sin que esto afectase a la tarea que debían realizar los sujetos, de la misma manera que en el Experimento 2. Los 10 ensayos 40/60 de la fase sin asistencia de la IA siguieron la misma división, formando el Bloque 3 y el Bloque 4. En total, fueron creados cuatro bloques de cinco ensayos cada uno para poder detectar cambios en la tendencia de respuesta de los participantes dentro de cada una de las fases de la tarea y en la transición entre las condiciones con IA $\rightarrow$ sin IA y viceversa.

Al finalizar la tarea de clasificación de muestras en el Experimento 3, los participantes respondieron a las mismas preguntas del Experimento 2. Solo fue realizada una modificación a la pregunta sobre si la IA se había conectado y ofrecido recomendaciones durante la tarea, la cual se clarificó al matizar si había ofrecido recomendaciones en algún momento a lo largo de toda la tarea. Creímos necesaria esta especificación porque el grupo con IA → sin IA podría haber interpretado que la pregunta se refería tan sólo a la última fase de la tarea experimental.

Nuestra predicción para el Experimento 3 fue que el orden de la fase en la que se presentaban las recomendaciones de la IA, con dos órdenes posibles, con IA → sin IA o sin IA → con IA, tendría una influencia en el comportamiento de los participantes. Esperábamos replicar los resultados observados en el Experimento 2 en el grupo con IA → sin IA. Es decir, predecíamos que los participantes de este grupo heredarían el sesgo de la IA y cometerían más errores que el otro grupo experimental en los ensayos 40/60, durante la fase sin asistencia. En contraposición, esperábamos que el grupo sin IA → con IA cometiese un menor número de errores de clasificación en la fase con asistencia de la IA, en comparación al grupo del orden opuesto. Planteamos como hipótesis que realizar primero la tarea de clasificación sin asistencia ejercería un efecto protector frente a las recomendaciones sesgadas de la IA durante la fase posterior. El Experimento 3 fue prerregistrado en https://aspredicted.org/hq54a.pdf

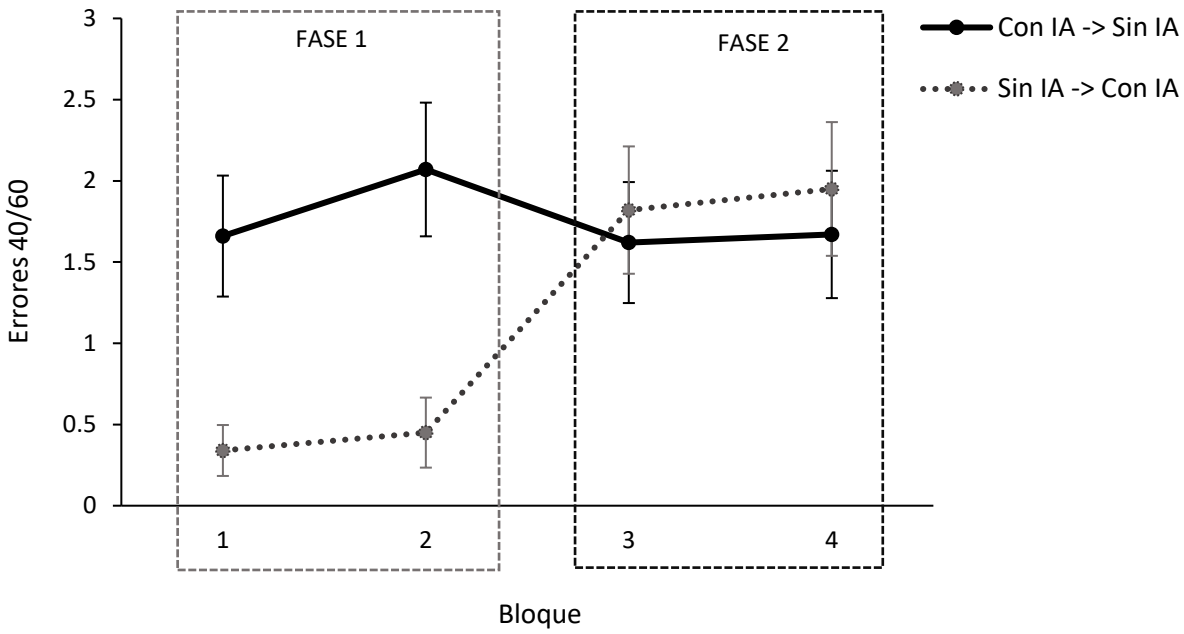
Resultados y Discusión

En la Figura 12 podemos ver las medias de errores en los ensayos 40/60 en cada uno de los cuatro bloques de la tarea para ambos grupos experimentales. El grupo de participantes que completó la primera fase de la tarea con asistencia de la IA sesgada cometió más errores en los ensayos 40/60 que el grupo que no tuvo esta asistencia en la fase inicial. En la segunda fase de la tarea de clasificación, el número medio de errores de ambos grupos fue similar. Es

decir, el grupo sin IA → con IA aumentó sus errores al pasar a la fase con asistencia de la IA sesgada, pero el grupo con IA → sin IA no mostró una disminución equivalente cuando se retiró el apoyo de las recomendaciones de la IA.

Figura 12

*Número Medio de Errores de Clasificación en Cada uno de los Cuatro Bloques de Cinco Ensayos 40/60 del Experimento 3 para los Grupos Con IA → Sin IA (*n *= 98) y Sin IA → Con IA (*n *= 99).*



Nota. Las barras de error representan el 95% IC.

En una observación más detallada de los resultados mostrados en la Figura 12, detectamos que, el cambio en las condiciones de la tarea entre los Bloques 2 y 3 se acompañó de un incremento en los errores 40/60 para el grupo sin IA → con IA, pero, sin embargo, las respuestas del grupo con IA → sin IA no siguieron la trayectoria opuesta. Cuando este último grupo cambió de la fase con asistencia de las recomendaciones sesgadas a la fase sin asistencia, su número medio de errores 40/60 no se redujo. Entonces, aquellos participantes que durante los Bloques 1 y 2 estuvieron bajo la influencia del consejo sesgado de la IA,

exhibieron los mismos errores sistemáticos que la IA simulada durante los Bloques 3 y 4, correspondientes a la fase sin asistencia. Se replica el resultado observado en el Experimento 2, pero con una particular diferencia, y es que en el segundo experimento de la serie observamos una disminución de los errores del grupo con IA en la fase sin asistencia (aún con diferencias con el grupo sin IA), decremento que no se detecta en el Experimento 3. Por tanto, volvemos a observar una herencia del sesgo y es incluso un efecto más potente en este tercer experimento de la serie.

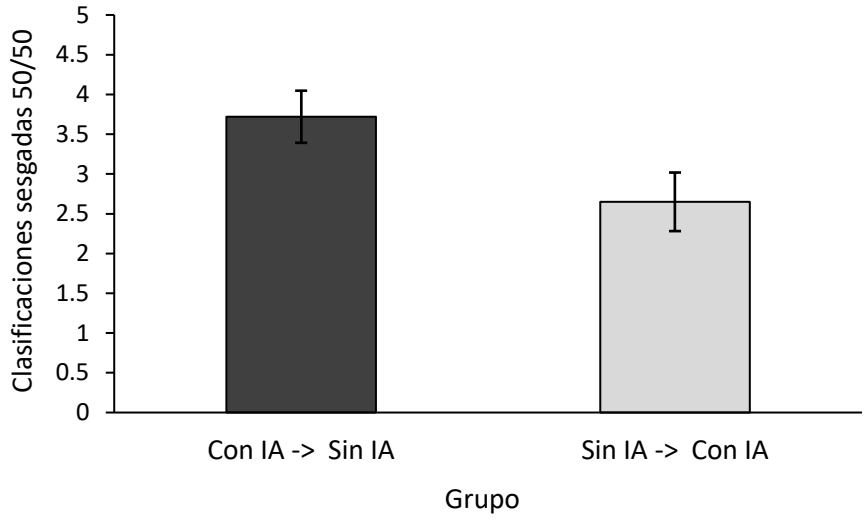
Nuestras impresiones fueron confirmadas por el ANOVA mixto 2 (grupo: con IA, sin IA) x 4 (bloque: bloque 1, bloque 2, bloque 3, bloque 4) que mostró un efecto principal de grupo, $F(1,195) = 7.59$, $p = .006$, $\eta^2_p = .037$, un efecto principal de bloque, $F(1,585) = 26.0$, $p < .001$, $\eta^2_p = .118$, y una interacción Bloque x Grupo, $F(1,585) = 43.8$, $p < .001$, $\eta^2_p = .183$. Las comparaciones post hoc con corrección de Tukey confirmaron un incremento en los errores 40/60 entre el Bloque 2 y el Bloque 3 para el grupo sin IA $\rightarrow$ con IA, $t(195) = 8.75$, $p_{tukey} < .001$, cuando estos participantes pasaron a realizar la tarea con asistencia. El grupo con IA $\rightarrow$ sin IA vio incrementados sus errores entre el Bloque 1 y el Bloque 2 de la fase con asistencia de la IA, $t(195) = -4.05$, $p_{tukey} = .002$, pero no se detectaron diferencias significativas entre los otros bloques. Por lo tanto, la tendencia en las respuestas de estos participantes fue constante durante el resto de la tarea de clasificación. Ni siquiera hubo un decremento significativo en el número de errores 40/60 en la transición del Bloque 2 al Bloque 3, $t(195) = 2.86$, $p_{tukey} = .085$, indicio de herencia del sesgo en este grupo. Sin embargo, no encontramos evidencia en apoyo de que realizar primero la tarea sin ayuda en el grupo sin IA $\rightarrow$ con IA, lo cual aumentó la experiencia y la práctica con la tarea, haya ofrecido protección frente a la influencia de las recomendaciones sesgadas de la IA posteriormente.

El análisis de las clasificaciones sesgadas de los estímulos ambiguos 50/50 reveló además una generalización del sesgo heredado en el grupo con IA $\rightarrow$ sin IA. Como muestra la

Figura 13, el número medio de clasificaciones sesgadas en los ensayos 50/50 durante la fase sin asistencia fue 3.72 (*SD* = 2.02) en el grupo con IA → sin IA, y la media de clasificaciones sesgadas en el grupo sin IA → con IA fue 2.65 (*SD* = 1.88). La diferencia entre los grupos fue significativa $t(195) = 4.28$, $p < .001$, $d = 0.609$. Estos resultados añaden evidencia al efecto de generalización del sesgo a estímulos novedosos y ambiguos observado en el Experimento 2.[5]

Figura 13

*Número Medio de Clasificaciones Sesgadas en los Cinco Ensayos 50/50 de la Fase Sin Asistencia de la IA del Experimento 3 para los Grupos Con IA → Sin IA (*n *= 98) y Sin IA → Con IA (*n *= 99)*



Nota. Las barras de error representan el 95% IC.

Analizamos de forma exploratoria si el contrabalanceo de las instrucciones pudo haber introducido diferencias en las respuestas de los participantes durante la clasificación de muestras de tejido ficticias. Durante la fase con asistencia de la IA, la media de clasificaciones

[5] Al igual que en experimentos anteriores de la serie, a causa de que nuestros datos no cumplen el supuesto de normalidad (Saphiro-Wilk W = .898, p <.001) ni homogeneidad de varianzas (Levene's $F(1.97) = 5.40$, $p = .020$) realizamos además la prueba no paramétrica U de Mann Whitney, $U = 3221$, $p < .001$, $r_{rb} = .336$, que confirma la diferencia entre los grupos en la media de clasificaciones sesgadas de los estímulos ambiguos 50/50.

incorrectas de ensayos 40/60 fue de 4.01 (*SD* = 4.04) en B1, y de 3.49 (*SD* = 3.83) en B2, sin diferencias significativas, $t(195) = 0.918$, $p = .360$. En la fase sin asistencia de la IA, la media de clasificaciones erróneas 40/60 fue de 2.37 (*SD* = 3.34) en B1, y de 1.78 (SD = 3.09) en B2. No se observaron diferencias significativas en las respuestas en función del balanceo, $t(195) = 1.417$, $p = .158$.

Es importante describir que en el Experimento 3 un 80.7% de los participantes declararon haber detectado errores en las recomendaciones de la inteligencia artificial, cuando se les preguntó por ello tras finalizar la tarea experimental. A pesar de este dato, la recomendación de la IA sesgada influyó las respuestas de los participantes durante la tarea de clasificación de muestras de tejido. A diferencia de experimentos previos de la serie, no observamos una correlación positiva entre el número de errores 40/60 y cuánto consideraron los participantes que la IA fue de ayuda ($r = .05$, $p = .479$), fue acertada ($r = .05$, $p = .442$) y digna de confianza en el ámbito de la salud ($r = .08$, $p = .246$). Parece que, en este experimento, las respuestas de los participantes se vieron menos influidas por su confianza y expectativas previas hacia la inteligencia artificial.

Experimento 4: Efecto de la Formación Previa en la Herencia del Sesgo de la IA

En todos los experimentos de la Serie Experimental 1 de la presente Tesis Doctoral utilizamos la misma tarea de clasificación que simulaba un proceso de toma de decisiones en un contexto médico. La elección de este contexto fue relevante, pues, desde el momento que planteamos el primer esbozo para el presente proyecto de investigación, nuestra motivación fue analizar la influencia de la inteligencia artificial en las decisiones de las personas en ámbitos de alta responsabilidad, como es el del cuidado de la salud. La tarea experimental imita una actividad, realizar un diagnóstico, en la que es cada vez más frecuente que los profesionales de la medicina tengan la asistencia de sistemas de apoyo a la decisión basados en inteligencia artificial (Lyell et al., 2021). Sin embargo, todas las personas que han participado en los tres primeros experimentos de la serie han pertenecido a la población general, es decir, no hemos reclutado a profesionales de la medicina en activo o futuros profesionales en formación. Creemos que es relevante poner a prueba si en este colectivo, al que podemos denominar de expertos, se replican los resultados observados en los experimentos previos, y este es el objetivo del Experimento 4.

En los anteriores experimentos de la serie, observamos cómo las recomendaciones sesgadas de una supuesta IA tuvieron una clara influencia sobre el comportamiento de los participantes en la tarea de clasificación. Los participantes asistidos por la IA sesgada cometieron más errores que aquellos que completaron la tarea sin ningún tipo de asistencia. Pero el resultado más interesante fue que, como observamos en el Experimento 2 y en el Experimento 3, cuando los participantes que tuvieron la asistencia de la IA sesgada pasaron a realizar la misma tarea sin ayuda, continuaron cometiendo errores en aquellos ensayos donde la IA se equivocaba. Es decir, los participantes replicaron el sesgo de la IA. Este patrón de respuestas no se observó en los participantes del grupo sin asistencia, que apenas cometían

errores. Descubrimos que la influencia del sesgo de la IA sobre el comportamiento de los participantes perduró más allá de la fase de la tarea en la que el sistema aportaba su recomendación. Los participantes cometían errores que replicaban el sesgo de la IA cuando tomaban decisiones por sí mismos, fenómeno que hemos dado en denominar herencia del sesgo.

En el presente Experimento 4 queremos conocer si estos resultados se observan también en estudiantes de medicina. Dado que serán los usuarios de sistemas de apoyo a la decisión en el ámbito sanitario basados en IA en el futuro, es necesario analizar cómo reaccionan ante recomendaciones erróneas cuando realizan una tarea de diagnóstico en colaboración con un modelo de IA sesgado. Con este objetivo invitamos a participar en nuestra investigación a estudiantes del Grado en Medicina de la Universidad de Deusto.

Método

Participantes

Participaron en el experimento 86 voluntarios (77.9% género femenino, 20.9% género masculino, y el 1.2% prefirió no contestar) con una edad media de 19.8 años (*SD* = 1.62, mín. = 18, máx. = 28) estudiantes de primer curso (43% del total de la muestra), y tercer curso (57% de la muestra) del Grado en Medicina de la Universidad de Deusto.

Distribuimos a los participantes al azar en los grupos con IA (*n* = 43) y sin IA (*n* = 43). El análisis de sensibilidad reveló que con un tamaño muestral de 86 participantes podíamos detectar un tamaño del efecto *f* = 0.13 o superior con una potencia del 80% para un ANOVA de medidas repetidas con factores intra e intersujetos.

Materiales y Procedimiento

En la creación del Experimento 4 seguimos el diseño y procedimiento del Experimento 2, sin apenas alteraciones más allá de una muestra de participantes diferente. La única

modificación en el procedimiento del Experimento 4 con respecto al diseño del Experimento 2 fue la utilización de las instrucciones e imágenes avisando del inicio de cada fase e indicando si la IA iba o no a estar conectada que ya utilizamos en el Experimento 3. Creímos que estas instrucciones serían útiles para delimitar el inicio de cada fase y las expectativas de los participantes para cada una de ellas.

Al igual que en experimentos previos, con el fin de analizar cambios en las tendencias de respuesta de los participantes de ambos grupos a lo largo de la tarea, decidimos dividir los 10 estímulos 40/60 de la Fase 1 en dos bloques, sin que esto afectara a los participantes: el Bloque 1 con los 5 primeros ensayos 40/60 y el Bloque 2 con los 5 últimos ensayos 40/60 que aparecieron en la serie de 60 ensayos entremezclados con estímulos de otras proporciones. Los 5 estímulos 40/60 de la Fase 2 pasaron a configurar el Bloque 3.

Deseamos destacar que en el Experimento 4 la devolución de información a los voluntarios tras finalizar su participación en la investigación cobró una mayor importancia y fue más extensa que en los anteriores experimentos de la serie. Deseábamos que la experiencia fuese lo más enriquecedora y beneficiosa posible para el alumnado del Grado en Medicina, y que les motivase a realizar una reflexión crítica sobre la aplicación de IA al ámbito sanitario. Una vez finalizaron su participación en la tarea experimental, explicamos a los estudiantes de forma detallada cuál fue el propósito de nuestro estudio y describimos el procedimiento realizado. Los alumnos también fueron informados de los resultados que encontramos en experimentos previos y los resultados que esperábamos observar en sus respuestas. Después de esta explicación, reflexionamos sobre la relevancia de la investigación de la interacción humano-IA en psicología y abrimos debate para conocer la opinión y preocupaciones de los estudiantes respecto a la utilización de herramientas de IA en medicina. La participación en nuestro estudio no solo dio a los alumnos del Grado en Medicina la oportunidad de conocer más de cerca el ámbito de la investigación, sino que, además, los

enfrentó a retos reales que encontrarán en su futuro, al menos en una situación simulada.

Se podría creer que los estudiantes de medicina serían más cautos a la hora de confiar en la recomendación de un sistema artificial para realizar una tarea que simula un diagnóstico, ya que estos estudiantes invierten años de formación para ser capaces de realizar dicha función con precisión. Al menos, puede que sean más cautos a la hora de confiar en una IA en un contexto clínico simulado en comparación a alumnos del Grado en Psicología (muestra del Experimento 1) y personas pertenecientes a población general, con diferentes contextos de formación y profesional (muestra de los Experimentos 2 y 3). Sin embargo, los estudiantes de medicina, al igual que cualquier otra persona, están cada vez más acostumbrados a interactuar con inteligencias artificiales en su vida diaria. A esto se suma que, actualmente, los escenarios de la medicina del futuro que se presentan a estudiantes colocan en una posición central a la inteligencia artificial, tecnología para la que se vislumbran posibilidades muy optimistas (Moor et al., 2023; Rajpurkar et al., 2022; Topol, 2023). Por ello, la percepción de la IA en los estudiantes de medicina puede ser bastante positiva o, al menos, creemos que otorgarán a la IA credibilidad al igual que otros grupos de población.

Nuestra predicción en el Experimento 4 fue que las respuestas de los alumnos del Grado en Medicina replicarían los resultados del Experimento 2. Esperamos que, en la Fase 1, el grupo de participantes asistidos por una IA sesgada cometa más errores en la tarea de clasificación que el grupo que no contó con ningún tipo de asistencia. En la Fase 2, esperamos que el grupo que realizó la tarea asistido por la IA en la fase anterior cometa también más errores que el grupo sin asistencia, clasificando los estímulos 40/60 en la misma dirección que estableció el sesgo de la IA, y que además clasifique de forma sesgada los estímulos ambiguos 50/50, evidenciando la generalización de ese sesgo a estímulos nuevos.

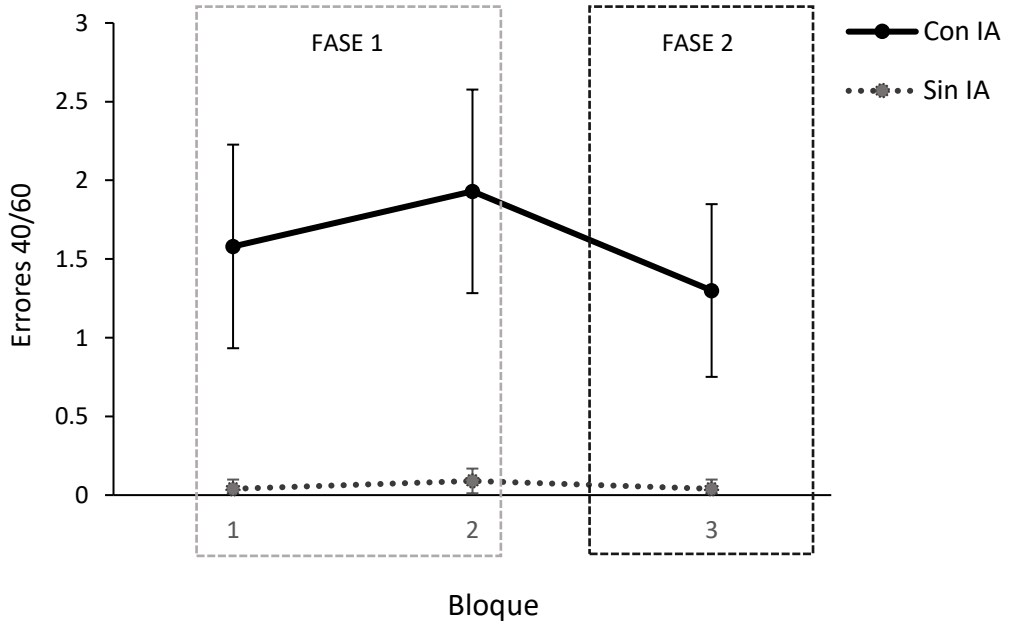
Resultados y Discusión

Como podemos ver en la Figura 14, y tal como esperábamos, el grupo que durante la

Fase 1 realizó la tarea de clasificación asistido por las recomendaciones sesgadas de la IA cometió más errores en los ensayos 40/60, tanto durante la Fase 1 como durante la Fase 2, que el grupo que realizó la tarea sin asistencia. La media de errores del grupo con IA fue mayor en todos los bloques de la tarea que la media del grupo sin asistencia. En el grupo sin IA, observamos una tendencia estable de las respuestas a lo largo de los tres bloques. Por el contrario, el comportamiento de los participantes en el grupo con IA mostró cambios en las diferentes fases de la tarea: los errores de este grupo aumentaron a lo largo de la fase con asistencia de las recomendaciones sesgadas, en los Bloques 1 y 2, para después disminuir en el Bloque 3, durante la fase de la tarea sin asistencia. A pesar de esta disminución, el grupo con IA aún cometió más errores que el grupo sin IA en el Bloque 3, si bien ambos grupos clasificaron muestras sin asistencia en esta fase de la tarea.

Figura 14

Número Medio de Errores de Clasificación en Cada uno de los Tres Bloques de Cinco Ensayos 40/60 del Experimento 4 para los Grupos con Asistencia de la IA (Con IA, n *= 43) y Sin Asistencia (Sin IA,* n *= 43).*



Nota. Las barras de error representan el 95% IC.

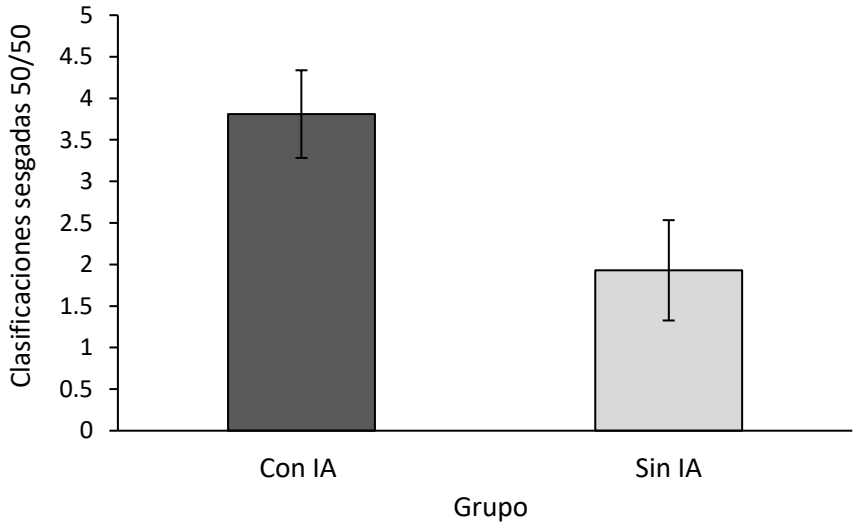
Para confirmar las impresiones descritas, realizamos un ANOVA mixto 2 (grupo: con IA, sin IA) x 3 (bloque: bloque 1, bloque 2, bloque 3) sobre la variable número de clasificaciones incorrectas de los estímulos 40/60 que reveló un efecto principal de bloque, $F(2,168) = 4.86$, $p = .009$, $\eta^2_p = .055$, un efecto principal de grupo, $F(1,84) = 28.1$, $p < .001$, $\eta^2_p = .251$, y una interacción Bloque x Grupo, $F(2,168) = 3.57$, $p = .030$, $\eta^2_p = .041$. Las comparaciones post hoc con corrección de Tukey confirmaron un aumento en los errores 40/60 del grupo con IA entre el Bloque 1 y el Bloque 2, $t(84) = -3.61$, $p_{tukey} = .007$, y un descenso en la media de errores de ese mismo grupo entre el Bloque 2 y el Bloque 3, $t(84) = 3.70$, $p_{tukey} = .005$, cuando pasaron a realizar la tarea sin asistencia de la IA. En el Bloque 3, el grupo con IA cometió más errores de clasificación que el grupo sin IA, $t(84) = 4.43$, $p_{tukey} < .001$. El grupo con IA continuó cometiendo errores en los ensayos 40/60 en la Fase 2, reproduciendo el mismo error sistemático de las recomendaciones de la IA de la fase anterior, a pesar de que en el Bloque 3 ya no estuvo guiado por recomendaciones sesgadas.

Con respecto a los estímulos novedosos y ambiguos 50/50 que los participantes encontraron en la Fase 2, nuestra predicción fue que serían clasificados de forma sesgada por los participantes del grupo con IA, mientras que serían clasificados al azar por aquellos participantes que no tuvieron asistencia de la IA en ninguna de las fases de la tarea. La comparación entre las medias de clasificaciones sesgadas de ambos grupos, que podemos ver en la Figura 15, confirma esta predicción. La media de clasificaciones sesgadas de los cinco estímulos 50/50 de la Fase 2 fue de 3.81 ($SD = 1.76$) para el grupo con IA y de 1.93 ($SD = 2.02$) para el grupo sin IA, siendo la diferencia entre las medias de ambos grupos significativa, $t(84) = 4.61$, $p < .001$, $d = 0.995$[6].

[6] Los datos no cumplieron el supuesto de normalidad (Shapiro-Wilk $W = .910$, $p < .001$), por lo que decidimos utilizar también la prueba no-paramétrica U de Mann-Whitney para confirmar la diferencia entre ambos grupos en la media de clasificaciones sesgadas de los estímulos ambiguos, $U = 450$, $p < .001$, $r_{rb} = .514$.

Figura 15

Número Medio de Clasificaciones Segadas en los Cinco Ensayos 50/50 de la Fase 2 del Experimento 4 para los Grupos con Asistencia de la IA (Con IA, n *= 43) y Sin Asistencia (Sin IA,* n *= 43).*



Nota. Las barras de error representan el 95% IC.

El 80.6% de los participantes del grupo con IA afirmó detectar errores en las recomendaciones del sistema artificial ficticio, o al menos así lo declararon en sus respuestas a las habituales preguntas que realizamos en los experimentos de la Serie 1 tras la finalización de la tarea de clasificación. Al igual que sucedió en el Experimento 1 y en el Experimento 2, observamos una correlación positiva y significativa entre el número de clasificaciones erróneas de estímulos 40/60 del grupo con IA en la Fase 1 de la tarea y cuánto consideraron a la IA de ayuda *(r* = .68, *p* < .001), cuánto consideraron que se habían basado en la recomendación de la IA para clasificar las muestras (*r* = .73, *p* < .001), cómo de acertado consideraron el consejo de la inteligencia artificial (*r* = .49, *p* = .005), y su confianza en los sistemas de inteligencia artificial en general en el ámbito de la salud (*r* = .29, *p* = .006). Cuando observamos las correlaciones entre las preguntas post experimentales y los errores en los ensayos 40/60 de la Fase 2, también encontramos que los participantes que más consideraron a la IA de ayuda (*r* = .59, *p* <

.001), se habían basado en la recomendación de la IA para clasificar las muestras (r = .57, p < .001), y la habían considerado más acertada (r = .59, p < .001), fueron también los que acumularon un mayor número de errores de clasificación de los ensayos 40/60.

Al igual que en los experimentos previos, en el presente Experimento 4 analizamos de forma exploratoria si el contrabalanceo de las instrucciones tuvo algún tipo de influencia en el número de clasificaciones incorrectas de los ensayos 40/60 realizadas por los participantes. El ANOVA 2 (grupo: con IA, sin IA) x 2 (balanceo: B1, B2) sobre los errores 40/60 cometidos por los participantes en la Fase 1 confirmó un efecto principal de grupo $F(1,82) = 26.19$, $p < .001$, $\eta^2_p = .242$. No se detectó un efecto principal del balanceo, $F(1,82) = 0.03$, $p = .843$, $\eta^2_p = .000$, ni una interacción Grupo x Balanceo, $F(1,82) = 0.008$, $p = .926$, $\eta^2_p = .000$. Las dos diferentes versiones de las instrucciones B1 y B2 en el Experimento 4 no introdujo ningún efecto extraño en los resultados de la Fase 1. El balanceo tampoco tuvo influencia alguna en los resultados observados en la Fase 2 de la tarea de clasificación, como también confirmó el ANOVA 2 (grupo) x 2 (balanceo), pues observamos un efecto principal de grupo, $F(1,82) = 19.95$, $p < .001$, $\eta^2_p = .196$, pero no un efecto principal del balanceo, $F(1,82) = 1.19$, $p = .279$, $\eta^2_p = .014$, ni una interacción Grupo x Balanceo, $F(1,82) = 1.21$, $p = .275$, $\eta^2_p = .015$.

Los resultados obtenidos en el Experimento 4, en el cual participaron estudiantes de medicina, replican los resultados observados en experimentos anteriores de la serie, realizados en población general. En la Fase 1, el grupo de voluntarios asistidos por una IA sesgada cometió más errores en la clasificación de los ensayos 40/60 que el grupo que no tuvo ningún tipo de asistencia. En la Fase 2, el grupo que durante la fase anterior tuvo la asistencia de la IA cometió también más errores que el grupo sin IA, aunque durante esta fase ambos grupos realizaron la tarea sin ningún tipo de apoyo. El grupo con IA clasificó los estímulos 40/60 en la misma dirección que estableció el sesgo de la IA, y, además, clasificó también de forma sesgada los estímulos novedosos y ambiguos 50/50. Por lo tanto, en el Experimento 4

encontramos, una vez más, evidencia de un efecto de herencia del sesgo de la IA pero en esta ocasión detectamos el efecto en un grupo de futuros profesionales de la medicina.

2.2. Discusión de la Serie Experimental 1

Las Personas Heredan Sesgos de la Inteligencia Artificial

Los resultados que observamos en los experimentos de la primera serie experimental de la presente Tesis Doctoral evidencian que las recomendaciones sesgadas de una herramienta basada en inteligencia artificial pueden sesgar las decisiones de las personas. Existe el riesgo de que los humanos reproduzcan los mismos errores sistemáticos que han mostrado previamente las recomendaciones de una inteligencia artificial, tiempo después de que haya finalizado su colaboración con ese sistema, e incluso en respuesta a estímulos novedosos e inciertos. A pesar de las graves implicaciones que podría alcanzar este efecto, de ocurrir en entornos reales, no hemos encontrado entre la literatura especializada, generada en torno a la interacción humano–IA, ninguna investigación que haya tratado de analizar de forma empírica este fenómeno antes de la publicación de nuestro trabajo de investigación (Vicente & Matute, 2023).

En los cuatro experimentos de la Serie Experimental 1, de la forma que predecíamos, los participantes asistidos por una IA sesgada cometieron más errores en la tarea que simulaba un diagnóstico médico, que aquellos participantes que realizaron la tarea sin asistencia. Los participantes que tuvieron asistencia de la IA ficticia tomaron una decisión equivocada en aquellos casos en los que la IA aportaba una sugerencia errónea, mostrando conformidad con el agente artificial. En cambio, los participantes que completaron la tarea sin asistencia no tuvieron ningún problema para clasificar exactamente esos mismos casos de forma correcta. Además, los participantes del grupo asistido por la IA cometieron los mismos errores sistemáticos que el modelo hipotético no solo en la fase de la tarea en la que tuvieron recomendaciones, sino también en una fase posterior, donde ya no contaron con este apoyo. Es decir, heredaron el sesgo de la IA cuando debían tomar decisiones por sí mismos. Nuestros

resultados son robustos, pues observamos el mismo efecto en un experimento de laboratorio realizado en condiciones controladas con una muestra de estudiantes universitarios del Grado en Psicología (Experimento 1), en experimentos online para los cuales reclutamos una muestra de población general internacional a través de la plataforma Prolific (Experimentos 2 y 3) y en una muestra de estudiantes del Grado en Medicina (Experimento 4).

Creemos que el hallazgo de que las personas pueden heredar el sesgo de la inteligencia artificial y que, además, puede darse una generalización de este sesgo a estímulos ambiguos, son los resultados más novedosos e interesantes de la presente Tesis Doctoral. Los sesgos de la inteligencia artificial tendrían el potencial de propagarse a través de las decisiones humanas. Desde hace tiempo, es bien conocido que los modelos de inteligencia artificial pueden contener sesgos y errores (Fry, 2018; O'Neil, 2019). Tal como describimos en la sección Sesgos en la Inteligencia Artificial de la Introducción, los algoritmos de aprendizaje automático son entrenados a partir de conjuntos de datos que son un registro histórico de acciones y decisiones humanas en el pasado. Si estos datos esconden patrones de errores sistemáticos, o si los datos representan de forma parcial la realidad, los modelos de IA que son entrenados a partir de dichos datos desarrollarán sesgos. En definitiva, la inteligencia artificial hereda sesgos de las personas. Nuestra presente investigación apunta a que puede ocurrir también el efecto opuesto: las personas podrían heredar sesgos de la inteligencia artificial tras una interacción prolongada con uno de estos sistemas sesgados. Si tenemos en cuenta que el comportamiento humano sesgado por los resultados de la IA puede a su vez dar retroalimentación a ese, u otro, algoritmo de aprendizaje automático, continuará el ciclo de transmisión y propagación de sesgos, haciéndose cada vez más difícil de detener.

En nuestra serie de experimentos, las etiquetas de recomendación de la IA que indicaban una categoría errónea en las muestras 40/60, influyeron claramente a los participantes, pues erraron específicamente en la clasificación de estas muestras, en

comparación al grupo sin asistencia, en la fase de la tarea con asistencia de la IA sesgada (Experimentos 1 a 4). Observamos, además, un incremento en los errores 40/60 de estos participantes a lo largo de la Fase 1 de la tarea de clasificación (evidenciado por un incremento significativo de los errores entre el Bloque 1 y el Bloque 2 de ensayos en los Experimentos 2, 3 y 4). Este dato puede sugerir que el esfuerzo depositado por los participantes en el procesamiento y monitorización de la información disminuyó, y al mismo tiempo la tendencia a confiar en la etiqueta de recomendación de la IA para realizar la clasificación aumentó, conforme avanzaba la serie de ensayos y se prolongaba la colaboración con la IA.

Las etiquetas de recomendación que se mostraban junto a la muestra de tejido a clasificar, con unos llamativos colores azul y naranja, fueron probablemente unos estímulos más salientes para la atención de los participantes que las muestras de tejido que eran, en comparación, más complejas y demandaban de un mayor esfuerzo de observación y de procesamiento de la información para realizar el juicio correcto sobre la presencia (la asignación de la muestra de tejido a la categoría positivo) o ausencia (asignación a la categoría negativo), del síndrome a detectar. Por lo tanto, las etiquetas de recomendación quizá puedan haber interferido con la evaluación y procesamiento de la información objetiva (la diferente proporción de células oscuras y claras presente en las muestras) que debían realizar los participantes (Howard et al., 2020). Esto puede haber ocurrido de dos maneras que pasamos a discutir a continuación.

En primer lugar, en la argumentación que planteamos en el párrafo anterior encontramos paralelismos con el efecto de ensombrecimiento, en el cual dos o más estímulos compiten por el acceso a los procesos de aprendizaje. Cuando dos estímulos tratan de condicionarse al mismo tiempo, aquel que pueda entrenarse más fácilmente (porque es más intenso o saliente) dificultará el aprendizaje del otro estímulo (Domjan, 2016). Dado que en nuestros experimentos ambos estímulos, la muestra de tejido a clasificar y la etiqueta de

recomendación, se presentaban de forma simultánea, de forma similar a lo que ocurre en el ensombrecimiento (Chamizo et al., 2003; Quigley & Haselgrove, 2023), el estímulo más saliente, en este caso la llamativa recomendación de la IA, pudo haber dificultado el aprendizaje de los participantes en la clasificación de muestras de tejido. Los voluntarios del grupo sin asistencia, en ausencia de otro estímulo distractor, ganaron habilidad en la asignación de muestras de tejido a las categorías según el criterio que indicaron las instrucciones al inicio del experimento. Si esto fuera cierto, habría que tener en cuenta que podría producirse este fenómeno en cualquier situación en la que una persona trabaja con una IA que simplifique una labor determinada ofreciendo recomendaciones claras, salientes y, al igual que en nuestro experimento, en principio bastante fiables.

En segundo lugar, los resultados observados en la Fase 1 de nuestra serie de experimentos pueden abordarse también desde otra perspectiva complementaria: algunos autores han utilizado argumentos próximos a las teorías de los dos sistemas de procesamiento para explicar la tendencia de las personas a mostrar conformidad con una IA que yerra (Buçinca et al., 2021; Vered et al., 2023). Recordemos que las teorías del procesamiento dual postulan la existencia de dos tipos de sistemas de procesamiento cognitivo: el Sistema 1 o Tipo 1 especializado en operaciones rápidas, automáticas, asociativas y sin esfuerzo es el responsable del uso de heurísticos o de reglas simples para resolver problemas complejos, y el Sistema 2 o Tipo 2 es el responsable de un procesamiento controlado o deliberado, analítico, flexible y que requiere esfuerzo (Kahneman, 2003). En la continua búsqueda de la estrategia que demande menor esfuerzo cognitivo (Kool & Botvinick, 2018), las personas tenderían a usar las claves que proporciona la IA como un heurístico que sustituye a un procesamiento de la información de forma analítica (Skitka et al., 1999). En nuestros experimentos, las sugerencias de categorización que proporcionaba la IA constituían una clave que permitía resolver la tarea de una forma eficiente (Kahneman, 2003), evitando a los participantes comprometerse con una exhaustiva y costosa evaluación de la corrección de cada una de las recomendaciones en

cada caso concreto (Buçinca et al., 2021; Lai & Tan, 2019). Pero esta estrategia tiene inconvenientes, y sumada a la acción de sesgos cognitivos, como por ejemplo el efecto de autoridad o el sesgo de anclaje, dificultarían a las personas detectar y corregir información errónea aportada por la IA durante la toma de decisiones en la colaboración humano–IA (Agudo et al., 2024; Baudel et al., 2020; Rastogi et al., 2022). En relación al efecto de autoridad, la IA tiende a ser percibida como una fuente experta, lo cual puede disminuir la probabilidad de que las personas contravengan sus preceptos (Araujo et al., 2020; Baudel et al., 2020; Rebitschek et al., 2021). El sesgo de anclaje es la tendencia en las personas a ajustar sus juicios en base a una pieza de información que han recibido inicialmente, de forma que este punto de partida ejerce como ancla (Tversky & Kahneman, 1974). En el momento en que una persona debe realizar una evaluación o tomar una decisión con asistencia de una IA, si la recomendación de la IA se muestra antes de que la persona emita su propio juicio independiente, esta recomendación puede actuar como ancla. Esto supondría que la decisión del humano es simplemente un ajuste a, o ratificación de, la evaluación de la IA (Agudo et al., 2024).

Las dos explicaciones que hemos propuesto hasta este punto, si bien se construyen sobre teorías diferentes, generan predicciones similares acerca del comportamiento que cabría esperar de los participantes de nuestros experimentos en la transición de la fase con asistencia de la IA a la fase sin asistencia. Además, en ambos casos esperaríamos también resultados similares en situaciones en que las personas trabajaran con inteligencias artificiales sesgadas en situaciones reales. Si la etiqueta de recomendación ensombreció el aprendizaje sobre cómo clasificar de forma correcta la muestra de tejido en la fase con asistencia de la IA, en la segunda fase sin asistencia estos participantes aún cometerían fallos en la categorización durante el inicio de su aprendizaje con las muestras de tejido (presentadas en este momento sin la interferencia de ningún otro estímulo adicional). Con el entrenamiento, en los sucesivos ensayos de la Fase 2, podrían reducir sus errores. De forma similar, si los participantes se

atuvieron a seguir automáticamente la etiqueta de recomendación de la IA durante la Fase 1 en sustitución de un procesamiento analítico y costoso de la información, tardarían en adaptarse a un procesamiento controlado en la Fase 2, cuando ya no contaban con la etiqueta como apoyo. El grupo con IA acumularía más clasificaciones incorrectas que el grupo que realizó toda la tarea sin asistencia al inicio de esta segunda fase, pero con el entrenamiento, podrían retomar por completo una forma de procesamiento analítica y reducir los errores a lo largo de la Fase 2.

Cuando los participantes de los Experimentos 2 y 4 pasaron de la fase con asistencia de la IA a la fase de la tarea de clasificación que debían completar sin ayuda, detectamos una reducción en sus errores, aunque, a pesar de la mejora, no alcanzaron la ejecución casi perfecta demostrada por los voluntarios que no tuvieron asistencia en ninguna fase de la tarea. Es decir, el grupo con IA aún clasificaba con más errores que el grupo sin IA también en la fase en la que ambos grupos no tenían asistencia. En principio, este resultado parece coherente con las predicciones que planteábamos. Sin embargo, tanto la explicación construida sobre la economía cognitiva, que se intuye en la teoría de los dos sistemas de procesamiento (Kahneman, 2003), como la explicación que se asienta en el efecto de ensombrecimiento resultan insuficientes para explicar los resultados clave de nuestros experimentos: la herencia de sesgo y la generalización del sesgo a estímulos ambiguos.

En primer lugar, ninguna de las dos argumentaciones acertaría a explicar por qué en la Fase 2 los errores de los voluntarios que tuvieron asistencia de la IA sesgada se produjeron específicamente en aquellos casos para los que la IA mostró sistemáticamente una recomendación sesgada durante la fase anterior: los estímulos 40/60. Si, por ejemplo, la dificultad radicase en un problema para cambiar rápidamente de un procesamiento automático a un procesamiento controlado, los participantes podrían haber tenido también dificultades para clasificar otro tipo de estímulos, como aquellos de proporciones 60/40 o

30/70, y sin embargo no fue así (ver Tabla 1B del Apéndice B para consultar los índices de error de cada uno de los tipos de estímulos en todos los experimentos de la Serie Experimental 1).

En segundo lugar, en el Experimento 3 destaca que no observamos un decremento en los errores de los participantes del grupo con IA en la transición de la Fase 1 a la Fase 2 de la tarea experimental. De hecho, los errores tampoco se reducen a lo largo de la Fase 2. En el Experimento 3 los participantes clasificaron un total de 10 muestras 40/60 durante la Fase 2, mientras que en los Experimentos 2 y 4 fueron cinco las muestras 40/60 a clasificar durante esta fase. Esto significa que en Experimento 3 la fase sin asistencia de la IA contó con más ensayos que nos revelaron qué trayectoria siguieron los errores de los voluntarios a lo largo de dicha fase. Los participantes que fueron asistidos por la IA sesgada replicaron los errores del sistema a lo largo de toda la fase sin asistencia, sin mostrar un atisbo de reducción o intento de corrección de dicho error sistemático. Puede construirse una explicación más sólida para este fenómeno si consideramos que los participantes adquirieron un aprendizaje a partir de la observación del comportamiento de la inteligencia artificial simulada de nuestros experimentos.

Probablemente los participantes aprendieron un nuevo criterio de clasificación basado en las sugerencias sesgadas de la IA. Este aprendizaje explicaría por qué el número medio de errores en ensayos 40/60 se mantuvo elevado de forma constante en el contexto sin asistencia para el grupo con IA en el Experimento 3 y por qué este sesgo heredado se generalizó a estímulos ambiguos nuevos (los estímulos 50/50) en el Experimento 2 y en el Experimento 3. Creemos que los participantes en nuestros experimentos adaptaron su estrategia de toma de decisiones a los resultados mostrados por la IA, aunque en ocasiones contradecían el criterio de clasificación en el que fueron instruidos al inicio de la tarea. Podría decirse que los participantes fueron entrenados por la IA, aprendiendo tanto sus aciertos como sus errores.

En un estudio de caso, Kwong et al. (2024) relatan un hallazgo en un entorno clínico

real que es similar al observado en nuestra serie de cuatro experimentos. Estos autores diseñaron un modelo de IA para predecir hidronefrosis obstructiva en niños a partir de la observación de ecografías del riñón. Para comprobar la precisión de su modelo, decidieron realizar una evaluación comparando las predicciones de la IA para las ecografías con las que realizaban los clínicos (es necesario señalar que estos expertos daban primero su evaluación y después veían el juicio de la IA, es decir, los investigadores tomaron precauciones para proteger los diagnósticos de los clínicos de la influencia de la IA). Sin embargo, durante este estudio, detectaron un cambio no intencional y significativo en la toma de decisiones del personal sanitario que había interactuado con el modelo y se había visto expuesto a sus predicciones (específicamente se detectó una reducción significativa en el número de veces que se decidía realizar una determinada prueba médica). No se identificaron modificaciones en el flujo de trabajo, en el personal, en las guías de práctica clínica o en las características de los pacientes que pudieran explicar este cambio en la toma de decisiones de los profesionales. Los autores llegaron a la conclusión de que la exposición de forma repetida a las predicciones del modelo de IA, es decir, a las etiquetas que este modelo atribuía a las imágenes a evaluar, fue la causa más probable del cambio detectado. Kwong et al. (2024) creen que los clínicos intuyeron cuales podían ser los criterios en los que se basaba la IA para proporcionar sus predicciones en la evaluación de imágenes, y se adaptaron a ellos. Es decir, adquirieron un aprendizaje, del que no eran completamente conscientes, después de exponerse de forma repetida a emparejamientos imagen-etiqueta de la IA. Este estudio de caso ofrece evidencia en un entorno real del efecto que encontramos en nuestros estudios de laboratorio (Vicente & Matute, 2023). Kwong et al. (2024) concluyen que la IA entrenó a las personas, y también nosotras señalamos al aprendizaje como el mecanismo más plausible para explicar los resultados observados en nuestros experimentos. Creemos que los participantes aprendieron el sesgo de la inteligencia artificial y lo aplicaron como quien sigue los consejos de un buen maestro cuando este ya no está presente.

Capítulo 3. Mitigar la Influencia de las Recomendaciones Sesgadas de la Inteligencia Artificial

Si existe el riesgo de que las personas hereden sesgos de la inteligencia artificial, creemos que, sin duda, la siguiente pregunta que deberíamos hacernos es cómo podríamos mitigar ese riesgo. Por ello, nos planteamos explorar en nuestro entorno controlado de laboratorio si advertir a las personas de potenciales errores en la inteligencia artificial podría atenuar la influencia de las recomendaciones sesgadas del algoritmo en las decisiones humanas y prevenir la herencia de este sesgo. Este constituyó el principal objetivo de la Serie Experimental 2 de la presente Tesis Doctoral.

3.1. Serie Experimental 2

En una serie de tres experimentos pusimos a prueba si alertar a las personas de la posible presencia de errores en la inteligencia artificial podría motivarlas a evaluar de forma crítica la información que aporta este agente.

Los tres experimentos que forman la Serie Experimental 2 tuvieron los siguientes objetivos específicos:

a. En el Experimento 5 comprobamos si proporcionar una advertencia, de forma general, acerca de que los modelos de inteligencia artificial también pueden cometer errores y contener sesgos, disminuiría la tendencia de las personas a confiar en sus recomendaciones.

b. En el Experimento 6 revelamos a los participantes cuál era el valor de fiabilidad del modelo ficticio de inteligencia artificial con el que iban a interactuar. Es decir, en las instrucciones de este experimento indicamos de forma clara el porcentaje de recomendaciones erróneas, o correctas, de la IA ficticia que ofrecía asistencia a los participantes. Esperábamos que especificar el dato de la precisión real del sistema concreto redujese la influencia del sesgo en las decisiones de los participantes, previniendo también su posterior herencia.
c. En el último experimento de la Serie Experimental 2, también analizamos, junto con la estrategia de informar sobre posibles errores, si el diferente coste del error que cometerían los participantes en caso de seguir la recomendación sesgada de la IA podría determinar la tendencia a aceptar este apoyo o rechazarlo. En un diagnóstico médico, tarea que simulamos en nuestros experimentos, la detección de un falso positivo, esto es, identificar de forma errónea a una persona como enferma cuando realmente no lo está, no tiene las mismas consecuencias que un falso negativo, es decir, fallar en detectar a una persona enferma al considerarla sana. En este escenario, las consecuencias de un falso negativo serían más graves, puesto que una persona enferma que no es correctamente diagnosticada pierde el acceso a la atención sanitaria que necesita. Nuestra predicción en el Experimento 7 fue que los voluntarios estarían más dispuestos a aceptar la recomendación errónea de la IA cuando conducía a un falso positivo, mientras que serían más reticentes a aceptarla cuando supusiese incurrir en un falso negativo.

En los tres experimentos de la Serie Experimental 2 utilizamos la misma tarea de clasificación empleada en la Serie Experimental 1. Por lo tanto, todos los experimentos que describimos en la presente Tesis Doctoral compartieron similitudes en su diseño y procedimiento. En la Tabla 4 podemos ver una descripción del diseño seguido en cada uno de los tres experimentos de la Serie Experimental 2. La principal diferencia con los experimentos

de la Serie Experimental 1 es que, en esta nueva serie, todos los participantes realizaron la tarea de clasificación con asistencia de una IA sesgada. La única diferencia entre los grupos en los experimentos de la Serie 2 fue la diferente información sobre los errores de la IA con la que partía cada uno de ellos al inicio de la tarea de clasificación, además de los diferentes tipos de error (falso positivo y falso negativo) que se manipulan en el último experimento de la serie.

Todos los experimentos de la Serie 2 fueron prerregistrados en AsPredicted. El enlace a los prerregistros correspondientes se indica en la sección Método de cada uno de los experimentos.

Tabla 4

Resumen del Diseño Seguido en los Tres Experimentos de la Serie Experimental 2

Experimento	**Grupo**	***n***	**Tarea de clasificación**			
			Fase 1	**E**	**Fase 2**	**E**
Experimento 5	Info-error	98	Asistencia IA	60	Sin asistencia	25
	No-info	99				
Experimento 6	Info-error	64	Asistencia IA	60	Sin asistencia	25
	Info-acierto	52				
	No-info	59				
Experimento 7	Info-error FP	50	Asistencia IA	60	Sin asistencia	25
	Info-error FN	50				
	No-info FP	49				
	No-info FN	49				

Nota. n, es el número final de participantes en cada grupo. E, es el número de ensayos en cada una de las fases de la tarea de clasificación. FP, falso positivo; FN, falso negativo.

Experimento 5: Advertencia de Error para Mitigar la Influencia de una IA Sesgada

En la Serie Experimental 1 obtuvimos evidencia de que las personas pueden heredar sesgos de la inteligencia artificial. En el Experimento 5 nos proponemos hallar una estrategia que pueda mitigar este efecto. En la actualidad, las herramientas de inteligencia artificial están siendo instauradas en multitud de ámbitos profesionales y, con ello, aumenta la necesidad de desarrollar estrategias que promuevan una colaboración efectiva entre humanos e inteligencias artificiales (Cabitza, Campagner, Ronzio, et al., 2023; Harbarth et al., 2024). Una colaboración efectiva se definiría por la capacidad del humano para integrar en su toma de decisiones la información aportada por la IA cuando es correcta, ética y útil, manteniendo a su vez la capacidad de rechazar información de la IA que pueda ser incorrecta o estar sesgada (Cabitza et al., 2021; Harbarth et al., 2024; Reverberi et al., 2022). Es decir, el humano no mostraría ni un exceso de confianza hacia la IA, ni tampoco mostraría una reticencia injustificada hacia su utilización, logrando un ideal de confianza completamente calibrada con la capacidad y precisión reales del modelo de IA.

Para la ayudar a las personas a calibrar su confianza en la IA se ha propuesto dar prioridad a la transparencia promoviendo, por ejemplo, el diseño de sistemas de IA que sean fácilmente explicables (Famiglini et al., 2024; Lai & Tan, 2019; Pierce et al., 2022). Esta estrategia hace referencia a proporcionar explicaciones comprensibles para personas no expertas sobre cómo funciona un determinado modelo de IA, cómo ha llegado a una determinada decisión, y cuáles son las limitaciones, o riesgo de error, en el modelo. La utilización de explicaciones arroja resultados contradictorios, pues si bien en algunos estudios han mostrado atenuar la conformidad de las personas con la IA, en otras investigaciones han aumentado la confianza en el sistema (Cabitza, Campagner, Angius, et al., 2023; Chen et al., 2023; Vered et al., 2023).

Otras intervenciones destinadas a obtener una colaboración efectiva han tenido en cuenta el momento en el que las personas tienen acceso a la recomendación de la IA. La recomendación de la IA puede ser mostrada antes de que el humano realice su propia evaluación (IA primero) o, por el contrario, la sugerencia de la IA puede ser presentada una vez el humano ha emitido su propio juicio, el cual puede modificar en función de la información aportada por la IA (humano primero). Encontramos evidencia tanto a favor de que la estrategia IA primero conduce a mejores decisiones que la opuesta (Cabitza, Campagner, Ronzio, et al., 2023), como a favor de que la estrategia humano primero es la adecuada para minimizar los errores en las decisiones en colaboración con la IA (Agudo et al., 2024).

Por lo tanto, a día de hoy, aún no está claro cuál podría ser la estrategia más adecuada para una colaboración efectiva entre humano e IA. Factores como el contexto de colaboración, el dominio de la tarea en la que trabaja el equipo humano–IA o la gravedad de las consecuencias de las decisiones tomadas por el equipo, por citar solo algunos ejemplos, modulan la confianza de las personas en la IA (Araujo et al., 2020; Ashoori & Weisz, 2019; Bogert et al., 2022; Fahnenstich et al., 2024). Entonces, cuál puede ser el mejor método para obtener una adecuada calibración de la confianza del humano en la IA dependerá también de estos factores.

En el Experimento 5, nosotras apostamos por una estrategia simple pero que puede ser efectiva para mitigar la influencia de la IA en las personas y la herencia del sesgo. Utilizamos un mensaje general y sencillo para alertar a las personas de posibles sesgos y errores en la IA. Consideramos que, si las personas tienen que realizar un esfuerzo cognitivo para comprender las explicaciones sobre el funcionamiento de la IA y su riesgo de sesgo, o para verificar si la información aportada por la IA es correcta, disminuirá la probabilidad de que esa información realmente motive una supervisión consciente de las recomendaciones de la IA (Bertrand et al., 2022; Vasconcelos et al., 2023).

Los mensajes en forma de recordatorio o advertencia sencillos han demostrado que pueden ser efectivos para inducir a las personas a decantarse por estrategias más exhaustivas de exploración de la información, al menos en el contexto de un buscador de internet (Epstein & Robertson, 2017). En el estudio de Epstein y Robertson (2017), los participantes que vieron un mensaje de alerta indicando que el orden de los resultados del buscador estaba sesgado inspeccionaron más las entradas inferiores de la página y no se conformaron con consultar los enlaces que se mostraban en primer lugar.

Creemos que nuestro procedimiento experimental nos brinda un entorno controlado idóneo para evaluar la efectividad de dar a los individuos una pieza informativa alertando sobre sesgos en la IA para mitigar una confianza excesiva en el modelo durante la toma de decisiones en un contexto alto impacto: un diagnóstico médico. Por lo tanto, en el Experimento 5 nos propusimos como objetivo comprobar de forma empírica si mostrar a los participantes un mensaje, general y sencillo, alertando de la posible presencia de errores y sesgos en la inteligencia artificial, podría reducir su tendencia a seguir acríticamente las recomendaciones sesgadas de la IA, mitigando también de esta forma la herencia y generalización del sesgo.

Método

Participantes

La muestra final estuvo formada por 197 participantes (51.8 % género femenino, 47.2% masculino, 0.5% no binario y 0.5% prefirió no responder) con una edad media de 26.6 años (*SD* = 6.49, rango de edad 18-57 años). Aunque completaron el experimento un total de 200 participantes, reclutados a través de Prolific, los datos de 3 de ellos fueron eliminados siguiendo los criterios de exclusión de datos descritos en el Experimento 1. En Prolific establecimos como filtro para la selección de participantes que hablasen inglés con fluidez y no hubiesen tomado parte en experimentos previos de nuestro laboratorio. Los participantes

recibieron una recompensa de 1 libra por realizar una tarea experimental que duró unos 12 minutos de media.

Los 197 participantes fueron distribuidos al azar en los grupos info-error (*n* = 98) y no-info (*n* = 99). El análisis de sensibilidad a posteriori reveló que con ese tamaño muestral podíamos detectar un tamaño del efecto tan pequeño como *f* = 0.09 con una potencia del 80% para un ANOVA mixto.

Materiales y Procedimiento

Los materiales y procedimiento del presente Experimento 5 fueron idénticos a los utilizados en el Experimento 4, con dos diferencias principales: (a) todos los participantes contaron con la asistencia de la IA sesgada durante la primera fase de la tarea y realizaron la segunda fase sin asistencia y (b) uno de los grupos, denominado info-error, recibió un mensaje antes de realizar la tarea de clasificación, indicando que los sistemas de IA pueden cometer errores y tener sesgos, mientras que el otro grupo, denominado no-info y que ejerció como grupo control, no recibió este tipo de información (por lo tanto fue equivalente al grupo con IA del Experimento 4). Predecimos que aquellos participantes alertados de posibles errores en la inteligencia artificial cometerían menos clasificaciones incorrectas en los ensayos 40/60, en los que recaía el error sistemático de la IA, en comparación con el grupo que no recibió dicho aviso.

La principal manipulación novedosa en el presente Experimento 5 fue la introducción de una pantalla de información, al finalizar las habituales instrucciones utilizadas en los experimentos previos, con un mensaje y una imagen, de la forma en la que se ejemplifica en la Figura 16 para el grupo no-info y en la Figura 17 para el grupo info-error. Decidimos incluir una imagen junto al mensaje porque pensamos que una clave visual motivaría a los participantes a prestar más atención al contenido de la información y podrían recordarla mientras realizaban la tarea de clasificación.

Figura 16

Captura de Pantalla que Muestra el Mensaje y la Imagen Mostrados al Grupo No-Info en las Instrucciones Previas a la Realización de la Tarea de Clasificación.

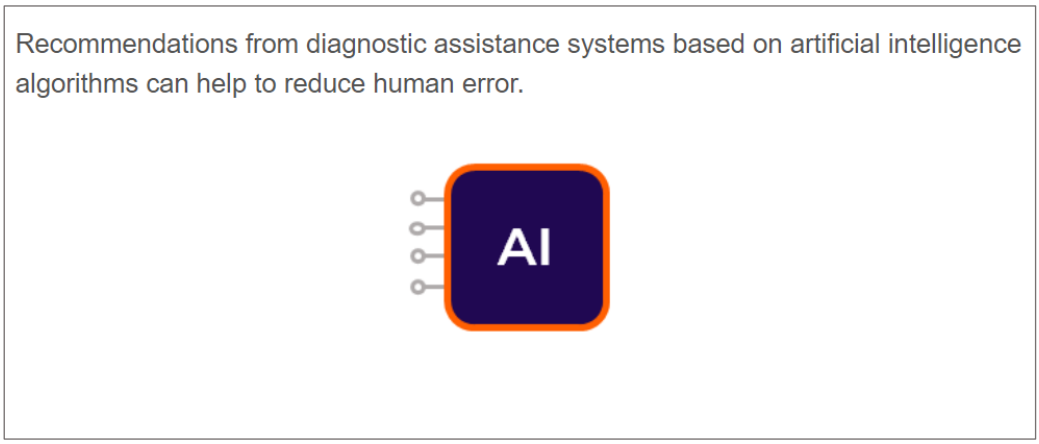
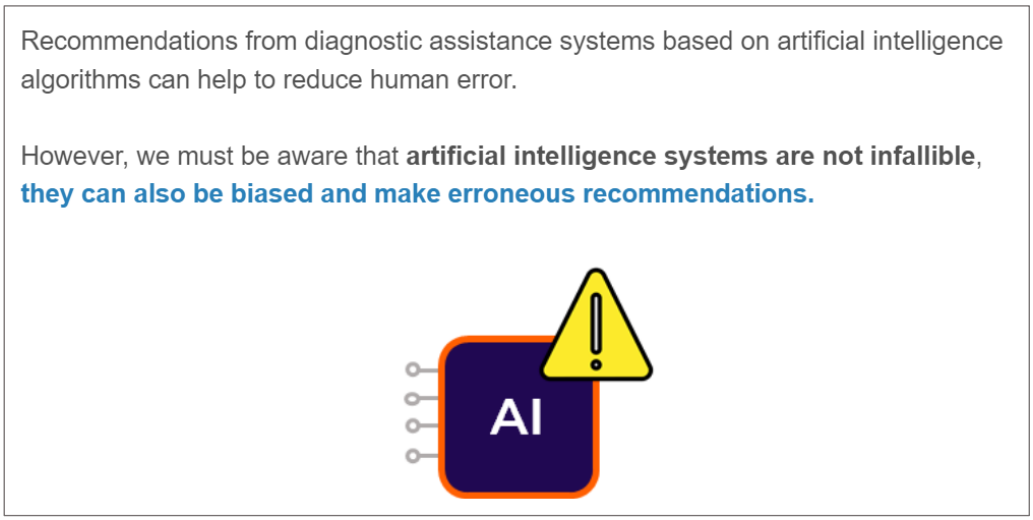


Figura 17

Captura de Pantalla que Muestra el Mensaje y la Imagen Mostrados al Grupo Info-Error en las Instrucciones Previas a la Realización de la Tarea de Clasificación.



De la misma forma que hicimos en el Experimento 4, en el presente experimento analizamos también si se producían cambios en el comportamiento de los sujetos a lo largo de la tarea de clasificación. Por ello, y tan sólo por el propósito de los análisis, optamos por dividir en tres bloques el total de ensayos 40/60 de la tarea. Los 10 ensayos 40/60 de la Fase 1 fueron

divididos en dos bloques de cinco ensayos, mientras los cinco ensayos 40/60 de la Fase 2 constituyeron el tercer bloque, para poder realizar los análisis estadísticos pertinentes, sin que ello afectase de ninguna forma a la tarea de clasificación que completaron los participantes.

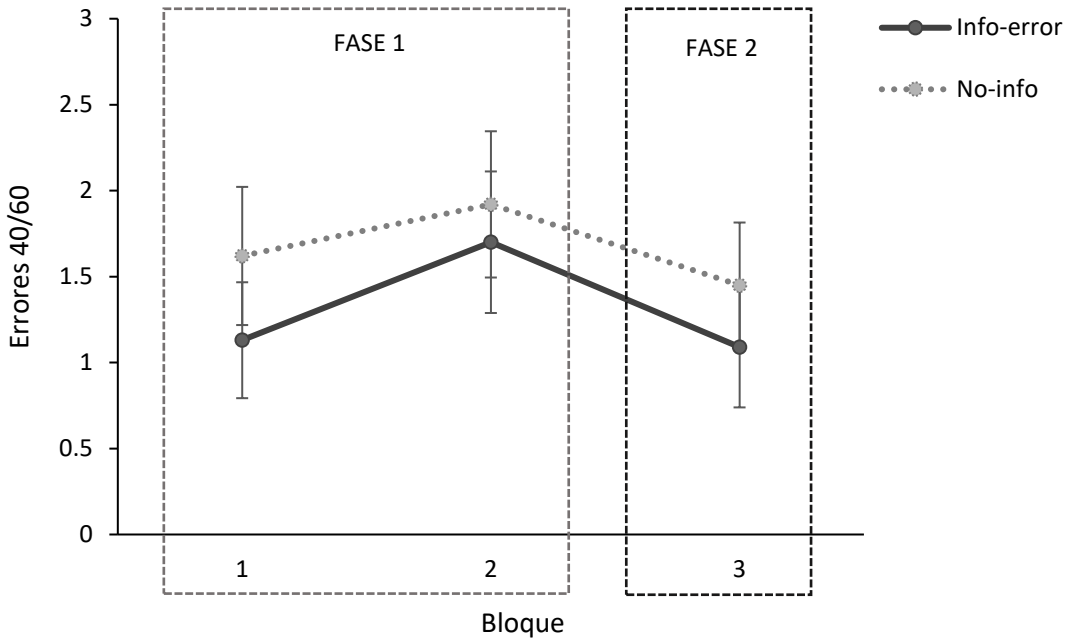
En el Experimento 5, predecíamos que, entre los participantes asistidos por una IA sesgada, aquellos que no recibieron ningún tipo de información sobre los errores del sistema cometerían más errores en la tarea de clasificación, específicamente en los ensayos 40/60 donde la IA mostraba una sugerencia errónea, en comparación a los participantes que sí recibieron información advirtiendo sobre los posibles errores de la IA. Los resultados del grupo no-info deberían replicar los observados para los grupos con IA de los experimentos de la Serie 1. Esperábamos que la media de errores del grupo no-info fuese superior a la del grupo info-error, tanto en la Fase 1 como en la Fase 2. En esta segunda fase, ambos grupos carecieron de asistencia durante la tarea de clasificación. En otras palabras, esperábamos que el aviso sobre potenciales errores en la IA demostrase también un efecto protector frente a la herencia del sesgo. El Experimento 5 fue prerregistrado en aspredicted.org https://aspredicted.org/vw6z9.pdf

Resultados y Discusión

Como puede verse en la Figura 18, la manipulación experimental, proporcionar o no una breve pieza de información alertando sobre sesgos en la IA, no introdujo diferencias entre los grupos ni durante la fase con asistencia de la IA (Fase 1) ni durante la fase sin asistencia (Fase 2) de la tarea de clasificación. La Figura 18 muestra las medias de errores de cada uno de los tres bloques de ensayos 40/60 de la tarea de clasificación para el grupo info-error y el grupo no-info. Llama la atención comprobar que ambos grupos siguieron la misma tendencia de respuesta, con un aumento de los errores entre los Bloques 1 y 2, y un descenso de los errores en el Bloque 3.

Figura 18

Número Medio de Errores de Clasificación en Cada uno de los Tres Bloques de Cinco Ensayos 40/60 del Experimento 5 para los Grupos con Información Sobre Sesgos en la IA (Info-Error, n *= 98) y Sin Información (No-Info,* n *= 99).*



Nota. Las barras de error representan el 95% IC.

El ANOVA mixto 2 (grupo: info-error, no-info) x 3 (bloque: bloque 1, bloque 2, bloque 3) confirmó nuestras impresiones iniciales, con un efecto principal de bloque, $F(2,390) = 19.62$, $p < .001$, $\eta^2_p = .091$, pero no un efecto principal de grupo, $F(1,195) = 1.91$, $p = .168$, $\eta^2_p = .010$, ni una interacción de Bloque x Grupo, $F(2,390) = 1.08$, $p = .340$, $\eta^2_p = .006$. Ambos grupos replicaron el comportamiento que habíamos detectado en los individuos asistidos por la IA sesgada en experimentos anteriores, pero, incumpliendo nuestra predicción inicial, la información sobre los errores solo produjo una ligera atenuación en las clasificaciones incorrectas de los voluntarios, sin diferencias significativas con la condición sin información en la Fase 1 de la tarea de clasificación (Bloques 1 y 2 en la Figura 18). A consecuencia de esa ausencia de diferencias entre ambos grupos en la fase con asistencia de la IA, era esperable

también una ausencia de diferencias en la segunda fase de la tarea, la fase sin asistencia (Bloque 3 de la Figura 18). Por lo tanto, parece que la información sobre posibles errores de la inteligencia artificial tampoco tuvo efecto alguno para atenuar la herencia del sesgo en la Fase 2.

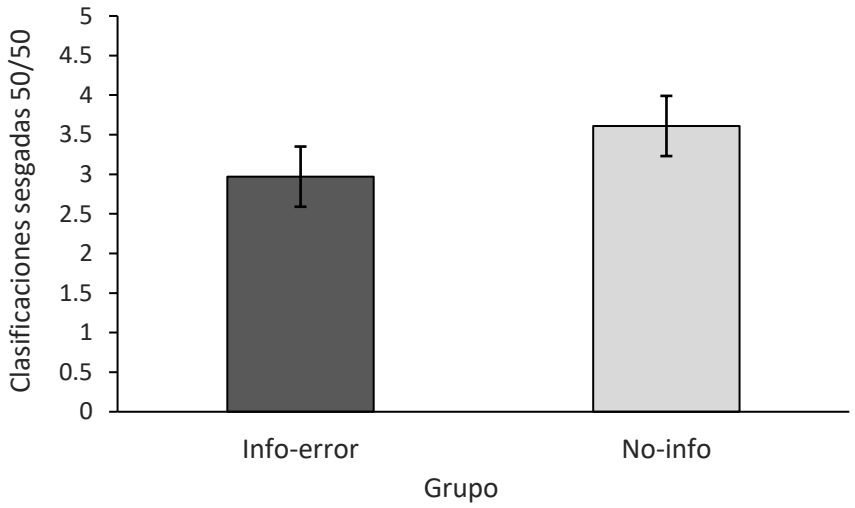
Realizamos adicionalmente un análisis de Bayes Factor, para probar la hipótesis nula y poder afirmar de esa manera que realmente no hubo un efecto de la información avisando de posibles errores (grupo: info-error, no-info) sobre el número medio de errores en ensayos 40/60 de la tarea de clasificación. El Bayes Factor en favor de la hipótesis nula fue BF_{01} = 2.16, %error = 1.04, un Bayes factor mayor que uno muestra evidencia anecdótica en favor de que el grupo (info-error, no info) efectivamente no influyó sobre el número de errores. Por lo tanto, consideramos que los datos del Experimento 5 no permiten extraer una conclusión definitiva sobre el efecto en el comportamiento de los participantes de información general advirtiendo de errores y sesgos en la IA. Es decir, aunque no encontramos diferencias significativas entre los grupos info-error y no-info, tampoco podemos afirmar de forma rotunda que advertir de forma general de errores en la IA no funciona para mitigar la tendencia de los participantes a seguir la recomendación sesgada de la IA.

Durante la fase de la tarea con asistencia de la IA sesgada, ambos grupos tendieron a clasificar de forma incorrecta las muestras de tejido 40/60 con frecuencia. Dado que era en estas muestras, de proporción células oscuras/claras 40/60, donde la IA ofrecía un consejo erróneo, interpretamos que los participantes se dejaron orientar por esa recomendación, tal como ocurrió en la Serie Experimental 1. Los errores cometidos por estos participantes en otros tipos de estímulos a lo largo de los ensayos de la tarea fueron mínimos (ver índices de error para los diferentes tipos de estímulos utilizados en la tarea de clasificación del Experimento 5 en la Tabla 2B del Apéndice B).

Cuando analizamos la tendencia a clasificar los cinco estímulos ambiguos 50/50 de la Fase 2 de forma sesgada, en la Figura 19, observamos que la media de clasificaciones sesgadas de los ensayos 50/50 en el grupo info-error fue de 2.97 (*SD* = 2.18), y en el grupo no-info fue 3.61 (*SD* = 1.93). La diferencia entre ambas medias, aunque pequeña, resultó ser significativa $t(195) = 2.17$, $p = .031$, $d = .309$[7].

Figura 19

Número Medio de Errores de Clasificación en los Cinco Ensayos 50/50 de la Fase 2 del Experimento 5 para los Grupos con Información Sobre Sesgos en la IA (Info-Error, n *= 98) y Sin Información (No-Info,* n *= 99).*



Nota. Las barras de error representan el 95% IC.

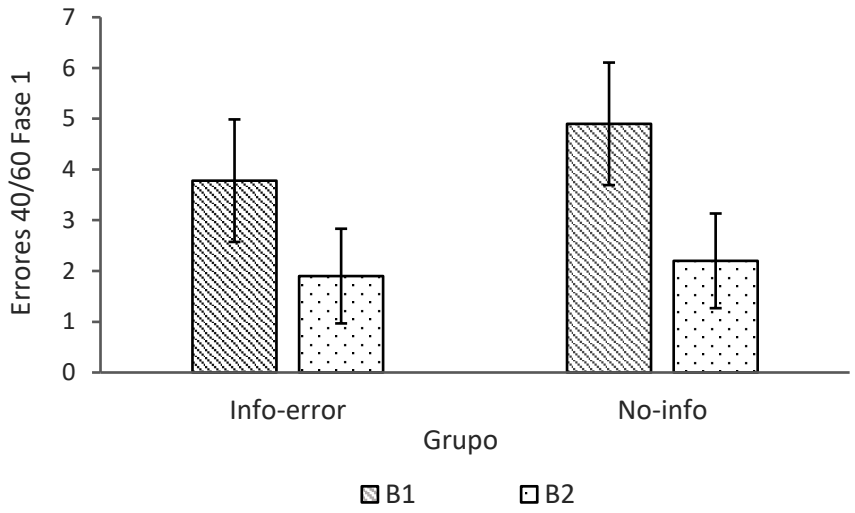
Al igual que en los experimentos previos, en el Experimento 5 decidimos explorar si podrían existir diferencias en los errores cometidos en la tarea de clasificación en función del contrabalanceo de las instrucciones. Recordemos que, tal como fue descrito en el Método del Experimento 1, las instrucciones fueron contrabalanceadas, de forma que, para la mitad de los

[7] La prueba U de Mann-Whitney confirmó esta diferencia, $U = 4083$, $p = .037$, $r_{rb} = .158$. Utilizamos esta prueba adicional porque nuestros datos no se ajustan a la normalidad (Shapiro-Wilk $W = 0.807$, $p < .001$) no homogeneidad de varianzas (Levene's $F(1,195) = 12.0$, $p < .001$).

participantes indicaban que una mayor presencia de células oscuras en la muestra de tejido correspondía con la categoría positivo (versión de las instrucciones que denominamos balanceo B1), mientras que para la otra mitad indicaban que el indicio de un positivo era una mayor presencia de células claras en la muestra a clasificar (versión denominada balanceo B2). El objetivo de este análisis era simplemente descartar algún posible efecto de las instrucciones sobre las respuestas de los participantes. De forma inesperada, en el Experimento 5 observamos diferencias en función del balanceo de las instrucciones en el número medio de clasificaciones incorrectas 40/60 durante de la Fase 1 (Figura 20).

Figura 20

Número Medio de Errores de Clasificación en los Diez Ensayos 40/60 de la Fase 1 del Experimento 5 en Función del Grupo (Info-Error, No-Info) y en Función del Balanceo (B1, B2).



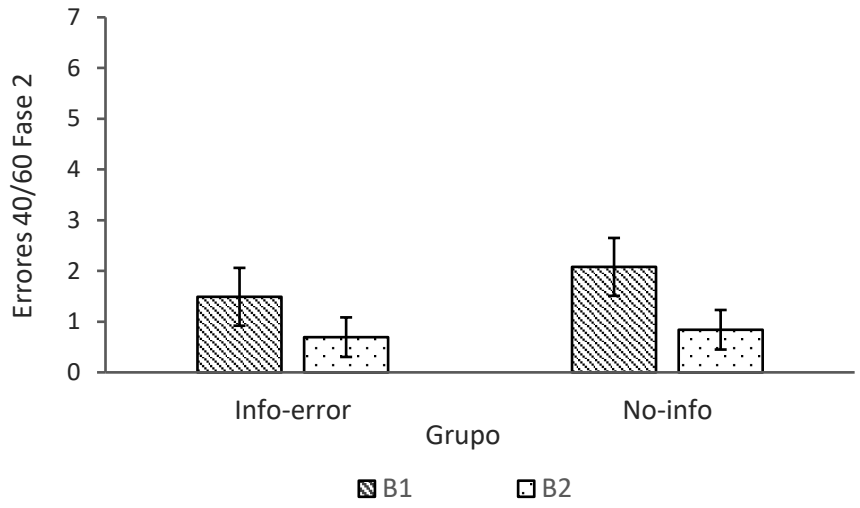
Nota. Las barras de error representan el 95% IC.

El ANOVA 2 (grupo: info error, no info) x 2 (balanceo: B1, B2) confirmó un efecto principal de balanceo, $F(1,193) = 19.240$, $p < .001$, para la Fase 1, pero no se observó ni un efecto principal de grupo, $F(1,193) = 1.865$, $p = .174$, ni una interacción Grupo x Balanceo, $F(1,193) = 0.619$, $p = .433$.

Encontramos diferencias en función del balanceo en los errores que cometieron los participantes también durante la Fase 2 (Figura 21). En experimentos previos, en los que utilizamos exactamente las mismas instrucciones y la misma tarea de clasificación, no habíamos detectado en ninguna de las fases de la tarea ni para ninguno de los grupos, con IA o sin IA, diferencias entre los participantes asignados a B1 y B2.

Figura 21

Número Medio de Errores de Clasificación en los Cinco Ensayos 40/60 de la Fase 2 del Experimento 5 en Función del Grupo (Info-Error, No-Info) y en Función del Balanceo (B1, B2).



Nota. Las barras de error representan el 95% IC.

El mismo análisis de la varianza con los errores 40/60 de la Fase 2 como variable dependiente reveló también un efecto principal de balanceo, $F(1,193) = 16.804$, $p < .001$. Tampoco en esa segunda fase detectamos un efecto principal de grupo, $F(1,193) = 2.204$, $p = .139$, ni una interacción Grupo x Balanceo, $F(1,193) = 0.804$, $p = .371$.

Si colapsamos los grupos info-error y no-info, podemos ver que el total de participantes asignados a B1 cometieron 4.73 errores de media ($SD = 4.65$) en los 10 ensayos 40/60 de la Fase 1 de la tarea de clasificación, mientras que los participantes asignados a B2

cometieron 2.32 errores de media (*SD* = 3.22). En la Fase 2, los participantes asignados a B1 cometieron 2.16 errores de media (*SD* = 2.25) en los cinco ensayos 40/60, y los participantes que fueron asignados a B2 cometieron una media de 1.07 errores (*SD* = 1.67) en esos ensayos. Los participantes que fueron asignados de forma aleatoria a B1 cometieron más errores que aquellos participantes asignados a B2, en ambas fases de la tarea experimental. Una diferencia tan grande en las respuestas de los participantes asociada al contrabalanceo de las instrucciones es un resultado sumamente extraño, que no habíamos detectado en ninguno de los anteriores experimentos de la Tesis Doctoral.

Comprobamos si podría haberse dado alguna anomalía en la investigación, durante la creación del programa de la tarea experimental o la recogida de datos. No detectamos nada, a excepción de algunos participantes aislados con respuestas extremas, que o bien tardaron un tiempo excesivo en completar la tarea de clasificación, o cometieron un número elevado de errores en el total de ensayos de la tarea, pero sin alcanzar el criterio de exclusión de datos. Sin embargo, incluso tras excluir a participantes que tardaron más de 20 minutos en realizar la tarea, que de media requirió 12 minutos, aún continuamos observando los mismos resultados. Por lo tanto, optamos por reportar los datos sin estas exclusiones.

Las preguntas que realizamos a los participantes una vez finalizada la tarea de clasificación nos informaron de que el 24.9% de los participantes declaró no haber tenido recomendaciones de la IA en ningún momento, aunque todos los participantes del Experimento 5 tuvieron la asistencia de la IA en la Fase 1. Por otro lado, un 84.8% de la muestra declaró haber detectado errores en las recomendaciones. Es curioso que, entre los participantes asignados al balanceo de las instrucciones B2, un 91.9% detectó algún error en las recomendaciones de la IA. En B1 este porcentaje se redujo al 77.6%. Si nos limitamos a evaluar estos porcentajes solo en el grupo no-info, que no fue advertido previamente de posibles errores, el 92% de los participantes de B2 detectó errores en las recomendaciones,

mientras que solo el 61.22% de los participantes de B1 afirmó detectar errores. Es posible que los participantes de B2 (en el que, recordemos, una mayor proporción de células claras en la muestra de tejido indicaba un positivo y mayor proporción de células oscuras debía clasificarse como negativo) se hayan dado cuenta más fácilmente del error, y hayan evitado seguir la recomendación de la IA. Por ahora, no podemos saber si este factor explica la diferencia observada entre los balanceos de las instrucciones.

Encontramos una correlación positiva entre los errores cometidos por los participantes en la clasificación de las muestras de tejido 40/60 durante la Fase 1 y cuánto consideraron a la IA de ayuda (r = .43, p < .001), el acierto que atribuyeron a la IA (r = .51, p < .001), y cuánto confiaban en la IA en el ámbito de la salud (r = .34, p < .001). Estas correlaciones positivas entre los errores de los participantes en la clasificación de muestras de tejido 40/60 y cuánto consideraron a la IA de ayuda (r = .41, p < .001), el acierto que le atribuyeron (r = .49, p < .001), y la confianza que depositaron en la IA en el ámbito de la salud (r = .35, p < .001), también se observaron en la Fase 2, en la que los participantes ya no contaron con la asistencia de la IA.

En el Experimento 5, el mensaje advirtiendo, de forma general, sobre sesgos en la inteligencia artificial no redujo significativamente el número de errores cometido por los participantes en la tarea experimental. Sorprendentemente, en el Experimento 5 sí encontramos diferencias significativas en los errores cometidos por los participantes en la clasificación de los estímulos 40/60 en función del contrabalanceo de las instrucciones. Los participantes asignados al azar a la versión de instrucciones B1 cometieron más errores que los participantes asignados al azar a B2. A partir de los resultados del Experimento 5 no podemos proponer una posible causa clara para este extraño efecto, que no habíamos detectado en experimentos previos. Sin embargo, sí creemos que estas diferencias en las respuestas de los participantes a consecuencia del contrabalanceo de las instrucciones han podido enmascarar el posible efecto del mensaje general advirtiendo de errores y sesgos en la IA. Por lo tanto, a

partir de los resultados obtenidos en el Experimento 5, no podemos conocer realmente si este mensaje general funciona, o no, para mitigar la conformidad de las personas con la IA. Sería conveniente poner a prueba de nuevo el efecto de este mensaje en un procedimiento sin interferencia del misterioso efecto del contrabalanceo de las instrucciones.

También debemos considerar que quizá el efecto del contrabalanceo de las instrucciones ha sido solo un resultado extraño aislado en el Experimento 5, ya que no se había manifestado en ningún otro experimento de la Tesis Doctoral con el mismo procedimiento. En el Experimento 5 algunos participantes reclutados a través de la plataforma Prolific mostraron un comportamiento que podemos describir como sospechoso, ya que tardaron un tiempo excesivo en completar la tarea experimental. Aunque Prolific es la plataforma que permite obtener datos de mejor calidad entre las disponibles para reclutamiento online (Douglas et al., 2023) quizá, en esta ocasión, algunos de los participantes no se involucraron en nuestra tarea experimental con el compromiso que habríamos deseado. Es decir, aunque no llegaron a cometer el número de errores que habría sido causa de exclusión de datos de los participantes, quizá tampoco prestaron plena atención a la tarea experimental, como releva el excesivo tiempo que invirtieron algunos de los participantes en completar la tarea (el máximo de tiempo que tardó un participante en completar la tarea fue 53 minutos, aunque podía realizarse en 12 minutos). Este tiempo excesivo podría indicar que algunos de los participantes interrumpieron la tarea de clasificación y luego la retomaron, o que estaban realizando varias tareas a la vez. En cualquier caso, si se dieron interrupciones, podrían haber contribuido especialmente al olvido de la información del mensaje de advertencia de error porque se mostró al inicio de la tarea.

En el Experimento 6 decidimos introducir un cambio y explorar otra estrategia que podría mitigar la influencia de las recomendaciones sesgadas de la IA en las decisiones de los participantes: información específica comunicando el porcentaje de aciertos, o errores, de la

IA. Este tipo de mensaje quizá pueda ser más efectivo para invitar a los participantes a ser cautos ante posibles recomendaciones erróneas de la IA, pues puede que resulte más saliente, en comparación a un mensaje más general. Para este siguiente experimento decidimos reclutar participantes voluntarios entre los estudiantes del Grado en Psicología de la Universidad de Deusto. De esta forma, también pondremos a prueba si en un procedimiento similar, pero con muestra diferente y con un método de recogida de datos más controlada, se detecta de nuevo un efecto del contrabalanceo de las instrucciones como el identificado en el Experimento 5.

Experimento 6: Efecto del Enmarcado de la Información

Cuando un profesional cuenta con el apoyo de una determinada herramienta para realizar su trabajo, como, por ejemplo, un clasificador basado en inteligencia artificial que ayuda a analizar pruebas de diagnóstico por imagen, es necesario que el profesional conozca datos exactos de la fiabilidad y riesgo de sesgo de dicha herramienta. Con esta información, el usuario puede calibrar de forma más realista qué confianza debe depositar en las sugerencias de la inteligencia artificial. Por ello, en el Experimento 6 comprobamos si proporcionar información explícita sobre la precisión de la inteligencia artificial ficticia de nuestros experimentos reduciría la vulnerabilidad de los participantes a la influencia de las recomendaciones sesgadas. Siguiendo este objetivo, un grupo de participantes recibió información explícita indicando el porcentaje de errores en las recomendaciones del sistema de inteligencia artificial que les apoyaba durante la tarea de clasificación, mientras que otro de los grupos recibió información explícita sobre el porcentaje de acierto de ese mismo sistema de IA. Finalmente, el tercer grupo, el grupo control, no recibió ningún tipo de dato sobre la fiabilidad de la IA que les asistió durante la tarea.

En psicología es bien conocido el efecto del enmarcado de la información en la toma de decisiones de las personas (Kahneman, 2003). Un mismo problema, que demande realizar una elección entre varias opciones, admite variaciones en el enmarcado de la información, por ejemplo, en la forma de presentar la probabilidad de los resultados de cada una de las acciones que pueden tomarse. Cuando las personas tratan de resolver un problema de este tipo, sus decisiones se ven claramente orientadas en función de si las consecuencias de una determinada elección son presentadas en términos de ganancias o de pérdidas. De una forma predecible, poner el énfasis en las ganancias conduce a una mayor tendencia a asumir riesgos, mientras que la focalización en las pérdidas, algo muy temido en determinados escenarios, suscita aversión al riesgo (Kahneman, 2003; McNeil et al., 1982; Tversky & Kahneman, 1981).

Inspirándonos en el efecto de enmarcado, en nuestro experimento deseábamos poner a prueba si el énfasis en el porcentaje de aciertos o en el porcentaje de errores al informar sobre los resultados de la IA provocaría diferencias en las decisiones de los participantes, pese a que todos ellos recibían en realidad la misma información sobre la fiabilidad del hipotético modelo de IA. Esperábamos que aquellos participantes informados sobre el porcentaje de error fuesen más reticentes a aceptar las recomendaciones de la IA, en comparación al grupo de voluntarios informados sobre el porcentaje de acierto. Si pensamos que el porcentaje de categorizaciones incorrectas, en el tipo de tarea que utilizamos en nuestros experimentos, lleva implícito el porcentaje de pacientes que no serían diagnosticados de forma correcta y que no recibirían tratamiento frente a una enfermedad, el error se puede conceptualizar como una pérdida. Como consecuencia, quizá los participantes informados sobre el porcentaje de error muestren aversión al riesgo que conllevaría la decisión de seguir las sugerencias de la IA sesgada. Los resultados del Experimento 6 pueden ayudarnos a entender cuál es la mejor forma de transmitir a los usuarios información sobre la fiabilidad de la IA para que les resulte fácil de interpretar.

En el Experimento 6 también era necesario comprobar si volvería detectarse un efecto del contrabalanceo de las instrucciones como en el Experimento 5. En el experimento anterior, un mensaje general advirtiendo sobre la posible presencia de sesgos en la inteligencia artificial no mitigó la tendencia de los participantes a imitar el error de la IA, en comparación al grupo sin información. Sin embargo, los resultados del Experimento 5 no fueron concluyentes debido, sobre todo, a la diferencia observada en el comportamiento de los participantes en función del contrabalanceo de las instrucciones. Por ello, en el siguiente experimento de la Serie Experimental 2 decidimos proporcionar información más específica sobre la IA ficticia que asistió a los participantes durante la tarea de clasificación y poner a prueba la efectividad de esta estrategia en una muestra diferente, en comparación al Experimento 5. La muestra del Experimento 6 estuvo formada por estudiantes universitarios que completaron la tarea

experimental en aulas de ordenadores bajo supervisión de las investigadoras. Este control de las condiciones en las que los voluntarios tomaron parte en la investigación nos permitiría dilucidar si el efecto del contrabalanceo del Experimento 5 fue solo un artefacto extraño asociado específicamente a factores asociados a la recogida de datos online en dicho experimento, o si, por el contrario, podría ser un efecto constante en los experimentos de la Serie 2.

Método

Participantes

Participaron en el experimento 179 estudiantes del Grado en Psicología de la Universidad de Deusto. Excluimos los datos de 4 participantes de la muestra siguiendo los mismos criterios de exclusión de datos utilizados en todos los experimentos de la presente Tesis Doctoral, que fueron descritos en la sección Método del Experimento 1. La muestra final estuvo compuesta por 175 participantes con una media de edad de 18.5 años (*SD* = 0.83, rango de edad 18-23 años). El 82.3% de la muestra se describió como género femenino, el 16.6% como género masculino y el 2% se autodenominó no binario. Los voluntarios tardaron 12.5 minutos de media (*SD* = 2.8) en completar la tarea experimental.

Los 175 participantes fueron distribuidos de forma aleatoria entre los grupos info-error (*n* = 64), info-acierto (*n* = 52) y no-info (*n* = 59). El análisis de sensibilidad a posteriori reveló que con dicho tamaño muestral obteníamos una potencia del 80% para detectar un tamaño del efecto pequeño, *d* = 0.10 en un ANOVA mixto 3 (grupo: info-error, info-acierto, no-info) x 3 (bloques: bloque 1, bloque 2, bloque 3).

Materiales y Procedimiento

El principal cambio en el procedimiento del Experimento 6, con respecto al seguido en el Experimento 5, fue la creación de tres grupos, info-error, info-acierto y no-info, en lugar de

dos. A excepción de este cambio, el procedimiento fue el mismo en ambos experimentos. Antes de realizar la tarea de clasificación de muestras de tejido, entre las instrucciones comunes a todos los experimentos de la serie, se mostró una pieza de información diferente a cada uno de los tres grupos del experimento: el grupo no-info recibió una breve indicación general sobre la inteligencia artificial que le ofrecería recomendaciones, pero no fue informado sobre el acierto o el error del sistema (Figura 22); para los grupos info-error e info-acierto, se añadió una frase explícita a las instrucciones generales advirtiendo sobre el 20% de error (Figura 23) o sobre el 80% de acierto (Figura 24) del sistema de IA, respectivamente.

Figura 22

Captura de Pantalla que Muestra el Mensaje y la Imagen Mostrados al Grupo No-Info en las Instrucciones Previas a la Realización de la Tarea de Clasificación.

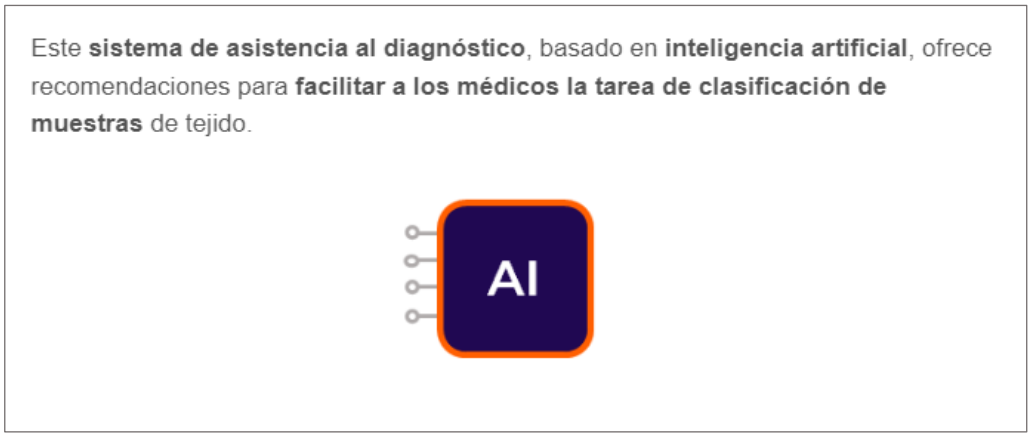


Figura 23

Captura de Pantalla que Muestra el Mensaje y la Imagen Mostrados al Grupo Info-Error en las Instrucciones Previas a la Realización de la Tarea de Clasificación.

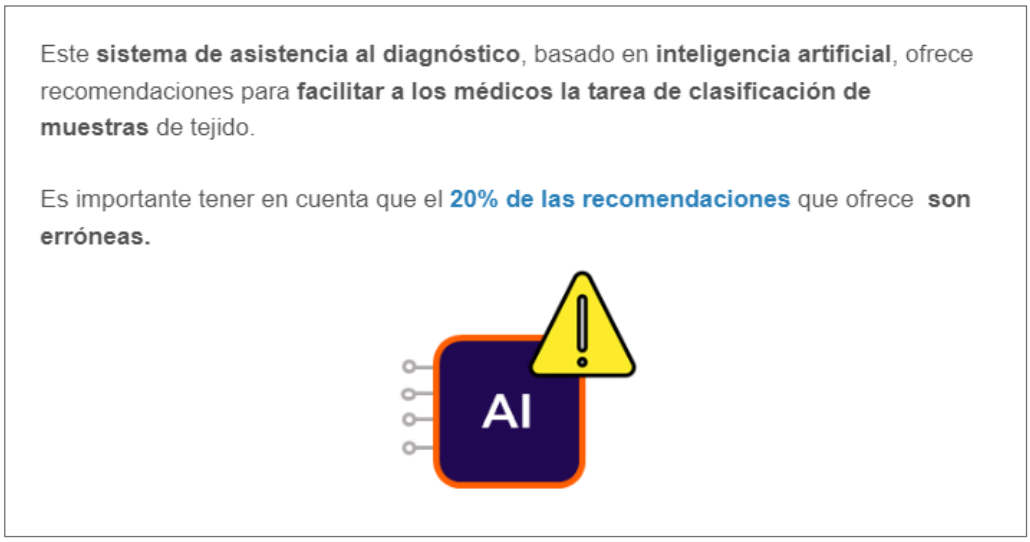
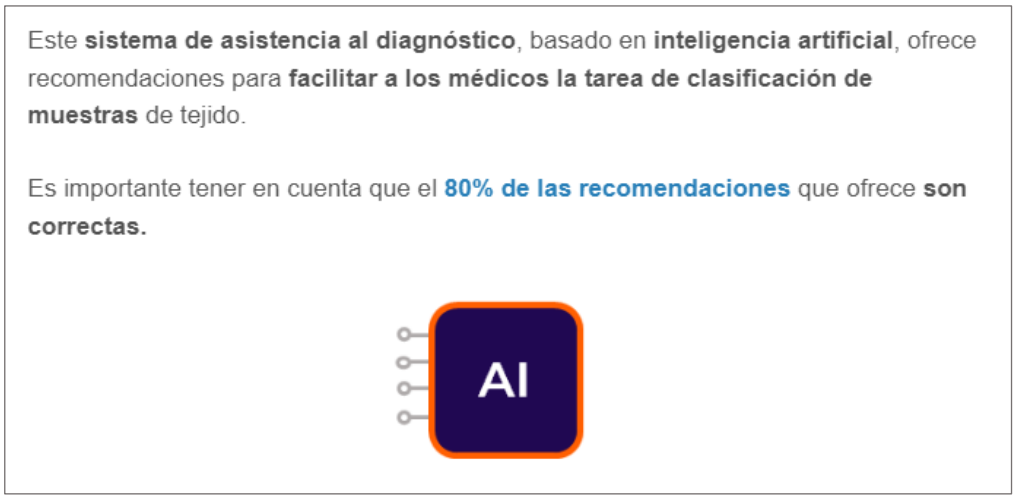


Figura 24

Captura de Pantalla que Muestra el Mensaje y la Imagen Mostrados al Grupo Info-Acierto en las Instrucciones Previas a la Realización de la Tarea de Clasificación.



El segundo cambio principal en el Experimento 6, en comparación al procedimiento seguido en el Experimento 5, fue una modificación en las preguntas que realizamos a los participantes al finalizar la tarea de clasificación. En los anteriores experimentos la primera de

las preguntas fue "¿Se ha conectado el algoritmo y has tenido recomendaciones en algún momento mientras hacías la tarea?", después de esta primera pregunta, y en pantalla a parte, en el presente Experimento 6 añadimos una nueva escala que obligaba a los participantes a realizar una estimación, en un rango de 0 a 100, del porcentaje de recomendaciones ofrecidas por la IA que habían percibido como erróneas. Esta estimación sustituirá a las preguntas "¿Has detectado algún error en las recomendaciones de la IA? (Sí/No)" y "¿Cómo de acertado ha estado el algoritmo de inteligencia artificial en la clasificación de las muestras de tejido? (1 – 9)" utilizadas en experimentos previos. Con la escala que añadimos en el Experimento 6 pretendíamos conseguir una evaluación más sensible de la capacidad de nuestros participantes para estimar el porcentaje de error real de la IA. En lugar de preguntar tan sólo cómo de acertado percibieron al algoritmo de IA, con un rango de respuestas acotado a los límites muy acertado o poco acertado, en el Experimento 6 queríamos saber si los participantes estimarían de forma correcta un error próximo al 20% (la cifra de error real) cuando se enfrentasen a indicar un dato en una escala de 0 a 100, y si existirían diferencias entre los grupos en dicha estimación.

A continuación, tras estimar el porcentaje de error, los participantes respondieron al resto de preguntas que hemos utilizado en otros experimentos, a excepción de las dos eliminadas, siempre en el mismo orden. A los participantes asignados a los grupos info-error e info-acierto se les preguntó también qué porcentaje de recomendaciones erróneas en la IA indicaron las instrucciones al inicio de la tarea. Esta pregunta se presentó al final, cuando los participantes ya habían respondido al resto de preguntas, y en una pantalla separada, con el objetivo de tratar de comprobar si los participantes habían prestado atención a las instrucciones y recordaban el porcentaje de error, o acierto, que leyeron al inicio del experimento. En resumen, se requirió de los participantes del Experimento 6 que estimasen el porcentaje de error de la IA en dos ocasiones. Primero estimaron el error percibido u observado en las recomendaciones de la IA a lo largo de la serie de ensayos, y después

indicaron el porcentaje declarado que recordaban haber leído en las instrucciones antes del inicio de la tarea.

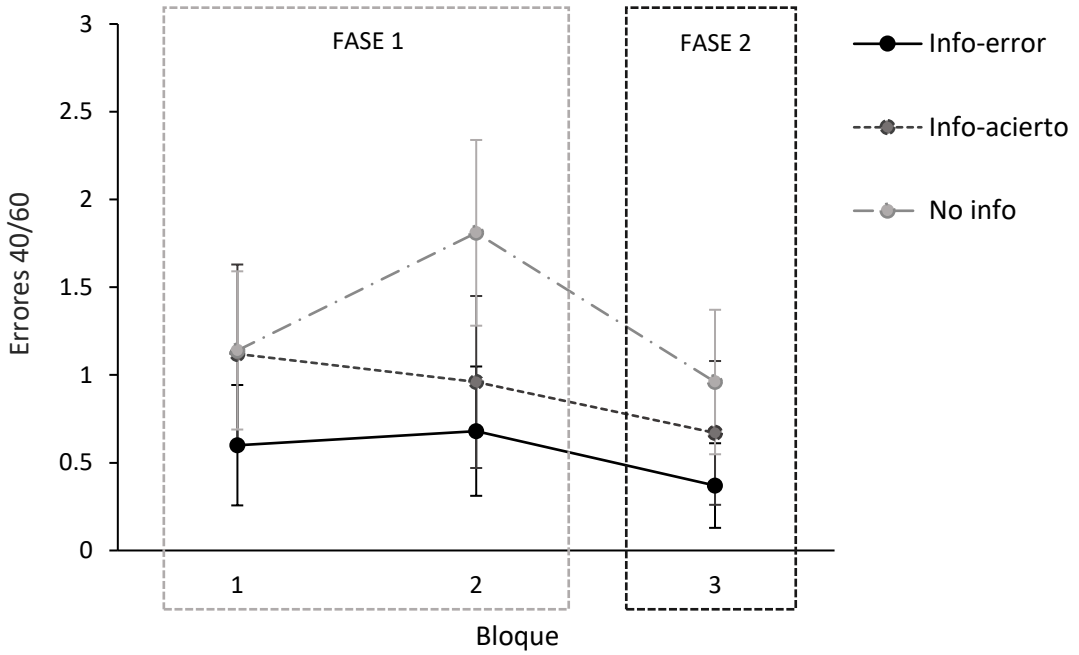
Todos los participantes del Experimento 6 realizaron la clasificación de muestras asistidos por la recomendación de la IA sesgada durante la Fase 1, mientras que en la Fase 2 continuaron realizando la misma tarea sin asistencia. Esperábamos que el grupo al que se le indicó que el 20% de las recomendaciones de la IA eran erróneas cometiese menos errores en la clasificación de los estímulos 40/60 (el estímulo en el que la recomendación de la IA se equivocaba) en ambas fases en comparación al grupo que no recibió ningún tipo de información sobre el error. En cuanto al grupo que recibió la información sobre fiabilidad con énfasis en los aciertos de la IA, especificando que el 80% de las recomendaciones eran correctas, esperábamos que cometiesen un número de errores en la tarea de clasificación menor al del grupo no-info, pero mayor al número de errores cometidos por el grupo info-error. Creíamos que enfatizar el porcentaje de error de la IA haría a los participantes menos proclives a aceptar acríticamente sus recomendaciones y serían más capaces de detectar el sesgo. En el caso de los participantes que recibieron la información complementaria, probablemente considerarían que el 80% de acierto en la IA era un dato suficientemente bueno y que podían confiar en sus recomendaciones en la mayoría de los casos, aunque el sistema no tuviera una ejecución perfecta. El Experimento 6 fue prerregistrado en Aspredicted.org https://aspredicted.org/xg7h5.pdf

Resultados y Discusión

La Figura 25 muestra el número medio de errores de clasificación en ensayos 40/60 en cada uno de los grupos a lo largo de toda la tarea experimental.

Figura 25

Número Medio de Errores de Clasificación en Cada uno de los Tres Bloques de Cinco Ensayos 40/60 del Experimento 6 para los Grupos Info-Error (n = 64), Info-Acierto (n = 52) y No-Info (n = 59).



Nota. Las barras de error representan el 95% IC.

Como podemos observar en la Figura 25, el grupo info-error cometió menos errores, de media, en la clasificación de los estímulos 40/60 que los otros dos grupos, en cada uno de los tres bloques de la tarea. El grupo que acumuló un mayor número de clasificaciones incorrectas de los ensayos 40/60 fue el grupo no-info, tal como esperábamos. En este grupo, que no recibió ningún tipo de información sobre la precisión de la IA, observamos un incremento en las clasificaciones erróneas entre el primer bloque de cinco estímulos 40/60 y el segundo bloque de estímulos 40/60 de la fase con asistencia de la IA (Fase 1). Es en este segundo bloque donde la diferencia en los errores entre los grupos info-error y no-info fue más clara. En el grupo no-info, también se observó un decremento en los errores en el bloque de estímulos 40/60 de la Fase2, sin asistencia de la IA. Como consecuencia, las diferencias en

clasificaciones erróneas entre los grupos se atenuaron en la última fase. El grupo no-info replicó el patrón de respuestas propio de los grupos con asistencia de la IA de la Serie Experimental 1. Sin embargo, los grupos info-error e info-acierto parece que sí estuvieron más alerta ante posibles errores de la IA, de modo que el número medio de clasificaciones erróneas que cometieron estos grupos fue menor en comparación al grupo no-info. Además, los grupos info-error e info-acierto no tuvieron cambios tan claros a lo largo de la tarea, es decir, aunque el número medio de errores disminuyó un poco en la última fase, las respuestas de estos dos grupos siguieron un patrón estable a lo largo de los tres bloques.

Para comprobar estas impresiones realizamos un ANOVA mixto 3 (grupo: info-error, info-acierto, no-info) x 3 (bloque: bloque 1, bloque 2, bloque 3) que mostró un efecto principal de bloque, $F(2,344) = 14.27$, $p < .001$, $\eta^2_p = .077$, un efecto principal de grupo, $F(2,172) = 3.82$, $p = .024$, $\eta^2_p = .043$, e interacción Bloque x Grupo, $F(4,344) = 4.05$, $p = .003$, $\eta^2_p = .045$. Las comparaciones post hoc con corrección de Tukey confirmaron una diferencia significativa en el Bloque 2 entre las medias de clasificaciones erróneas de los grupos no-info e info-error, $t(172) = 3.14$, $p_{tukey} = .022$. En este segundo bloque, no hubo diferencias significativas entre los grupos no-info e info-acierto, $t(172) = 2.45$, $p_{tukey} = .262$, ni entre los grupos info-error e info-acierto, $t(172) = -0.80$, $p_{tukey} = .997$. Además, las comparaciones post hoc no revelaron diferencias entre las medias de clasificaciones erróneas de los grupos en el Bloque 1, ni entre los grupos no-info e info-error, $t(172) = 1.72$, $p_{tukey} = .073$, ni para los grupos no-info e info-acierto, $t(172) = 0.06$, $p_{tukey} = 1.00$, ni mucho menos para los grupos info-error e info-acierto, $t(172) = -1.60$, $p_{tukey} = .802$. Tampoco se observaron diferencias significativas entre los grupos en el Bloque 3: no entre los grupos no-info e info-error, $t(172) = 2.23$, $p_{tukey} = .331$, tampoco entre no-info e info-acierto, $t(172) = 1.09$, $p_{tukey} = .974$, ni entre los grupos info-error e info-acierto, $t(172) = -1.13$, $p_{tukey} = .968$.

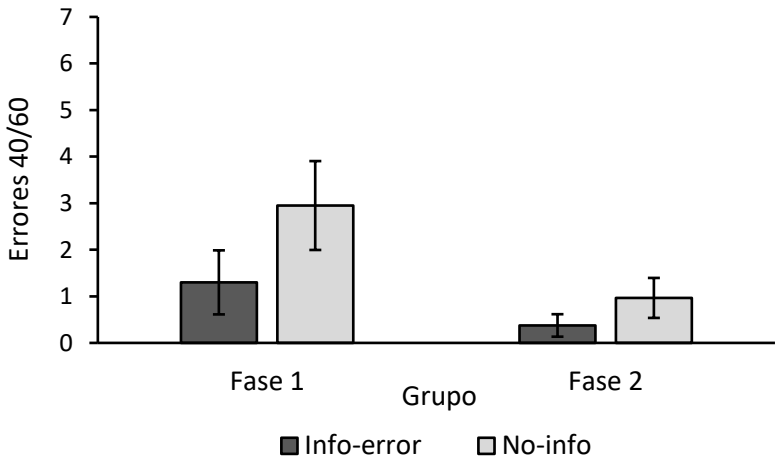
Por otro lado, las comparaciones post hoc sí confirmaron la tendencia observada en la

Figura 25 para el grupo no-info, con un incremento significativo en los errores 40/60 entre el Bloque 1 y el Bloque 2, $t(172) = -4.68$, $p_{tukey} < .001$, y una reducción significativa en sus errores entre el Bloque 2 y el Bloque 3, $t(172) = 4.86$, $p_{tukey} < .001$. Esta tendencia no se observó ni para el grupo info-error, ni para el grupo info-acierto. Es decir, efectivamente el grupo no-info replicó el patrón de respuesta habitual en los grupos con IA de la Serie Experimental 1, mientras que los otros dos grupos tuvieron menos errores y además ausencia de diferencias entre las medias de errores en los diferentes bloques a lo largo de la tarea.

De forma exploratoria, decidimos analizar posibles diferencias en los errores 40/60 de la Fase 1 y de la Fase 2 solo para los grupos info-error y no-info. Las pruebas t de Student para muestras independientes mostró diferencias significativas entre estos dos grupos en la Fase 1, $t(121) = 2.79$, $p = .006$, $d = .503$ y en la Fase 2, $t(121) = 2.40$, $p = .018$, $d = .434$. La comparativa entre las medias de errores 40/60 de ambos grupos en las dos fases de la tarea se pueden ver en la Figura 26.

Figura 26

*Número Medio de Errores de Clasificación en los Diez Ensayos 40/60 de la Fase 1 y en los Cinco Ensayos 40/60 de la Fase 2 del Experimento 6 para los Grupos Info-Error (*n *= 64) y No-Info (*n *= 59).*



Nota. Las barras de error indican el intervalo de confianza de la media al 95%.

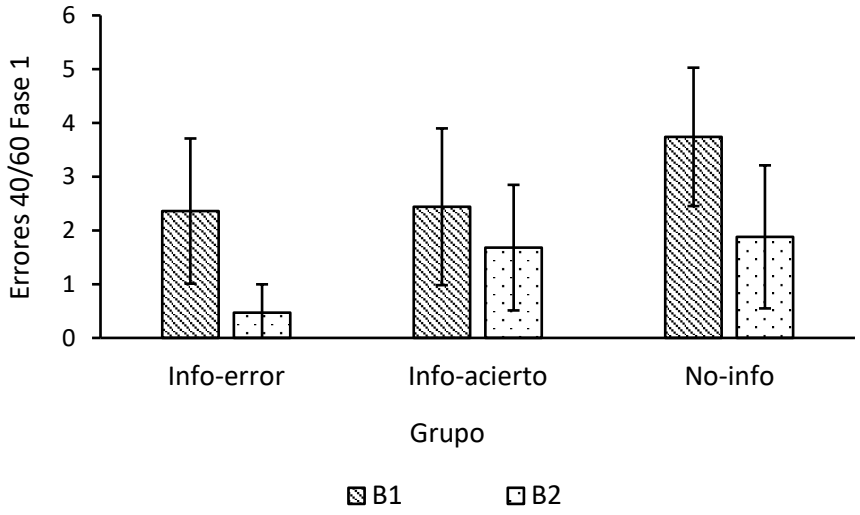
Realizamos también un análisis de varianza para explorar diferencias entre los grupos en la tendencia a clasificar los estímulos ambiguos 50/50 en la misma dirección que el sesgo de la recomendación de la IA, durante la fase de la tarea en la que los participantes ya no tuvieron recomendaciones. El ANOVA simple reveló que no hubo diferencias en la tendencia a clasificar de forma sesgada los estímulos ambiguos en función del grupo, $F(2,172) = .935$, $p = .395$, $\eta^2_p = .011$.

Antes de continuar con la descripción de los resultados, decidimos detenernos a analizar un posible efecto del contrabalanceo de las instrucciones también en el Experimento 6. En el Experimento 5, los participantes asignados al balanceo B1 de las instrucciones cometieron de media más errores 40/60 en la tarea de clasificación, tanto en la fase asistida por la IA como en la fase sin asistencia, que los participantes asignados al balanceo B2, en ambos grupos experimentales, info-error y no-info. Recordemos que en B1, una mayor proporción de células oscuras en la muestra de tejido indicaba la presencia de enfermedad y por tanto su clasificación como positivo, mientras que una mayor proporción de células claras significaba que el paciente del que provenía dicha muestra no estaba enfermo, y la muestra de tejido se clasificaba como negativo. En B2, era una mayor proporción de células claras en la muestra de tejido la señal que indicaba la presencia de enfermedad, y por tanto su clasificación como positivo, mientras que una mayor proporción de células oscuras era señal de ausencia de enfermedad y correspondía clasificar esa muestra como negativo.

En el Experimento 6, también los participantes que fueron asignados aleatoriamente a B1 cometieron más errores de clasificación que aquellos asignados a B2 durante la Fase 1 de la tarea, impresión derivada de la observación de la Figura 27.

Figura 27

Número Medio de Errores de Clasificación en los Diez Ensayos 40/60 de la Fase 1 del Experimento 6 en Función del Grupo (Info-Error, Info-Acierto y No-Info) y en Función del Balanceo (B1, B2).

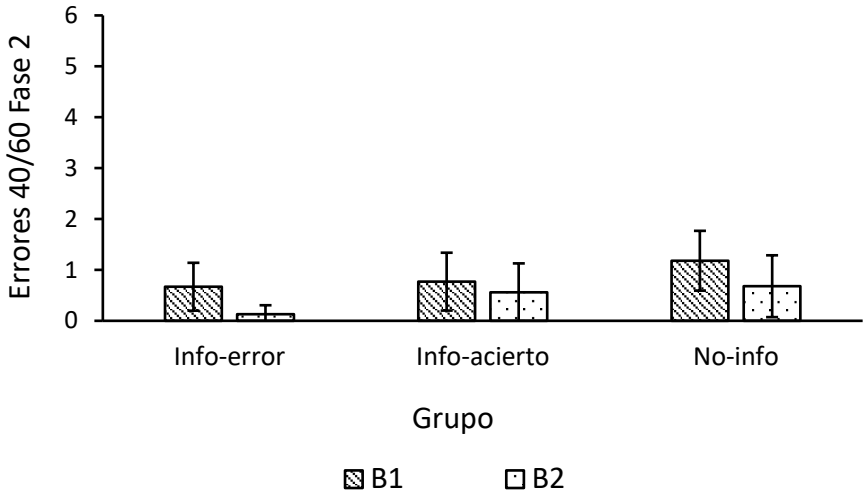


Nota. Las barras de error indican el intervalo de confianza de la media al 95%.

Realizamos un ANOVA 3 (grupo: info-error, info-acierto, no-info) x 2 (balanceo: B1, B2) con el número de errores de clasificación en ensayos 40/60 de la Fase 1 como variable dependiente, que confirmó un efecto principal de balanceo, $F(1,169) = 9.061$, $p = .003$, $\eta^2_p = .051$, pero no se observó ni un efecto principal de grupo, $F(1,169) = 2.738$, $p = .068$, $\eta^2_p = .031$, ni una interacción Grupo x Balanceo, $F(1,169) = 0.521$, $p = .595$, $\eta^2_p = .006$. Realizamos el mismo ANOVA 3 x 2 sobre el número de clasificaciones incorrectas en ensayos 40/60 de la Fase 2, que no reveló un efecto principal de balanceo, $F(1,169) = 3.817$, $p = .052$, $\eta^2_p = .022$, tampoco un efecto principal de grupo, $F(1,169) = 2.071$, $p = .129$, $\eta^2_p = .024$, ni una interacción Grupo x Balanceo, $F(1,169) = 0.213$, $p = .808$, $\eta^2_p = .003$. Por lo tanto, en la Fase 2 no se detectaron diferencias significativas en función del contrabalanceo de las instrucciones, como se muestra en la Figura 28.

Figura 28

Número Medio de Errores de Clasificación en los Diez Ensayos 40/60 de la Fase 2 del Experimento 6 en Función del Grupo (Info-Error, Info-Acierto y No-Info) y en Función del Balanceo (B1, B2).



Nota. Las barras de error indican el intervalo de confianza de la media al 95%.

En el Experimento 6, los participantes estimaron el porcentaje de error que habían percibido en las recomendaciones de la IA en una escala de 0 a 100. Se observaron diferencias entre los grupos en los porcentajes de error percibido, $F(2,172) = 7.48$, $p < .001$, $\eta^2_p = .080$, el grupo info-acierto percibió, de media, un 35% de error (*SD* = 23.5) en las recomendaciones de la IA, el grupo info-error percibió un 27.3% de error (*SD* = 14.1) y el grupo no-info curiosamente fue el más preciso en su estimación con un 22.3% de media (*SD* = 13.4).

La media en la estimación del porcentaje de error que los participantes recordaban haber leído en las instrucciones de la tarea fue de un 21.2% (SD = 10.5) para el grupo info-error, y de un 31.1% (*SD* = 24.4) para el grupo info-acierto. Algunos de los participantes del grupo info-acierto respondieron tanto a la pregunta sobre el porcentaje de error percibido, como a la pregunta sobre el porcentaje de error recordado (el indicado en las instrucciones al inicio de la tarea), con cifras cercanas a 80, lo cual explica por qué las estimaciones son más

altas para este grupo. Interpretamos que algunos participantes del grupo info-acierto recordaron haber leído dicha cifra en las instrucciones, sin recordar correctamente que se refería al porcentaje de aciertos, y este número actuó probablemente como anclaje elevando sus estimaciones del porcentaje de error en las recomendaciones. Las instrucciones especificaban que el 80% se refería a la proporción de recomendaciones correctas de la IA, no a las recomendaciones erróneas, pero parece que la interpretación, o bien de estas instrucciones, o bien de las preguntas referidas a los porcentajes de error, fue difícil para algunos participantes, o simplemente no prestaron la debida atención.[8]

De forma similar a lo observado en experimentos previos, en el Experimento 6 también encontramos correlaciones positivas entre el número de errores 40/60 cometidos en la Fase 1 de la tarea de clasificación y cuánto consideraron los participantes a la IA de ayuda (r = .46, p < .001), cuánto se basaron en la recomendación de la IA para dar sus respuestas (r = .58, p < .001), y su confianza en general en la inteligencia artificial en el ámbito de la salud (r = .37, p < .001). Las mismas correlaciones fueron observadas también en la Fase 2, de forma que los participantes que más se equivocaron en ensayos 40/60 fueron también quienes más consideraron a la IA de ayuda (r = .37, p < .001), se basaron más en la recomendación de la IA para clasificar (r = 47, p < .001), y confiaban más en las posibilidades de la IA aplicada al ámbito de la salud (r = .29, p < .001).

En el Experimento 6, aquellos individuos a quienes revelamos que el 20% de las recomendaciones de la IA eran erróneas tendieron a cometer menos errores en comparación a los participantes sin información alguna sobre la precisión de la IA. El comportamiento del grupo info-acierto, sin embargo, no manifestó diferencias significativas con respecto a los dos

[8] Si algunos participantes del grupo info-acierto encontraron problemática la interpretación del porcentaje dado en las instrucciones, este fenómeno podría haber introducido un efecto extraño en los resultados. Repetimos por tanto todos los análisis estadísticos del Experimento 6 excluyendo a los participantes que estimaron un error percibido y un error recordado mayores de 60%. No observamos ningún cambio en los resultados principales por lo que optamos por reportar los datos sin esta exclusión.

otros grupos. De nuevo, en el Experimento 6 detectamos un efecto del contrabalanceo de las instrucciones en el comportamiento de los participantes que quizá pudo haber confundido de alguna forma el efecto de nuestra manipulación experimental, como ya sucedió en el Experimento 5. Dado que identificamos el mismo efecto derivado del contrabalanceo de las instrucciones en dos experimentos distintos, con diferente muestra (participantes de Prolific y estudiantes universitarios) confirmamos que no fue un artefacto extraño aislado en el Experimento 5, si no que estamos ante un efecto constante asociado al procedimiento seguido en la Serie Experimental 2 y debemos indagar cuales pueden ser las causas.

Ante la inesperada diferencia entre los balanceos detectada en los Experimentos 5 y 6 nos preguntamos si, quizá, en B2 los participantes fueron más capaces de detectar el error sistemático de las recomendaciones. Si observamos los porcentajes de error percibido solo para el grupo no-info del Experimento 6 (grupo al que no se le indicó en las instrucciones una cifra de error o acierto que influyese su percepción del error) encontramos que los participantes asignados a B1 estimaron un error percibido del 18.1% (*SD* = 10.9) y los participantes de B2 estimaron un error del 28.2% (*SD* = 14.5). Los voluntarios del balanceo de instrucciones B2 quizá tendieron a detectar más los errores en las recomendaciones de la IA a lo largo de la tarea de clasificación. Quizá la diferencia entre los balanceos pudo deberse a algún efecto perceptivo o a que las instrucciones resultaron más intuitivas en un balanceo que en otro. O quizá la diferencia se debió a que las consecuencias del error no eran las mismas en B1 y en B2.

En los experimentos de la Serie Experimental 1 no detectamos diferencias en los errores que cometieron los participantes en la clasificación de muestras 40/60 en función del balanceo. Si la diferencia entre los balanceos observada en los Experimentos 5 y 6 se debiese a un factor perceptivo, o bien a que la clasificación fue más fácil de realizar bajo uno de los criterios, B1 o B2, creemos que deberíamos haber encontrado algún indicio de este efecto ya

en los primeros experimentos que describimos en esta Tesis Doctoral. Dado que no fue así, priorizamos la hipótesis de que quizá fue el diferente coste de las decisiones incorrectas en cada balanceo lo que ocasionó las diferencias.

Conforme al criterio de clasificación del balanceo B1, el error en la recomendación de la IA consistió en atribuir una etiqueta de positivo a una muestra de tejido que realmente debía ser clasificada como negativo, esto es, la IA cometía un falso positivo. En B2, sin embargo, el error de la IA consistió en sugerir la categoría negativo para una muestra de tejido que debía ser clasificada como positivo, es decir, un falso negativo. Quizá los participantes, ante la duda causada por la contradicción entre la información de la muestra y la etiqueta de recomendación, optaron por ser cautos y confiar menos en la recomendación de la IA cuando se arriesgaban a un falso negativo, en el balanceo B2. Optamos por explorar esta interesante hipótesis en el siguiente experimento de la serie.

Experimento 7: Efecto de la Advertencia de Error y Coste del Error

En el último experimento de la Serie Experimental 2 nos proponemos tratar de dar respuesta a dos incógnitas derivadas de los experimentos previos: (a) ¿Un mensaje que advierte de forma general de errores y sesgos en la IA realmente no funciona para mitigar la influencia de la IA en las decisiones de los participantes o no detectamos su efecto en el Experimento 5 por interferencia del contrabalanceo de las instrucciones? y (b) ¿Puede el diferente coste del error asociado a cada versión de instrucciones, B1 y B2, ser la causa de las diferencias observadas entre los participantes en función del contrabalanceo de las instrucciones en los Experimentos 5 y 6? El Experimento 7 tratará de dar respuesta a ambas preguntas y, con ello, proporcionar un cierre coherente a la Serie Experimental 2.

En el Experimento 5 no encontramos un efecto en el comportamiento de los participantes de un mensaje general advirtiendo sobre posibles errores y sesgos en la IA, pero probablemente fue debido al extraño efecto del contrabalanceo de instrucciones detectado en dicho experimento. Por ello, en el Experimento 7 decidimos poner de nuevo a prueba esa misma intervención para confirmar si podría mitigar la influencia de la IA en las decisiones de las personas, pero, esta vez, intentando neutralizar un posible efecto del contrabalanceo de las instrucciones asociado a los estímulos 40/60. En el Experimento 6, un mensaje específico enfatizando el porcentaje de errores en las recomendaciones de la IA sí fue efectivo para ayudar a los participantes a identificar y evitar seguir las recomendaciones sesgadas, pero la idea que intentamos defender en el Experimento 7 es que incluso un formato de información más general podría también tener ese efecto (y no lo pudimos observar previamente por interferencia del contrabalanceo).

Con este propósito, en el Experimento 7 modificamos el sesgo de la IA ficticia, que en lugar de una recomendación errónea siempre en estímulos 40/60, como en experimentos

anteriores, fue una recomendación equivocada siempre, y únicamente, para los estímulos 60/40. Con este sencillo cambio podríamos obtener más información sobre las posibles causas de las diferencias entre las dos versiones de instrucciones resultantes del contrabalanceo, B1 y B2, en los anteriores experimentos de la Serie Experimental 2. Por ejemplo, si la causa de estas diferencias se debiese a algún tipo de efecto perceptivo asociado a los estímulos 40/60, no se manifestaría asociado a los estímulos 60/40. Sin embargo, si la causa de las diferencias entre B1 y B2 fuese el diferente coste del error (falso positivo o falso negativo) en cada una de las dos versiones de instrucciones, estas diferencias seguirían manifestándose en el Experimento 7, aún con el cambio del sesgo de la IA, pero, como explicaremos en detalle en la sección de Método, esperaríamos más errores en B2 que en B1, al contrario que en los experimentos previos.

Como consecuencia, el segundo objetivo del Experimento 7 fue analizar si el tipo de error que cometía el sistema de inteligencia artificial, falso positivo o falso negativo, pudo influir en la tendencia de los participantes a seguir sus recomendaciones. Nuestra hipótesis en el Experimento 7 era que tanto el tipo de error (falso positivo y falso negativo) como el mensaje advirtiendo de la presencia de errores en la IA (información del error y no información) tendrían una influencia en la decisión de los participantes a adherirse a las recomendaciones sesgadas de la IA. Esperábamos que los participantes fuesen más reticentes a seguir la recomendación de la IA ante el riesgo de incurrir en un falso negativo, pero que, sin embargo, fuesen más proclives a aceptar su consejo en la condición falso positivo. También esperábamos que los participantes alertados de forma general de posibles sesgos en la IA cometiesen menos errores en la tarea de clasificación que los participantes que no recibieron este mensaje. Teniendo en cuenta ambos factores, esperábamos un mayor efecto de la información para mitigar los errores en ensayos con estímulos 60/40 (aquellos en los que la IA erraba) en la condición falso positivo. En la condición falso negativo, si los participantes tendían a ser más cautos, el margen para la mejora fruto de la intervención sería menor.

Método

Participantes

Reclutamos a 200 participantes a través de la plataforma Prolific, pero los datos de dos de ellos se perdieron por problemas técnicos. Por lo tanto, 198 participantes formaron la muestra final, cuya edad media fue 30.6 (*SD* = 8.99, rango de edad 19-72 años). El 58.1% de la muestra se autodefinió como género femenino, y el 41.9% como masculino. Ofertamos la participación en nuestro experimento a personas que hablasen inglés con fluidez y que no hubiesen tomado parte en experimentos previos de nuestro laboratorio. La recompensa establecida por realizar la tarea, que llevo unos 14 minutos de media, fue 1.29 libras.

Del cruce del factor de la información sobre los errores (info-error y no-info) y de los dos tipos de error (falso positivo y falso negativo) surgen cuatro grupos diferentes. Por lo tanto, los voluntarios fueron distribuidos al azar entre los grupos: info error – FP (*n* = 50), info error – FN (*n* = 50), sin info – FP (*n* = 49) o sin info – FN (*n* = 49). El análisis de sensibilidad reveló que con un tamaño muestral de 198 participantes podíamos detectar un tamaño del efecto tan pequeño como *f* = 0.20 con una potencia del 80% para un ANOVA 2 (información error: info-error, no-info) x 2 (tipo de error: falso positivo, falso negativo).

Materiales y Procedimiento

El procedimiento seguido en el Experimento 7 se basó en el utilizado para el Experimento 5. El único cambio introducido en el Experimento 7 fue el sesgo de la IA ficticia: decidimos que la IA clasificaría de forma errónea los estímulos con una proporción de células de color oscuro y células de color claro 60/40, mientras que en todos los experimentos previos de la presente Tesis Doctoral la IA se equivocó siempre en la clasificación de los estímulos 40/60.

En el Experimento 5 y en el Experimento 6 detectamos que los participantes asignados

a B1 cometieron más errores de media en la clasificación de los estímulos 40/60 que los participantes de B2, durante la fase de la tarea asistida por la IA. Una de las explicaciones posibles para este intrigante efecto es que se deba al diferente significado o coste del error que supondría seguir la recomendación errónea de la IA en cada una de las versiones de las instrucciones. Si se tratase de una situación real, dar un diagnóstico incorrecto de la presencia de una patología que luego resulta no estar (i.e., falso positivo) tiene un riesgo menor para la vida de una persona que fallar en detectar una patología que sí está presente (i.e., falso negativo). Por este motivo, planteamos que, en los dos experimentos anteriores, cuando los participantes tuvieron alguna duda sobre a qué categoría asignar la muestra 40/60, en B1 optaron por clasificarla como positivo, siguiendo la sugerencia de la IA, pues era el error menos costoso. Por otro lado, en B2, clasificar como negativo una muestra de tejido que parecía ser positiva tendría unas implicaciones graves, por lo que fue mejor optar por la categoría "positivo" contradiciendo la sugerencia de la IA.

En el presente Experimento 7 planteamos como hipótesis que el tipo de error, con su significado y su coste asociados, tendría una influencia en la tendencia de los participantes a seguir las recomendaciones de la IA durante la tarea de clasificación. En esencia, creemos que, ante la posibilidad de que la IA esté equivocándose en su recomendación, la decisión de seguir su consejo o ignorarlo va a verse influida por las consecuencias que se atribuyan al error. Las consecuencias serían percibidas como más graves en un falso negativo, (i.e., no detectar el síndrome en un paciente enfermo), que en un falso positivo.

Deteniéndonos un poco más en la descripción del cambio del sesgo a los estímulos 60/40 en lugar de los 40/60, debemos explicar qué implica este cambio en relación al tipo de error que comete ahora la IA en los balanceos B1 y B2, como se describe en la Tabla 5.

Tabla 5

Clasificación Correcta Según el Balanceo, Recomendación Errónea de la IA y Diferente Significado del Error para los Estímulos con Proporción de Células Oscuras y Claras 40/60 de los Experimentos 1 a 6 y 60/40 del Experimento 7.

Experimento	Estímulo	Balanceo	Clasificación correcta	Error IA	Tipo de error
1-6	40/60	B1	Negativo	Positivo	Falso positivo
		B2	Positivo	Negativo	Falso negativo
7	60/40	B1	Positivo	Negativo	Falso negativo
		B2	Negativo	Positivo	Falso positivo

Las instrucciones en el Experimento 7 fueron las utilizadas en el Experimento 5. Recordemos que en el Experimento 7 decidimos poner de nuevo a prueba el mismo mensaje de advertencia de error empleado en el Experimento 5, pero tratando de controlar el efecto del contrabalanceo de las instrucciones que interfirió en el experimento previo. Siguiendo estas instrucciones, en B1 la clasificación correcta del estímulo 60/40 era positivo, y, por tanto, el sesgo de la IA en el Experimento 7 consistió en sugerir siempre para estos estímulos la etiqueta “negativo”, es decir, el error que cometía la IA en las muestras 60/40 fue un falso negativo. De la misma manera, en B2 el error de la IA en las muestras 60/40 fue un falso positivo (ver Tabla 5). En los dos anteriores experimentos de la Serie Experimental 2, los participantes cometieron de media más errores en la clasificación de los ensayos 40/60 en B1 que en B2. Si estas diferencias se debieron al tipo de error, falso positivo en B1 y falso negativo en B2, entonces, el cambio en el sesgo de la IA debería reflejarse en los resultados del Experimento 7. Si es así, esperamos que la media de errores de clasificación en los ensayos 60/40 sea mayor en B2, donde el tipo de error es falso positivo, que en B1, ahora asociado al error falso negativo. En resumen, al cambiar el sesgo de la IA a otro estímulo cambia el tipo de error que se da en cada balanceo y esperamos observar un patrón de resultados inverso al

detectado en los Experimentos 5 y 6. Es decir, esperamos que los participantes cometan más errores siempre en la condición falso positivo, independientemente del balanceo (B1 o B2) con el que se asocie, lo que daría apoyo a la hipótesis de que el tipo de error y su coste influyen en la decisión de los participantes de seguir la recomendación de la IA.

En el Experimento 7 en el total de 60 ensayos de la Fase 1 de la tarea de clasificación, aparecieron 10 ensayos con estímulos 60/40. Entre los 25 ensayos de la Fase 2 de la tarea de clasificación, el estímulo 60/40 se mostró en un total de 5 ensayos (es decir, mismo número de ensayos y proporción de error que en experimentos previos en los que el sesgo se situó en los ensayos 40/60).

Al finalizar la tarea experimental, en el presente Experimento 7 realizamos a los participantes las mismas preguntas que en experimentos anteriores. Los participantes estimaron en una escala de 0 a 100 el porcentaje de recomendaciones ofrecidas por la IA que habían percibido como erróneas, como en el Experimento 6. Esta escala nos permitiría conocer si el grupo info-error estimaba el porcentaje de error en las recomendaciones de la IA con mayor precisión que el grupo no-info en el Experimento 7.

Sintetizando nuestras predicciones principales para el Experimento 7, esperábamos que en esta ocasión un mensaje general advirtiendo de posibles errores y sesgos en la IA sí mostrase un efecto, es decir, que los participantes del grupo info-error cometiesen menos errores que los participantes del grupo no-info. Creemos que no observamos este efecto en el Experimento 5 por el extraño efecto del contrabalanceo de las instrucciones y debido a un patrón de comportamiento también extraño de algunos participantes en ese experimento.

En el Experimento 7, esperábamos además que la tendencia de los participantes a aceptar las recomendaciones de la IA estuviese influida por el coste del error. Esperábamos que cuando el error sistemático de la IA era falso positivo los participantes siguiesen más su recomendación, cometiendo más errores, que cuando era falso negativo, condición en la que

los participantes serían más cautos en sus respuestas debido al mayor coste percibido del error. Por lo tanto, esperábamos que la media de errores 60/40 fuese menor en los participantes avisados de posibles errores en la IA que en los participantes que no observaron esta información. A la vez esperábamos que la media de errores 60/40 fuese menor en la condición falso negativo que en la condición falso positivo, para ambos grupos info-error y no-info. Teniendo en cuenta ambos factores, nuestra predicción fue que los participantes no-info FP serían los que cometerían más errores, seguido por los participantes no-info FN, los cuales no distarían mucho en número de errores de info-error FP. Finalmente, los participantes info-error FN serían los que cometerían menos errores en la clasificación de muestras 60/40.

Con el procedimiento planteado en el Experimento 7 comprobaríamos si las diferencias encontradas entre B1 y B2 podían ser atribuidas a una influencia del tipo de error, o si debíamos pensar en otros factores. Si el problema de los experimentos previos se debiese a un efecto perceptivo, o a que el criterio de clasificación fuese más intuitivo en uno de los balanceos, dado que ni los estímulos ni el criterio de clasificación cambiaron en el Experimento 7, el resultado debería ser el mismo al observado en los Experimentos 5 y 6, con más errores en B1. En cambio, si los resultados observados en los dos experimentos previos estuvieron influidos por el tipo de error hacia el que se orientaba el sesgo de la IA, en el Experimento 7 deberíamos observar más errores de media en B2, en el que el error de recomendación en los estímulos 60/40 era un falso positivo. En síntesis, en el Experimento 7 tratamos de separar un posible efecto del tipo de error de un posible efecto del balanceo de las instrucciones, para comprobar cuál ejerció una influencia en el comportamiento de los participantes. El Experimento 7 fue prerregistrado en Aspredicted.org https://aspredicted.org/yz79a.pdf

Resultados y Discusión

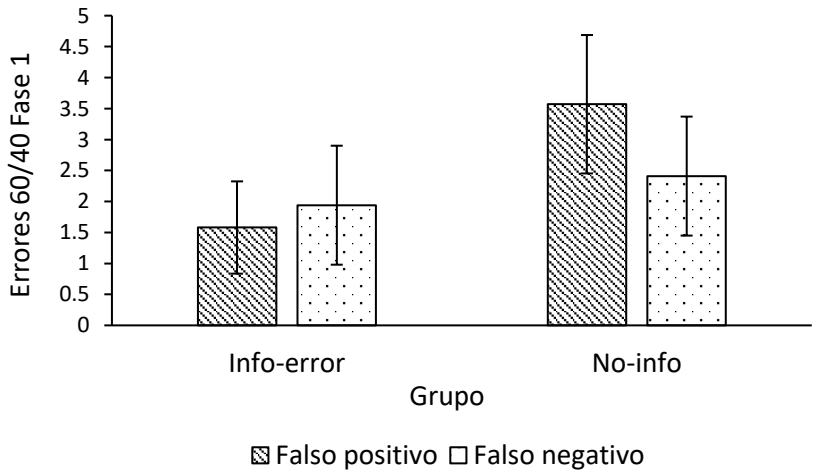
Para interpretar de forma adecuada los resultados del Experimento 7, y para poder establecer comparaciones con los resultados de los experimentos anteriores de la Serie

Experimental 2, en necesario tener presente que el tipo de error falso positivo se asoció al balanceo B2, y el tipo de error falso negativo al balanceo B1 cuando el sesgo de la IA consistió en un error sistemático en estímulos 60/40 (ver Tabla 5). El emparejamiento entre tipo de error y balanceo fue diferente en experimentos anteriores, cuando el sesgo se situó en estímulos 40/60: el falso positivo se equiparó a B1, y el falso negativo a B2. Para facilitar la comprensión y comparación de los datos, en esta sección tan sólo haremos referencia a la variable tipo de error (falso positivo, falso negativo). La información sobre el balanceo queda, por tanto, implícita.

En el presente Experimento 7, el grupo info-error tendió a cometer menos errores en la tarea de clasificación que el grupo no-info, como se muestra en la Figura 29. Parece que en el Experimento 7 proporcionar información de forma general sobre errores y sesgos en la IA sí motivó a los participantes a ser más cautos en el momento de aceptar sus recomendaciones, efecto que no encontramos en el Experimento 5.

Figura 29

Número Medio de Errores de Clasificación de los Ensayos 60/40 Durante la Fase de la Tarea con Asistencia de la IA Sesgada del Experimento 7 para los Grupos Info-Error FP (n = 50), Info-Error FN (n = 50), No-Info FP (n = 49) o No-Info FN (n = 49).



Nota. Las barras de error indican el intervalo de confianza de la media al 95%.

Los voluntarios que no vieron el mensaje advirtiendo sobre errores en la IA antes de realizar la tarea, o grupo control, tendieron a cometer más errores ante el sesgo falso positivo (*M* = 3.57, *SD* = 4.05) que ante el falso negativo (*M* = 2.41, *SD* = 4.05), aunque sin diferencias significativas. Por otro lado, aquellos participantes que sí fueron advertidos sobre errores en la IA apenas difirieron en función del tipo de error con una media de 1.58 errores (*SD* = 2.70) en la condición falso positivo, y una media de 1.94 errores (*SD* = 3.49) en la condición falso negativo. En definitiva, tras la observación de la Figura 29, el resultado más relevante fue un efecto de atenuación de los errores 60/40 en la Fase 1, tras informar sobre la presencia de posibles errores en las recomendaciones de la IA, pero especialmente para el tipo de error falso positivo.

El ANOVA 2 (información sobre los errores: info-error, no-info) x 2 (tipo de error: falso positivo, falso negativo) confirmó un efecto principal de la información sobre el error, $F(1,194) = 6.263$, $p = .013$, $\eta^2_p = .031$, pero no un efecto del tipo de error, $F(1,194) = 0.668$, $p = .415$, $\eta^2_p = .003$, ni interacción entre ambos factores, $F(1, 194) = 2.402$, $p = .123$, $\eta^2_p = .012$.

Aunque la interacción entre la información sobre el error y el tipo de error no fue significativa, decidimos igualmente realizar comparaciones post hoc con corrección de Tukey entre las medias de los cuatro grupos, para confirmar la tendencia observada en la Figura 29: si bien no parecen existir diferencias entre las medias de errores de los grupos info-error FN y no-info FN, cuando prestamos atención al tipo de error falso positivo vemos que el grupo info-error FP cometió menos errores que no-info FP. La diferencia de medias para estas comparaciones post hoc se muestran en la Tabla 6. Tras realizar comparaciones por pares de las medias de errores de los cuatro grupos, observamos diferencias significativas en el número medio de clasificaciones erróneas 60/40 solo entre el grupo info-error FP y no-info FP.

Tabla 6

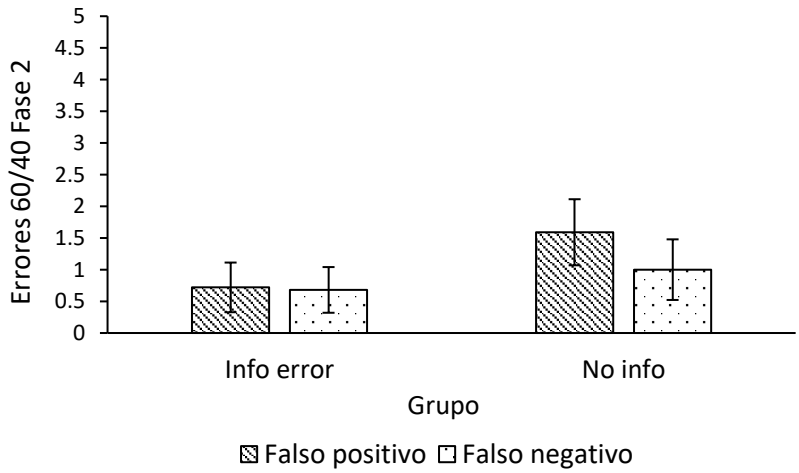
Comparaciones Post Hoc con Corrección de Tukey de la Interacción de Información Sobre el Error y el Tipo de Error para la Fase 1 con Asistencia de la IA.

Comparación						
Grupo	**Tipo de error**	**Grupo**	**Tipo de error**	**Diferencia de medias**	***t***	p_{tukey}
Info error	Falso positivo	Info error	Falso negativo	-0.360	-0.521	.954
		No info	Falso positivo	-1.991	-2.865	.024
		No info	Falso negativo	-0.828	-1.192	.633
	Falso negativo	No info	Falso positivo	-1.631	-2.347	.091
		No info	Falso negativo	-0.468	-0.674	.907
No info	Falso positivo	No info	Falso negativo	1.163	1.665	.345

En la fase sin asistencia de la IA o Fase 2, aún el grupo no-info cometió más errores que el grupo info-error, como se observa en la Figura 30.

Figura 30

*Número Medio de Errores de Clasificación de los Ensayos 60/40 Durante la Fase de la Tarea Sin Asistencia del Experimento 7 para los Grupos Info-Error FP (*n *= 50), Info-Error FN (*n *= 50), No-info FP (*n *= 49) o No-Info FN (*n *= 49).*



Nota. Las barras de error indican el intervalo de confianza de la media al 95%.

El gráfico sugiere que la información avisando de errores en las recomendaciones tuvo también un efecto preventivo sobre la herencia del sesgo. Para comprobarlo, realizamos un ANOVA 2 (información sobre el error: info-error, no-info) x 2 (tipo de error: falso positivo, falso negativo) pero esta vez con las clasificaciones erróneas de los estímulos 60/40 de la Fase 2 como variable dependiente. Observamos de nuevo un efecto principal de la información sobre el error, $F(1,194) = 7.00$, $p = .009$, $\eta^2_p = .035$, pero no un efecto del tipo de error, $F(1,194) = 4.94$, $p = .162$, $\eta^2_p = .010$, ni interacción entre ambos factores, $F(1, 194) = 1.50$, $p = .222$, $\eta^2_p = .008$.

Focalizándonos en el tipo de error, falso positivo y falso negativo, en la Figura 30 vemos una atenuación en el número medio de errores del grupo info-error FP en comparación al grupo no-info FP. Las comparaciones post hoc con corrección de Tukey, que pueden consultarse en la Tabla 7, confirmaron que esta diferencia fue significativa.

Tabla 7

Comparaciones Post Hoc con Corrección de Tukey de la Interacción de Información Sobre el Error y el Tipo de Error para la Fase 2 o Sin Asistencia de la IA.

Comparación						
Grupo	**Tipo de error**	**Grupo**	**Tipo de error**	**Diferencia de medias**	***t***	p_{tukey}
Info error	Falso positivo	Info error	Falso negativo	0.040	-0.521	.999
		No info	Falso positivo	-0.871	-2.865	.034
		No info	Falso negativo	-0.280	-1.192	.816
	Falso negativo	No info	Falso positivo	-0.911	-2.347	.024
		No info	Falso negativo	-0.320	-0.674	.747
No info	Falso positivo	No info	Falso negativo	0.591	1.665	.254

Entre nuestras predicciones planteábamos que cuando el coste del error era menor, en la condición falso positivo, los participantes serían más proclives a seguir la recomendación de la IA, sin embargo, serían más cautos cuando el sesgo de la IA se orientaba a un falso negativo,

dado que las consecuencias asociadas al error eran más graves. Parece que, en efecto, cuando no se proporcionó un aviso alertando de errores, de forma general el tipo de error falso negativo promovió una tendencia de respuesta más cauta. En contraposición, los participantes tendieron a seguir la recomendación errónea de la IA con más frecuencia cuando el error fue falso positivo. Dado que la tendencia de respuesta de los participantes en el falso positivo ya era cautelosa, la información sobre los errores no introdujo diferencias entre los participantes para esta condición, como muestra la ausencia de diferencias significativas entre info-error FN y no-info FN. Sin embargo, la advertencia de que la IA podía estar sesgada indujo a los participantes a ser más prudentes en sus decisiones específicamente cuando el sesgo consistió en un error falso positivo, lo cual se manifestó en una reducción en los errores de clasificación del grupo info-error FP en comparación al grupo no-info FP en la Fase 1. La mayor prudencia en las respuestas del grupo info-error FP durante la fase con asistencia de la IA se acompañó también de una atenuación de la herencia del sesgo cuando posteriormente estos participantes se enfrentaron a clasificar muestras de tejido sin asistencia durante la Fase 2. Así lo demuestra la reducción de los errores en ensayos 60/40 del grupo info-error frente al grupo control o no-info, cuando el error sistemático de la IA se orientó a un falso positivo, también en la segunda fase de la tarea de clasificación.

En el Experimento 5 y en el Experimento 6 hallamos diferencias en los errores que cometieron los participantes en la clasificación de las muestras 40/60 en función del contrabalanceo de las instrucciones y planteamos como hipótesis que el origen de estas diferencias podría ser el diferente coste del error en cada uno de los balanceos B1 y B2. Para indagar en las posibles causas de tal diferencia, en el Experimento 7 también decidimos explorar si se dieron diferencias en los errores cometidos en ensayos 40/60 en función del tipo de error falso positivo y falso negativo.

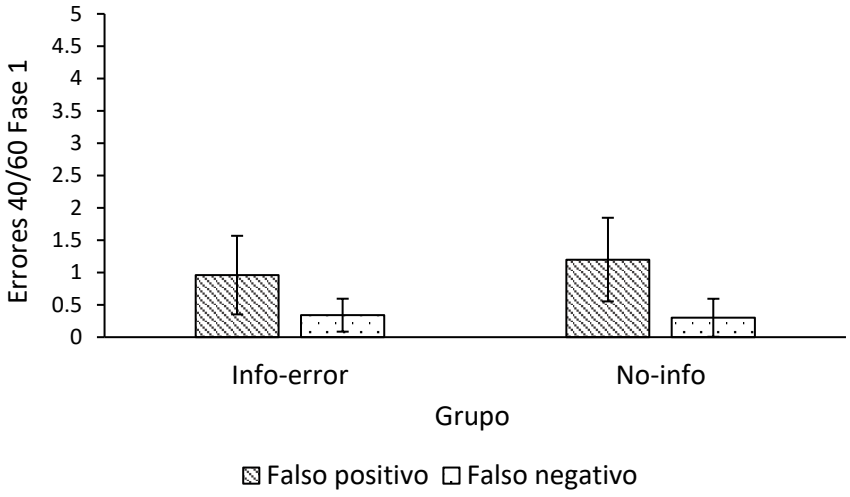
Recordemos que en el Experimento 7, el sesgo de la IA se situó en las muestras de

tejido 60/40, es decir, la recomendación de la IA fue errónea únicamente en los 10 ensayos con estímulos 60/40 de la tarea de clasificación. Esto significa que en los 10 ensayos con estímulos 40/60 la IA dio una sugerencia correcta. Por ello, esperábamos que los voluntarios apenas cometiesen errores en la clasificación de dichas muestras en el Experimento 7.

En la Figura 31 se muestra una comparación entre las medias de errores en los ensayos 40/60 en función del tipo de error por separado para los grupos info-error y no-info. Aunque de forma general los participantes cometieron pocos errores en la clasificación de muestras 40/60 tal como predecíamos, es curioso observar que también en este caso ambos grupos cometieron más errores falso positivo.

Figura 31

*Número Medio de Errores de Clasificación de los Ensayos 40/60 Durante la Fase 1 del Experimento 7 para los Grupos Info-Error FP (*n *= 50), Info-Error FN (*n *= 50), No-Info FP (*n *= 49) y No-Info FN (*n *= 49)*



Nota. Las barras de error indican el intervalo de confianza de la media al 95%.

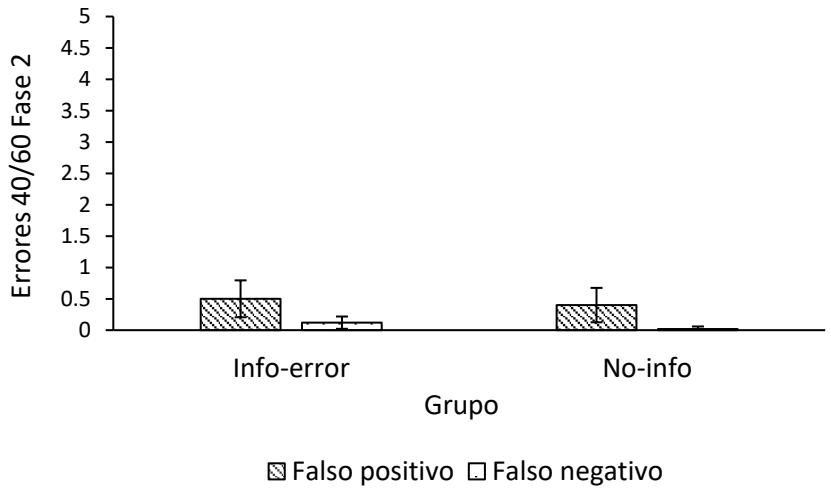
El ANOVA 2 (información sobre el error: info-error, no-info) x 2 (tipo de error: falso positivo, falso negativo), con los errores cometidos en la clasificación de los ensayos 40/60 de

la Fase 1 como variable dependiente, confirmó un efecto principal del tipo de error, $F(1,194) = 9.383$, $p = .003$, $\eta^2_p = .046$, para la Fase 1, pero no se observó ni un efecto principal de la información sobre el error, $F(1,194) = 0.180$, $p = .672$, $\eta^2_p = .001$, ni una interacción entre ambos factores, $F(1,194) = 0.315$, $p = .576$, $\eta^2_p = .002$.

También los participantes cometieron más errores en ensayos 40/60 en la condición falso positivo, que en la condición falso negativo, durante la fase de la tarea sin asistencia de la IA, como se puede ver en la Figura 32. En la Fase 2, el ANOVA 2 x 2 confirmó un efecto principal del tipo de error, $F(1,194) = 12.347$, $p < .001$, $\eta^2_p = .060$, pero no un efecto principal de la información sobre el error, $F(1,194) = 0.767$, $p = .382$, $\eta^2_p = .004$, ni interacción entre ambos factores, $F(1,194) = 0.001$, $p = .972$, $\eta^2_p = .000$.

Figura 32

*Número Medio de Errores de Clasificación de los Ensayos 40/60 Durante la Fase 2 del Experimento 7 para los Grupos Info-Error FP (*n *= 50), Info-Error FN (*n *= 50), No-Info FP (*n *= 49) y No-Info FN (*n *= 49).*



Nota. Las barras de error indican el intervalo de confianza de la media al 95%.

El tipo de error (falso positivo vs. falso negativo) que podían cometer los participantes

en la clasificación de las muestras en el Experimento 7 se derivó directamente de los criterios de clasificación descritos en las instrucciones para cada uno de los balanceos. Por lo tanto, explorar si se dieron diferencias en las respuestas de los participantes a ensayos 40/60 en función del tipo de error lleva implícito explorar si existieron diferencias en la clasificación de estos ensayos en función del contrabalanceo. Como ya sucedió en el Experimento 5 y en el Experimento 6, de nuevo en el Experimento 7 detectamos diferencias en la clasificación de las muestras 40/60 en función del tipo de error (o instrucciones B1 y B2).

En la sección de Resultados y Discusión del Experimento 6 planteábamos que quizá la diferencia observada entre los balanceos podría deberse a algún efecto perceptivo, a que las instrucciones resultaran más intuitivas en un balanceo que en otro, o al diferente coste del error asociado a seguir la recomendación sesgada de la IA en cada uno de los balanceos. En el Experimento 7 la etiqueta de recomendación de la supuesta IA indicaba una clasificación correcta en los ensayos 40/60. Si las diferencias que detectamos en experimentos previos en función del balanceo se debieron a aspectos perceptivos o si uno de los criterios de clasificación resultó más intuitivo y fácil de comprender, y recordar, estas variables habrían perdido importancia una vez que la recomendación de la IA era coherente con la información objetiva de la muestra de tejido. Es decir, la etiqueta de recomendación correcta habría sido una clave bastante saliente para la atención de los participantes, y si se hubiesen adherido a ella por completo no habrían cometido errores en la clasificación de muestras 40/60, atenuando la influencia de algún otro factor extraño (e.g., que uno de los criterios de clasificación resultase más sencillo) y homogeneizando las respuestas de los participantes. Los resultados observados tanto para las muestras 60/40 como para las muestras 40/60 podrían explicarse por una tendencia en los participantes a preferir siempre incurrir en un falso positivo ante casos en los que no tenían una certeza completa de cuál era el diagnóstico correcto. Es decir, muestras en las que no era extremadamente fácil discriminar qué color estaba presente en una mayor proporción, como en las muestras 60/40 y 40/60, y muestras

para las cuales la IA proporcionaba una recomendación errónea generando incertidumbre en los participantes, como en las muestras 60/40 en el Experimento 7. Por lo tanto, una explicación plausible para los datos observados podría ser que, ante cualquier caso de duda, los voluntarios preferían asignar las muestras de tejido a la categoría positivo, pues se aseguraban de elegir la opción menos arriesgada. En el contexto de nuestra tarea de clasificación, un diagnóstico médico, determinar que un paciente está afectado por una determinada enfermedad en una primera prueba de cribado, aunque posteriormente se descubra que realmente la persona no estaba enferma, es la opción más segura, al menos en comparación al riesgo de obviar bajo la etiqueta de un diagnóstico negativo a una persona enferma.

Volviendo a focalizarnos de nuevo solo en los resultados del Experimento 7, nos preguntamos a continuación si, quizá, los participantes que fueron informados de sesgos y errores en la IA detectaron más los errores de las recomendaciones que los participantes que no vieron esta información. Un 93% de los participantes del grupo info-error declaró haber detectado errores en las recomendaciones de la IA, este porcentaje descendió al 81,6% en el grupo no-info. Los participantes del grupo info-error reportaron haber detectado con más frecuencia errores en las recomendaciones de la IA que los participantes del grupo no-info, $\chi^2(1,198) = 5.80$, $p = .016$. No encontramos diferencias en el número de participantes que afirmaron haber detectado errores en las recomendaciones de la IA en función del tipo de error $\chi^2(1,198) = 0.412$, $p = .521$.

A aquellos participantes que afirmaron detectar errores se les pidió también que realizasen una estimación del porcentaje de error que habían percibido en las recomendaciones de la IA. Los participantes del grupo info-error estimaron de media un 24.4% (SD = 17) de recomendaciones erróneas de la IA, mientras que el porcentaje estimado por el grupo no-info de media fue 20.5 % (SD = 14.6), una estimación muy certera, pues el valor real

era 20%. No observamos diferencias entre las estimaciones de ambos grupos, $t(166) = 1.58$, $p = .117$, $d = 0.244$.

Al igual que en experimentos previos de la presente Tesis Doctoral, observamos una correlación positiva y significativa entre cuánto consideraron los participantes a la IA de ayuda ($r = .43$, $p < .001$), cuánto siguieron el consejo de la IA ($r = .49$ $p < .001$), el acierto que atribuyeron a la IA ($r = .40$, $p < .001$), la confianza en la IA en el área de la salud ($r = .35$, $p < .001$) y el número de errores de clasificación que cometieron en los estímulos 60/40. Observamos también para la Fase 2 que aquellos participantes que cometieron más clasificaciones incorrectas de los ensayos 60/40 fueron también los que más consideraron a la IA de ayuda ($r = .25$, $p < .001$), quienes más afirmaron haber seguido sus recomendaciones ($r = .33$, $p < .001$), quienes les atribuyeron más precisión ($r = .32$, $p < .001$) y mostraron una mayor confianza en la IA aplicada al ámbito de la salud ($r = .22$, $p = .002$).

3.2 Discusión de la Serie Experimental 2

Advertir a las Personas de Posibles Sesgos en la IA podría Mitigar la Herencia de esos Errores Sistemáticos

Los mensajes que adviertan a las personas de posibles errores y sesgos en la inteligencia artificial, antes de iniciar su interacción con el sistema, podrían minimizar la influencia de las recomendaciones sesgadas de la IA sobre sus decisiones y la herencia del sesgo. En la Serie Experimental 2 pusimos a prueba esta hipótesis, pero, ¿qué podemos concluir a partir de los resultados observados en los tres experimentos de la segunda serie?

En el Experimento 5, tanto el grupo que fue alertado sobre la presencia de posibles errores, como el grupo que no recibió esta información, se comportaron de forma muy similar en la tarea de clasificación, sin diferencias en el número medio de errores cometidos en los ensayos 40/60, aquellos donde la IA se equivocaba. Por lo tanto, en el Experimento 5 no observamos diferencias en las respuestas de los participantes en función de la información con la que partieron al inicio de la tarea, aunque consideramos que en este experimento el posible efecto de la información sobre el error pudo quedar enmascarado por el efecto del contrabalanceo de las instrucciones. En otras investigaciones, el mero hecho de advertir sobre posibles errores en los resultados del modelo de IA tampoco obtuvo un claro éxito para mitigar la tendencia de los individuos a seguir las recomendaciones erróneas de una máquina (Vered et al., 2023). Kupfer et al. (2023) observaron que informar sobre posibles errores en una IA que proporcionó asistencia a los participantes durante una tarea de selección de personal aumentó la verificación de información de los voluntarios, pero no supuso una mejora en la calidad objetiva de las decisiones, en comparación al grupo que realizó la misma tarea sin asistencia. Rezazade Mehrizi et al. (2023) pusieron a prueba otro tipo de intervenciones, aunque dirigidas al mismo objetivo. Estos autores optaron por (a) proporcionar información sobre cómo la IA

había alcanzado una determinada decisión, con variaciones en su forma (desde no información, a información parcial o información detallada) y (b) promover una determinada actitud hacia la IA antes de la interacción con ella, actitud que podía ser positiva, negativa o ambivalente. Sin embargo, ninguna de estas intervenciones tuvo un efecto, y en esta investigación radiólogos expertos siguieron tanto las sugerencias correctas como las incorrectas del sistema de IA para realizar diagnósticos tras el examen de mamografías.

En el Experimento 6, la información sobre el error que cometía la IA en sus recomendaciones fue más específica y fue enmarcada de dos formas diferentes: haciendo énfasis en el 20% de errores o en el 80% de aciertos. La información explícita señalando un 20% de error en los resultados de la inteligencia artificial sí pareció atenuar la influencia de este agente en las decisiones de los participantes, así como la herencia del sesgo. Al menos, los análisis de los datos obtenidos en el Experimento 6 sí revelaron diferencias en el número de errores en ensayos 40/60 entre el grupo que no recibió ningún dato sobre la precisión de la IA que le prestaría apoyo, y el grupo que fue informado sobre el porcentaje de error que solía cometer el sistema de IA en sus recomendaciones. Sin embargo, no se observaron diferencias entre el grupo al que se informó de que el sistema acertaba en el 80% de las ocasiones y el grupo sin información, o el grupo informado sobre el 20% de error (ver Tabla 2B del Apéndice B para consultar los índices de error de cada uno de los tipos de estímulos en función del grupo al que fueron asignados los participantes en los Experimentos 5 y 6 de la Serie Experimental 2).

En relación al porcentaje de acierto de la falsa IA de nuestro experimento debemos señalar que Yu et al. (2019) observaron que una precisión del 80% en un modelo de inteligencia artificial se asoció a confianza en el modelo, de forma que los individuos estuvieron dispuestos a aceptar y seguir su consejo. Por ello, creemos que en nuestro Experimento 6, proporcionar a los participantes el dato de la precisión de la IA simulada que

les iba a dar asistencia no atenuó la tendencia a aceptar recomendaciones sesgadas, ni tampoco la herencia del sesgo, en comparación al grupo sin información, porque un 80% de aciertos pudo haber sido considerado un índice bastante bueno.

Como consecuencia, quizá los voluntarios formaron una imagen mental global de esta IA como confiable, suprimiendo el esfuerzo de ejercer una supervisión de sus resultados caso a caso (Buçinca et al., 2021). Esta solución heurística, si bien pudo haber reducido la carga mental del participante, comprometió la calidad de la toma de decisiones, pues la realidad fue que en el 20% de las ocasiones la recomendación de la IA estaba equivocada. En relación a la anterior observación, cabe recordar que el Experimento 6 mostró también que enmarcar la misma información sobre la precisión de la IA enfatizando el 20% de error en las recomendaciones, en lugar del 80% de aciertos, sí logro atenuar los errores de los participantes en aquellos ensayos en los que la IA invitaba a tomar una decisión equivocada. Enfatizar el porcentaje de error también mostró mitigar la herencia del sesgo. Visibilizar el error posiblemente invitó a las personas a ejercer una supervisión más minuciosa de la IA porque hizo más saliente la posibilidad de que una maquina no sea siempre precisa. Por lo tanto, es relevante tener en cuenta el efecto del enmarcado de la información, un efecto robusto en Psicología y que se ha demostrado en múltiples contextos (Kahneman, 2003; Tversky & Kahneman, 1981), también a la hora elegir cómo reportar información sobre una herramienta de inteligencia artificial.

Parece que para promover que las personas depositen una confianza adecuada en la inteligencia artificial, deben ser capaces de conocer el riesgo de error. Con este fin, sería conveniente que la información sobre la precisión de la IA se focalizase en el porcentaje de ocasiones en las que el modelo tiende a equivocarse, más que en el porcentaje de aciertos que es capaz de alcanzar, para fomentar que los usuarios evalúen la información, detecten posibles sesgos y los corrijan. Continuando con posibles recomendaciones para promover que las

personas calibren de forma adecuada la confianza que pueden depositar en un determinado modelo de IA, proporcionar un índice de la precisión de la IA caso a caso podría ser más efectivo que dar un dato de su fiabilidad global. Algunos estudios indican que esta estrategia sería efectiva para conseguir que las personas discriminen de forma correcta cuándo confiar o no confiar en una determinada recomendación de la IA para cada caso específico (Banerji et al., 2023; Zhang et al., 2020). Por ejemplo, en el caso de nuestros experimentos, esto se traduciría en ofrecer un dato que indicase lo probable que es que la IA esté dando una recomendación errónea para cada muestra de tejido en cada ensayo. Vendría a ser algo así como permitir a la IA expresar cuándo tiene dudas en una recomendación tras la evaluación de cada caso concreto. De esta forma, la persona que interactúa con la IA podría identificar mejor en qué casos es necesario realizar una evaluación minuciosa para prevenir errores.

No podemos continuar esta discusión sin hacer mención al efecto del contrabalanceo de las instrucciones detectado en la Serie Experimental 2, y que ya introdujimos previamente al comentar la ausencia de diferencias entre los grupos en el Experimento 5. La diferencia en clasificaciones incorrectas de los estímulos 40/60 en función del balanceo de instrucciones B1 y B2 coarta la interpretación de los resultados observados en los Experimentos 5 y 6, puesto que quizá el extraño efecto pudo interferir con la manipulación que buscábamos poner a prueba: si avisar sobre errores en la IA podía prevenir a los participantes de seguir sus recomendaciones sesgadas. Por este motivo, en el Experimento 7 cambiamos el error sistemático de la IA a los estímulos con una proporción de células oscuras y claras 60/40 para poner a prueba nuestra manipulación evitando interferencias de algún posible efecto perceptivo asociado a las muestras 40/60.

En el Experimento 7, durante la fase de la tarea con asistencia de la IA, los participantes avisados sobre la posible presencia de errores en las recomendaciones cometieron menos errores en la clasificación de muestras 60/40 (aquellas en las que la IA

proporcionaba de forma sistemática una sugerencia incorrecta) en comparación al grupo control sin información. Pero, el resultado llamativo de este Experimento 7 fue que ese efecto de decremento de las clasificaciones incorrectas se dio específicamente cuando seguir la recomendación sesgada de la IA llevaba a incurrir en un falso positivo, que era el error menos costoso. Los voluntarios del grupo control, no advertidos sobre posibles errores, tendieron a equivocarse más en la clasificación cuando la IA sugería un falso positivo. Por ello, el aviso sobre posibles errores de la IA atenuó la tendencia a incurrir en falsos positivos. Sin embargo, cuando seguir la recomendación sesgada de la IA llevaba a un falso negativo, los participantes de este grupo no informado sobre errores tendieron a ser más cautos, mostrando un menor número medio de errores en ensayos 60/40, pues el posible error era más costoso. El mensaje de aviso no tuvo un efecto claro frente al sesgo falso negativo porque, si los participantes ya eran cautelosos en sus respuestas y frenaban la tendencia a seguir la etiqueta de recomendación en la clasificación, poco podía sumar el informar de errores en la IA a esta estrategia. Quizá no detectamos cambio en la condición falso negativo debido a un efecto suelo.

Las consecuencias que se derivan de las decisiones, en un determinado contexto de colaboración, parecen modular la confianza que las personas depositan en la IA, como han señalado algunas investigaciones (Ashoori & Weisz, 2019; Fahnenstich et al., 2024; Rieger et al., 2022). Sin embargo, aún se conoce poco acerca del efecto de este coste percibido de las decisiones y sus implicaciones. A día de hoy no es posible predecir si un alto coste percibido tendería a promover una mayor confianza en la IA o, por el contrario, desconfianza de la IA, pues se ha observado evidencia en ambas direcciones. Fahnenstich et al. (2024) observaron que los participantes se adherían más a las recomendaciones de la IA en un contexto de alto riesgo (estimar la presencia de bacterias en muestras de tejido humano), en comparación con un contexto de bajo riesgo (estimar la presencia de bacterias en muestras de tejidos de plantas). En la interpretación de los resultados del Experimento 7, nosotras planteamos que a

mayor coste percibido de la decisión (un posible falso negativo con la consecuencia de obviar un paciente que necesita atención médica) los participantes tendieron a adherirse menos a la recomendación de la IA. El papel que juega este coste o gravedad de las consecuencias de la decisión en la tendencia de los individuos a aceptar las recomendaciones de la IA es un factor muy interesante a explorar en próximas investigaciones.

Podemos concluir que, en el Experimento 7, avisar a los participantes de la posible presencia de errores del sistema los llevó a ser menos proclives a confiar en la etiqueta de recomendación, y a detectar más los errores durante la Fase 1 de la tarea de clasificación, efecto que se manifestó en la condición falso positivo, como hemos explicado antes. Esta atenuación de la influencia de la recomendación sesgada en las respuestas de los participantes en la primera fase de la tarea, también se acompañó de una reducción de la herencia del sesgo en la fase posterior. Los participantes del grupo informado sobre el error también cometieron menos errores que el grupo control, para la condición falso positivo, en la fase de clasificación sin asistencia. Quienes evitaron adherirse a la recomendación sesgada para resolver la tarea de diagnóstico, evitaron a su vez la adquisición por aprendizaje del sesgo de la inteligencia artificial.

Como explicamos anteriormente, creemos que el efecto de la información sobre el error se manifestó en la condición falso positivo, pues es la que ofrecía margen para la mejora, ya que en la condición falso negativo los participantes ya tendían a seguir una estrategia cautelosa y crítica en la tarea de clasificación. Creemos que este hallazgo es muy positivo, pues indicaría que incluso una advertencia general y sencilla sobre posibles errores y sesgos en la IA fue de ayuda para prevenir el aprendizaje del sesgo de la IA durante la Fase 1 de la tarea de clasificación, y a consecuencia, la herencia del sesgo en la Fase 2 en aquella condición en la que los participantes eran menos críticos y tendían a cometer más errores.

En el Experimento 7, la confirmación de una tendencia en los participantes a acumular

más errores en la condición falso positivo resultó extremadamente relevante y puede ayudarnos a comprender las diferencias observadas en función del contrabalanceo de las instrucciones en los experimentos anteriores de la Serie Experimental 2. Quizá de forma general, en caso de duda los participantes prefirieron optar por la categoría positivo en B1, ya que en este caso un falso positivo sería la opción menos arriesgada, pero, sin embargo, evaluaron de forma más crítica la recomendación de la IA y cometieron menos errores en B2, ya que un falso negativo conllevaba mayor riesgo (ver Tabla 3B del Apéndice B para consultar los índices de error de cada uno de los estímulos 40/60 y 60/40 en función de la información del error y el tipo de error en los Experimentos 6 y 7 de la Serie Experimental 2).

En los cuatro primeros experimentos de la presente Tesis Doctoral no se detectó diferencia alguna en las respuestas de los participantes en función del contrabalanceo de las instrucciones. Podemos descartar con seguridad algún efecto de esta variable en los resultados de la Serie Experimental 1. El discutido extraño efecto del contrabalanceo comenzó a manifestarse en la Serie Experimental 2 tras algunos cambios en el diseño: primero, en la Serie 2 todos los participantes pasaron a completar la Fase 1 con asistencia de la IA, no se dio una división entre un grupo con asistencia y otro grupo sin asistencia; además, en la segunda serie experimental, algunos participantes recibieron información advirtiendo sobre errores en las recomendaciones de la supuesta inteligencia artificial. Este mensaje aumentó el número de participantes que fueron conscientes de la posible aparición de errores en las recomendaciones de la IA. Dado que la muestra total de participantes de cada uno de los experimentos de la Serie 2 fue asistida por la IA sesgada, también aumentó el número de personas que podían detectar errores en las recomendaciones, no solo en los grupos que eran avisados de este hecho, sino también en los grupos control. Es decir, que estas dos circunstancias pudieron haber aumentado en general la conciencia, por información u observación, sobre una IA imperfecta que podía errar. La facilidad para detectar errores, una vez los participantes estaban en guardia para encontrarlos, quizá no fue la misma en los dos

balanceos de instrucciones B1 y B2. En el Experimento 5 y en el Experimento 6, detectamos que aquellos participantes asignados a B2 detectaban con más frecuencia los errores presentes en las recomendaciones, al menos así lo declararon en las preguntas que realizamos al finalizar la tarea de clasificación, aunque no observamos esa misma tendencia en el Experimento 7. En principio, este resultado parece coherente con la hipótesis de que los participantes desplegaron un procesamiento de la información más controlado y analítico ante la posibilidad de un error falso positivo en B2.

Con respecto a los resultados del Experimento 5 y del Experimento 6, argumentamos que quizá el efecto del contrabalanceo pudo haber ejercido una cierta influencia extraña en los resultados, obstaculizando la posibilidad de extraer alguna conclusión acerca de si efectivamente funcionó nuestra manipulación experimental en estos dos experimentos, debido, por ejemplo, a una diferencia perceptiva entre las dos condiciones. Sin embargo, en el Experimento 7 observamos una mitigación de la influencia de las recomendaciones sesgadas de la IA y la herencia del sesgo cuando se mostró a los participantes un mensaje avisando de posibles errores al inicio de la tarea de clasificación. En el Experimento 7 utilizamos el mismo mensaje de error que en el Experimento 5, y si bien en el Experimento 5 no causó el efecto esperado, en el Experimento 7 sí mitigó la influencia del sesgo de la IA. La única diferencia entre ambos experimentos fue que en el Experimento 5 el sesgo de la IA fue un error sistemático en los estímulos 40/60, y en el Experimento 7 se trasladó a un error sistemático en los estímulos 60/40. Este cambio del sesgo permitiría comprobar el efecto de la manipulación principal, el mensaje de error, en las respuestas de los voluntarios a estímulos 60/40, disociado de la influencia de un posible efecto perceptivo asociado a los estímulos 40/60. Los resultados del Experimento 7 nos permiten concluir que el mensaje avisando de errores de la IA fue efectivo, sin sospechas de la interferencia de algún extraño efecto perceptivo no controlado. En este último experimento, además, incluimos como variable el tipo de error, y por tanto pasamos a controlar el posible efecto de este factor en el comportamiento de los

participantes.

En un momento como el actual, en el que se está creando una regulación legal para la inteligencia artificial, el diseño de estrategias que puedan facilitar una adecuada supervisión humana de la IA es más relevante que nunca. La Comisión Europea establece que los sistemas de IA que se implementen en sistemas de alto riesgo, como son los sistemas de apoyo a la decisión en el ámbito de la salud, no pueden funcionar de forma autónoma, deben estar bajo la vigilancia de un ser humano, en las diferentes fases de su ciclo de creación (Alkorta, 2022; European Commission, 2021; Lazcoz & de Hert, 2023). En la última fase de dicho ciclo, la implementación de la IA en entornos reales, la persona que ejerza dicha supervisión será quien vele por la adecuación y la ética de las decisiones que toma en colaboración con una inteligencia artificial, constituyéndose como una salvaguarda frente a errores y sesgos de la máquina. Probablemente en muchas situaciones estas personas serán la última línea de defensa frente a sesgos y otro tipo de comportamientos problemáticos de la inteligencia artificial. Sin embargo, a día de hoy no existen pautas concretas que especifiquen cómo exactamente se debe llevar a cabo tal supervisión y cómo se va a regular su cumplimiento. Las tentativas de legislación también plantean que la herramienta de IA debería ofrecer facilidad para que casi cualquier persona pudiera supervisar su rendimiento sin necesidad de una especialización en tal materia. La realidad es que aún se conoce muy poco acerca de cuáles pueden ser las condiciones más idóneas para fomentar una adecuada supervisión humana de la IA, y diseñar con ello protocolos y estrategias (Cabitza, Campagner, Ronzio, et al., 2023).

Con frecuencia, se ha propuesto que proporcionar explicaciones sobre cómo funciona la inteligencia artificia en general, o más específicamente explicaciones acerca de por qué la IA opta por una recomendación determinada dadas ciertas circunstancias, sería una estrategia adecuada para ayudar a los usuarios a calibrar qué confianza depositar en la IA (Ghassemi et al., 2021). Estas explicaciones podrían frenar la confianza excesiva en el algoritmo y reducir el

sesgo de automatización (Rastogi et al., 2022). Este sesgo cognitivo ha sido definido como la tendencia de las personas a utilizar las claves que proporciona la máquina como un heurístico que sustituye a una búsqueda y procesamiento de la información analítica y controlada (Skitka et al., 1999).

Los estudios que han puesto a prueba la efectividad de la estrategia de explicar el funcionamiento de la IA a los usuarios, es decir, aumentar su transparencia para mejorar la toma de decisiones en el equipo humano–IA, han arrojado resultados contradictorios. Si bien algunos estudios señalan que, efectivamente, estas estrategias podrían atenuar el sesgo de automatización bajo unas determinadas condiciones, otros también apuntan a que las explicaciones pueden incluso aumentar esa excesiva confianza en la inteligencia artificial, y por ende el sesgo de automatización (Bertrand et al., 2022; Vered et al., 2023). Especialmente, captó nuestra atención el reciente estudio de Vered et al. (2023) porque estos autores indicaron de forma explícita a sus participantes cuáles eran las limitaciones y potenciales fuentes de error de la IA. Aunque Vered et al. (2023) advirtieron que el consejo de la IA no siempre era correcto y los participantes podían modificarlo o contradecirlo, igualmente detectaron sesgo de automatización, y la explicación sobre por qué la IA había proporcionado una recomendación determinada no redujo esta tendencia, e incluso, en algunas circunstancias, la aumentó. Por otro lado, Vered et al. (2023) sí observaron una mitigación del sesgo de automatización en aquellos participantes que se llevaron una primera impresión negativa de la IA. Esto significa que quienes vieron a la IA cometer errores en los primeros ensayos en los que interactuaron con ella, tendieron a confiar menos en el sistema (Nourani et al., 2021). Deseamos hacer una mención especial al estudio de Jabbour et al. (2023) que evaluó el impacto negativo de una inteligencia artificial sesgada (cometía un error sistemático) en la precisión de los diagnósticos realizados por profesionales del ámbito clínico. En esta investigación también pusieron a prueba si las explicaciones sobre cómo funcionaba el modelo de IA mitigaría la influencia negativa del sesgo en los clínicos. En una muestra de participantes

que incluía médicos y enfermeras de diversos hospitales de Estados Unidos, Jabbour et al. (2023) observaron cómo la precisión en los diagnósticos de los clínicos decaía, en comparación a la línea de base, cuando la IA que aportaba predicciones estaba sesgada. Las explicaciones sobre el funcionamiento del modelo de IA no fueron capaces de mitigar este efecto negativo.

En nuestros experimentos, por el contrario, sí observamos una atenuación del impacto negativo de las recomendaciones sesgadas de la IA en las decisiones de los participantes cuando fueron previamente alertados sobre la posible ocurrencia de errores, bien de una forma general o bien de una forma específica, ofreciendo el dato exacto de porcentaje de error en la IA. Es importante señalar que estas intervenciones mitigaron también la herencia del sesgo de la inteligencia artificial. Quizá aquellos participantes alertados sobre errores en las recomendaciones de la IA desplegaron una serie de estrategias para mantener su atención en las muestras de tejido a clasificar, evitando dejarse orientar completamente por la etiqueta de recomendación durante la tarea de diagnóstico. Esta forma de actuar habría dificultado aprender de la inteligencia artificial y como consecuencia redujo el riesgo de los participantes de heredar su sesgo.

Probablemente no exista un único tipo de estrategia óptimo para ayudar a las personas a calibrar una adecuada confianza en la IA, como evidencia la incoherencia entre los resultados observados en la literatura. Quizá para que las explicaciones sobre el funcionamiento de la IA sean efectivas necesiten adaptar su contenido y formato a las características de los usuarios y al contexto de colaboración del equipo humano–IA (Hudson & Franklin, 2023; Rago et al., 2021). Si existe una conclusión clara que podemos extraer de la revisión de la literatura y de los resultados de la Serie Experimental 2 es que es importante continuar investigando. Es necesario conocer que factores determinan que un determinado tipo de información, o una determinada forma de mostrar la información a los usuarios, sea efectiva para mitigar la influencia negativa de las recomendaciones de la IA en las decisiones

de las personas. Aún sabemos poco sobre cuáles son las mejores condiciones para promover una colaboración efectiva entre IA y humanos, y la creciente aplicación de la inteligencia artificial a cada vez más contextos crea urgencia por resolver la ausencia de conocimiento empírico.

Nuestros experimentos representan una pequeña contribución al esfuerzo colectivo de investigadores de todo el mundo que tratan de subsanar esa carencia. Nosotras hemos observado que cuando se trata de proporcionar información sobre la precisión de una IA, las personas podrían ser capaces de interpretar mejor esa información, y calibrar mejor su confianza, cuando se enfatiza el porcentaje de errores del modelo. De forma similar, incluso una información más simple y general pero que ponga énfasis en la falibilidad de la IA, advirtiendo de que la IA puede estar sesgada o puede cometer errores, también disminuiría la tendencia de las personas a aceptar recomendaciones sesgadas de una IA, especialmente en situaciones en que estas personas tienden de base a ser menos críticas con la máquina como, por ejemplo, cuando perciben que las consecuencias de una posible decisión errónea del equipo humano-IA no son especialmente graves.

Capítulo 4: Discusión General y Conclusiones.

GENERAL DISCUSSION AND CONCLUSIONS

People, as we have observed, could in fact inherit artificial intelligence biases, and warning them about the risk of biases and errors in AI outcomes seems to exercise a moderate preventive effect on this phenomenon. This statement synthetizes the two main conclusions that we can extract from the two series of experiments in the present Thesis Dissertation.

The first main purpose of our research project was to explore whether an individual's decisions could be biased by the skewed recommendations of a fictitious AI advisor. In addition, and most importantly, we wanted to know whether people can inherit the bias of the AI and retain it after the end of their interaction with the biased AI system. We answered this question in our first series of experiments.

All four experiments in the first experimental series shared similarities in their design: participants had to imagine that they were doctors with the mission of classifying tissue samples as affected or not by a fictitious syndrome. Thus, the task simulated a medical diagnosis. A group of participants completed this task with the assistance of an AI system, while the other group performed the task without assistance. The artificial intelligence system that assisted our participants made a systematic error in the classification of tissue samples with a particular configuration of the cells of two colours presented in the tissue: the wrong recommendations were always made for the samples with 40% dark cells and 60% light cells (i.e., 40/60 stimuli).

Experiment 1 evidenced the vulnerability of human decisions to biased advice by a hypothetical AI model. Individuals' decisions were correct and unbiased in the control group, whereas the AI-assisted group made errors in the 40/60 samples. That is, their decisions were skewed in the same direction and in the same stimuli as AI recommendations.

One possible criticism of the results observed in Experiment 1 is that our task was a simulation and a simplification of a clinical decision-making process, very different from a real medical environment. In support of Experiment 1 results, it is important to mention that studies in real clinical settings have also found a loss in the accuracy of clinicians' decisions when they were advised by erroneous recommendations from an AI (Adam et al., 2022; Dratsch et al., 2023; Jabbour et al., 2023; Rezazade Mehrizi et al., 2023). Jabbour et al. (2023) conducted a randomized clinical vignette survey study in which physicians and nurses from different hospitals at Unites States took part, to test the impact of a biased AI on clinicians' diagnostic accuracy. These authors found that diagnostic accuracy decreased in comparison with the base-line performance when clinicians were shown AI biased information. Therefore, similar results have been observed in controlled laboratory experiments as well as in studies of greater ecological validity, which provides strong support for Experiment 1 findings.

Experiment 2 and Experiment 3 added a new phase to the essential design of Experiment 1. In the second phase of Experiment 2 both groups, AI-assisted and unassisted, had to perform the same classification task without assistance. Thus, participants in the AI-assisted group initially performed the task with the AI recommendations but later, they faced the same classification task without any assistance. The previously AI-assisted participants still reproduced the same mistakes as the AI made in the previous phase, that is incorrect classification of 40/60 samples, when they had to make decisions on their own. Thus, those participants mimicked AI bias even when they changed to a context in which AI advice was absent. Furthermore, the AI bias of the previous phase also influenced how participants

classified new and ambiguous stimuli in the second phase of the task. These results showed that volunteers inherited the AI bias and pointed out a potential generalization of this bias to decisions in new and uncertain cases. The finding of an AI inherited bias effect represented the most relevant and novel contribution of our series of experiments. To our knowledge, this effect had not previously been tested nor documented in any empirical research.

In addition, Experiment 3 also showed that increasing the experience with the task did not prevent participants from absorbing AI bias, since the group that performed the task unassisted during the first phase and was AI-assisted in the second phase was equally influenced by the biased recommendations as those participants that encountered the artificial advisor in the first phase. Experiment 4 replicated the same results as Experiment 2 showing the inheritance of the AI bias in a sample of medical students. The participants that took part in Experiment 1 were psychology students, while the sample in Experiments 2 and 3 was a general and diverse population recruited through Prolific. We can conclude that our main results are robust as we detected the same effect in four experiments using different population samples.

After summarising the main results of the first experimental series, we discussed some psychological processes that might underlie the inherited bias effect observed in the first four experiments (see Discussion section of the Experimental Series 1). Arguments grounded on the two-systems model of cognitive processes could be useful to understand human compliance with imperfect AI systems (Buçinca et al., 2021; Vered et al., 2023). This theory postulates the existence of two types of cognitive processing systems: System 1, specialised in fast, automatic, associative and effortless operations, is responsible for the use of heuristics or simple rules to solve complex problems, and System 2, characterised for deliberate, analytical, flexible and effortful operations (Kahneman, 2003). In our experiments, we can interpret that the AI recommendation label would have functioned as a heuristic solution to avoid a

controlled and effortful processing of tissue samples information. Opting for this automatic processing in Phase 1 of the classification task would have difficulted to regain a mode of controlled processing in the unassisted phase of the task, Phase 2, impairing participants' performance. In addition, we considered another explanation, grounded in conditioning and learning principles, to our results: we suggested that the AI recommendation label, a more salient stimuli, could have overshadowed the processing and learning on the tissue samples, a comparatively more complex stimulus.

However, we concluded that both mechanisms fall short of explaining the inheritance and generalisation of the AI bias, and we posited that participants learned the AI bias, a theory that provides a more complete explanation for these phenomena. Although participants were instructed on the correct classification criteria at the beginning of the experiments, they learned from the AI system how to classify tissue samples during Phase 1, modifying their decision-making strategy to adapt to the AI suggestions. This would explain why in Phase 2 participants replicated exactly the same type of systematic errors that the AI made in the previous phase and its generalisation to ambiguous stimuli. Following the prediction of the heuristic or overshadowing theories, during Phase 2 participants would have made errors in the classification, but they would have improved after a certain number of trials, that is, we would have observed a progressive reduction in their errors. In addition, the error would have not generalized to novel stimuli if it had not been learned. Therefore, a learning of AI bias better explains the absence of a decay in the average number of 40/60 errors throughout the unassisted phase of the task observed in Experiment 3 (which has a longer Phase 2 than Experiments 2 and 4) and the generalization of the error to the novel 50/50 stimuli. Summing up our interpretation of the results, volunteers learnt the AI bias and applied it even when the advice of the AI was no longer present. This has very serious implications for any situation in which humans interact with AI and can learn from it.

Related to the above explanation, in a case study, Kwong et al. (2024) reported a similar phenomenon to the one we observed in our series of experiments. The authors were testing the accuracy of an AI model for diagnostic assistance in a real clinical setting. However, they found that after interacting with the AI model, clinicians changed their decision-making strategy to match the strategy followed by the AI in its predictions. This case study also supports our interpretation that people can learn from AI as if it were a mentor, acquiring strategies they will replicate in their future decision-making.

Once the influence of biased AI recommendations on human decisions was evidenced, the next step was to analyse whether informing participants about the risk of error in AI recommendations could mitigate the system's influence on their decisions, and, specifically, on the inheritance of the AI systematic error. This represented the second main purpose of the present Thesis. In a series of three experiments (Experimental Series 2), we tested whether a brief and simple message warning of possible errors or biases in artificial intelligence, in Experiments 5 and 7, or communicating the specific percentage of error in AI recommendations, in Experiment 6, could prevent people from uncritically adhering to the advice from the simulated artificial agent. Experiment 7 was also an opportunity to explore the influence of the cost of the error (i.e., the consequence of following an inadequate AI recommendation implying a false positive vs. false negative error) on the decision to accept the model's advice.

In Experiment 5, one group of participants were warned about possible AI errors at the beginning of the experiment, before they started to classify tissue samples, while the control group did not receive such a message. Contrary to our expectations, we did not find differences between both groups but we believe that this was because the differences in participants behaviour introduced by the counterbalanced instructions in this experiment interfered with the effect of the experimental manipulation (which is why we decided to test it

again in Experiment 7). A more specific message, informing about 20% of errors in the recommendations made by the AI, reduced the tendency of the informed group to follow those inaccurate recommendations labels in comparison with the control group in Experiment 6. Informed participants made fewer incorrect classifications of 40/60 samples than those uninformed in the first AI-assisted phase, and also in the second unassisted phase of the experimental task. This result seems to indicate that the message warning about possible errors had a preventive effect on the inheritance of the AI bias. However, informing participants that the AI made a ratio of 80% correct recommendations did not prevent them from following the skewed advice of the AI system, without differences with the control condition without information. In Experiment 6 the same accuracy information about AI outcomes framed in two different versions, emphasising error (20% of wrong recommendations) or correctness (80% of correct recommendations), showed a different efficacy to mitigate the impact of AI-biased advice on participants' decisions. The framing effect is a robust result in many areas of psychology (Kahneman, 2003; Tversky & Kahneman, 1981) and our finding suggest that how data about AI accuracy is framed would be relevant in determining whether people will be motivated or not to be more cautious when collaborating with an IA system facilitating the detection of its potential biases.

In Experiment 7, we meant to test the robustness of the inherited bias effect and the influence of the error warning by changing the AI bias to an incorrect recommendation always in 60/40 samples, instead of the 40/60 samples of the previous experiments. The change of the bias to 60/40 samples also implied a dissociation of the effect of the experimental manipulation from a possible extraneous perceptual effect associated with the 40/60 samples. Similarly, this change involved a switch in the meaning of the error associated with incorrect AI advice in each of the two different, but complementary, sets of instructions used in the experiments. In the B1 instructions, the classification criteria stated that a higher proportion of dark-coloured cells in the tissue sample meant that it should be classified as positive (and a

greater proportion of light-coloured cells was a negative). In contrast, for classification criterion B2 a higher proportion of light-coloured cells signalled that the tissue sample should be classified as a positive (and tissue sample with a greater proportion of dark-coloured cells would be classified as a negative). Thus, in B1, following the incorrect AI advice for 60/40 samples implied making a false negative error in Experiment 7, when in previous experiments it had been a false positive for 40/60 samples. In B2 the IA recommendation indicated a false positive for the 60/40 samples, when it was a false negative for 40/60 in previous experiments. In conclusion, there was a reversal in the type of error associated with each set of instruction in Experiment 7 to explore the effect of each variable (an effect of B1 and B2 classification criteria versus an influence of the cost of the error) separately.

In the context of our experiments the cost of the error was different for a false positive versus a false negative. Participants had to detect whether a particular disease was present by observing tissue samples from different patients. Failing to detect a patient affected by the disease, that is, misclassifying a positive tissue sample as negative (i.e. false negative) has more serious consequences, as this patient would not receive any treatment, than the opposite error. In comparison, diagnosing a given patient as affected by a given syndrome that is not present (i.e. false positive) would be perceived as less risky. We expected participants to follow the biased AI-suggestions more in the false positive condition, making more errors, than in the false negative condition (regardless of the association of type of error with instruction set B1 or B2, variable that we did not expect to influence participants' responses).

Interestingly, results from Experiment 7 suggest that the cost of the potential AI error influences participants' tendency to follow AI's suggestions and modulates the impact of the message alerting about potential errors in the AI model. When the AI recommendation labels offered false positives in 60/40 samples, participants incorrectly classified those samples more frequently. Thus, we believe they were more prone to follow AI advice in the false positive bias

condition. Consequently, informing volunteers about possible AI mistakes had an impact on the false positive condition. When the AI systematic error was a false positive, the group of participants with awareness of possible AI errors made fewer mistakes in 60/40 samples than those volunteers who were not alerted about the possibility of bias. When the AI was skewed toward false negatives, volunteers' responses were more cautious, committing fewer errors. Thus, the difference between informed and non-informed participants was not observed when the AI bias was oriented to false negatives and the participant's decisions already tended to be more conservative. The results from Experiment 7 support the interpretation that participants would perceive the misidentification of a healthy patient as a patient suffering from the syndrome as less risky than the opposite mistake of failing to detect an affected patient. Summing up, the results of our second series of experiments highlight that under certain circumstances people could be less prone to trust AI and to conform to its inaccurate suggestions.

From the findings of our second series of experiments, Experiments 5, 6 and 7, we can derive some recommendations to reduce people's compliance with AI and the risk of inheriting AI biases. Developers of artificial intelligence systems should responsibly communicate which realistic outcomes the model can achieve as users need clear and accurate information about the reliability of the tool they are going to work with (Mullainathan & Obermeyer, 2017). Expectations about the AI results shape the trust people place in the model. Although people take into consideration not only the information provided to them but also their interactions with the AI to decide the level of confidence they place in this advisor, it would be crucial to adequately convey the risk of error on AI predictions to help users to correctly calibrate their trust in the machine (Yin et al., 2019).

In the Introduction section, we mentioned the hope that lies in the synergy of the human-AI team to achieve better decisions than those each agent could have made on their

own. However, the excessive trust that humans often place in AI predictions is an obstacle to such promise. Interaction with artificial intelligence systems may also activate cognitive biases in the individuals which reduce the rationality of the decisions made by this team (Rastogi et al., 2022). Instead of combining their inferences with the artificial agent's results and modifying or rejecting inaccurate AI's recommendation, people tend to uncritically accept the model's advice, even when it is noticeably erroneous. In our experiments, those participants who perceived the AI as more trustworthy tended to follow more its incorrect advice. Although approximately 80% of participants in our experiments detected errors in AI suggestions, we still found an influence of the biased recommendations and inheritance of AI bias on those participant's decisions. Such overreliance has been evidenced in multiple studies that have pointed to the action of an authority effect in interaction with artificial intelligence algorithms as one possible cause (Baudel et al., 2020).

It is believed that explanations about how the AI works or how this agent has arrived at a certain result could help people identify when the algorithm's reasoning has been flawed, and mitigate their overreliance in artificial intelligence. But the reality is that such explanations have mixed results, sometimes improving people's ability to detect and correct inaccurate AI outcomes, and sometimes worsening it (Buçinca et al., 2021; Cabitza, Campagner, Angius, et al., 2023; Jabbour et al., 2023; Vered et al., 2023). It should be considered that the cognitive biases of users as, for example, the framing effect on AI we have observed in Experiment 6, can also influence the interpretation of information about the AI performance, its limitations and risk of error.

Dual-process theories of information processing can help to understand why people uncritically accept biased AI recommendations, and why alerting them about possible errors does not always have the expected effect. Humans would be expected to analytically process each AI recommendation as well as the explanation of how the AI has arrived at a particular

decision. However, this requires high cognitive effort, which humans often find aversive (Kool & Botvinick, 2018; Lyell & Coiera, 2017). Instead of assessing how much trust to place in the AI in each case, people seem to develop heuristics about the system's competence to determine whether or not to follow its advice in general (Buçinca et al., 2021). Similarly, the success of explanations about AI processes or limitations on reducing users' overreliance is mediated by the effort required to verify the correctness of a determined algorithm's reasoning in such explanation (Vasconcelos et al., 2023). Sometimes, providing information about how the AI algorithm works or data about its reliability induces an image of competence which can elicit an excessive trust in the AI. This might be the case for the group of participants informed of 80% of correct recommendations by the AI system in Experiment 6.

Cognitive interventions to motivate the analytical processing of AI recommendations could help users properly calibrate AI trust (Buçinca et al., 2021). Research on reasoning and decision-making processes could be useful to enhance the effectiveness of strategies that attempt to mitigate the influence of AI on human decisions. Thus, the way in which AI outputs, recommendations or assessments are presented to users is important to trigger, or counteract, human´s cognitive bias. For example, it could be better to present AI advice once people have processed information and made their own judgment to avoid the action of anchoring bias. This bias has been demonstrated to impair decision accuracy and the ability of humans to override wrong information in case of AI incorrect advice (Agudo et al., 2024). The same principle should be applied to explanations of AI limitations that should be designed to counteract the action of cognitive biases (Bertrand et al., 2022). In Experiment 6, a simple intervention such as presenting information about AI accuracy focused on its error rate (20% of wrong recommendations) was more effective in mitigating the influence of biased AI recommendations on participants' decisions than presenting the same information with an emphasis on the ratio of correct recommendations (80%). These results show that the way in which information is framed either promote or diminish an adequate human oversight of AI

outcomes. Similarly, the results of Experiment 7 suggest that making people more aware of the consequences of a wrong decision could prevent them from blindly following biased AI advice, although we still have a limited understanding of the effect of the cost of the decision on people's adherence to AI, as pointed out by Fahnenstich et al. (2024). In sum, as the results of this Thesis Dissertation highlight, research into the details of the psychological aspects of human-AI interaction is necessary to achieve the best possible benefits from the use of artificial intelligence tools.

Artificial intelligence is a human creation that should be prepared to deal with a biased world. Therefore, it is necessary to take measures to reduce the risk of bias in artificial intelligence models during its development. For example, it would be necessary to ensure that the data samples used to train the algorithms are representative and have adequate predictor variables (Buolamwini & Gebru, 2018; Larrazabal et al., 2020; Parikh et al., 2019). Beyond the control of data sets during the process of creating a new AI model, these tools also require strict evaluation and validation of results in controlled studies and real-world settings (Thomasian et al., 2021). Reliable monitoring mechanisms are needed to audit artificial intelligence algorithms (Mehrabi et al., 2022). To this end, it has even been proposed to develop artificial intelligence tools to detect and quantify biases in both data and in the predictions and decisions of other AI algorithms (Kleinberg et al., 2020; Zou & Schiebinger, 2018). We believe that the development of technological solutions is an incomplete answer to the problem of the dangerous circular propagation of biases from humans to artificial intelligence and the other way around. The auditing of AI tools needs to be complemented with extensive research on how humans interact with such tools. No matter how many precautions are taken during AI design and training, expecting an algorithm to remain unbiased when introduced in a biased real-world context would be unrealistic. Consequently, in certain cases, biases in AI may be difficult to control and prevent by technological means,

leaving the human watcher as the last line of defence to detect and correct such biases (Mittermaier et al., 2023).

To this end, the human supervisor must be alerted of the risk of bias in AI results to detect the systematic error and correct it. Therefore, rigorous research on how humans react to and are influenced by AI bias and how this problem may be mitigated is necessary. As our first series of experiments showed (Experiments 1, 2, 3 and 4), to rely on humans to critically evaluate AI's results, reject doubtful information and contradict AI suggestions a priori seems unwise. For humans to became effective bias detectors, they would need to be previously trained in the identification of the risk of bias in AI tools and be motivated to undertake such task. The results of Experiments 6 and 7 give us some hope, showing that people, once alerted of the risk of bias in artificial intelligence, might engage with the model in a more cautious and critical way.

Importantly, AI-based tools intended to be applied in sensitive contexts, as healthcare, should be thoroughly evaluated and monitored, which requires a regulation to set standards for their acceptance and implementation in real-world environments (Alkorta, 2022). It has been argued that artificial intelligence should be regulated by ethical codes, as other disciplines are, and ethics committees should enforce compliance with these codes ensuring sanctions for violators (Véliz, 2019). In this context, transparency would be essential to facilitate the monitoring of AI tools and ensure their compliance with the principles of ethical codes and regulations. Strategies such as open data and code sharing would facilitate the auditing of new technological developments and the validation of AI in different contexts (Norori et al., 2021; Rajpurkar et al., 2022). If artificial intelligence development follow ethical principles, it has the potential to become a force for good, for example, helping to combat prejudice if AI is used to compensate for and reduce human biases and historical inequalities

between population subgroups (Chen et al., 2020; Obermeyer et al., 2019; Pierson et al., 2021).

However, for AI to be beneficial, and not pose a risk to humans, it is not enough just to have an ethical code of conduct. More research is essential. It is necessary to investigate in detail in which contexts there is a greater vulnerability for humans to show overreliance on AI systems, which are the most appropriate formats to convey information about AI based tools to the user, at what point it is appropriate to show the AI suggestion to the human in a collaboration protocol, to mention just a few issues among a myriad of processes that are still unknown. The present Doctoral Thesis principally sought to highlight the importance of a deep understanding of the psychological processes that develop in people interacting with an agent described as artificial intelligence. Cognitive processes are involved in how people relate to machines. In a world in which AI will almost inevitably accompany human beings in most of their professional and domestic tasks, it is necessary to know what these tools mean to people, what influence AI algorithms can exert on their behaviour, and even what potential this technology has to modify their beliefs about the world. As our results showed, interaction with artificial intelligence tools can have an impact on how people process information, reason and make decisions. The most relevant contribution of our research has been to evidence that this impact can persist even time after the collaboration with the artificial intelligence has ended.

Referencias

Abd-Alrazaq, A., AlSaad, R., Shuweihdi, F., Ahmed, A., Aziz, S., & Sheikh, J. (2023). Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. *npj Digital Medicine, 6,* Article 84. https://doi.org/10.1038/s41746-023-00828-5

Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. *Nature Machine Intelligence, 3*(6), 461–463. https://doi.org/10.1038/s42256-021-00359-2

Adam, H., Balagopalan, A., Alsentzer, E., Christia, F., & Ghassemi, M. (2022). Mitigating the impact of biased artificial intelligence in emergency decision-making. *Communications Medicine, 2,* Article 149. https://doi.org/10.1038/s43856-022-00214-4

Adlung, L., Cohen, Y., Mor, U., & Elinav, E. (2021). Machine learning in clinical decision making. *Med, 2*(6), 642–665. https://doi.org/10.1016/j.medj.2021.04.006

Agudo, U., Arrese, M., Liberal, K. G., & Matute, H. (2022). Assessing emotion and sensitivity of AI artwork. *Frontiers in Psychology, 13,* Article 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*, Article 1. https://doi.org/10.1186/s41235-023-00529-3

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

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken,

H., Schlobach, S., van der Gaag, L., van Harmelen, F., … Welling, M. (2020). A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer, 53*(8), 18–28. https://doi.org/10.1109/MC.2020.2996587

Alaa, A. M., Gurdasani, D., Harris, A. L., Rashbass, J., & van der Schaar, M. (2021). Machine learning to guide the use of adjuvant therapies for breast cancer. *Nature Machine Intelligence, 3*(8), 716–726. https://doi.org/10.1038/s42256-021-00353-8

Alkorta, I. (2022). Five crucial challenges for regulation of medical artificial intelligence. *Revista Internacional de Los Estudios Vascos, 67*(2), 31–40. https://www.eusko-ikaskuntza.eus/es/riev/five-crucial-challenges-for-regulation-of-medical-artificial-intelligence/rart-24784/

Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022). Fairness and explanation in AI-informed decision making. *Machine Learning and Knowledge Extraction, 4*(2), 556–579. https://doi.org/10.3390/make4020026

Arasu, V. A., Habel, L. A., Achacoso, N. S., Buist, D. S. M., Cord, J. B., Esserman, L. J., Hylton, N. M., Glymour, M. M., Kornak, J., Kushi, L. H., Lewis, D. A., Liu, V. X., Lydon, C. M., Miglioretti, D. L., Navarro, D. A., Pu, A., Shen, L., Sieh, W., Yoon, H.-C., & Lee, C. (2023). Comparison of mammography AI algorithms with a clinical risk model for 5-year breast cancer risk prediction: An observational study. *Radiology, 307*(5), Article e222733. https://doi.org/10.1148/radiol.222733

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

Ashoori, M., & Weisz, J. D. (2019). *In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes.* ArXiv https://doi.org/10.48550/arXiv.1912.02675

Banerji, C. R. S., Chakraborti, T., Harbron, C., & MacArthur, B. D. (2023). Clinical AI tools must convey predictive uncertainty for each individual patient. *Nature Medicine, 29*(12), 2996–2998. https://doi.org/10.1038/s41591-023-02562-7

Baudel, T., Verbockhaven, M., Roy, G., Cousergue, V., & Laarach, R. (2020). *Addressing cognitive biases in augmented business decision systems*. ArXiv. https://doi.org/10.48550/arXiv.2009.08127

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve—Algorithm aversion and demonstrating the ability to learn. *Business and Information Systems Engineering, 63*(1), 55–68. https://doi.org/10.1007/s12599-020-00678-5

Berthet, V. (2022). The impact of cognitive biases on professionals' decision-making: A review of four occupational areas. *Frontiers in Psychology, 4*(12), Article 802439. https://doi.org/10.3389/fpsyg.2021.802439

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How cognitive biases affect XAI-assisted decision-making. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, USA*, 78–91. https://doi.org/10.1145/3514094.3534164

Blanco, F., Moreno-Fernández, M. M., & Matute, H. (2020). Are the symptoms really remitting? How the subjective interpretation of outcomes can produce an illusion of causality. *Judgment and Decision Making, 15*(4), 572–585. https://doi.org/10.1017/S1930297500007506

Blumenthal-Barby, J. S., & Krieger, H. (2015). Cognitive biases and heuristics in medical decision making. *Medical Decision Making, 35*(4), 539–557. https://doi.org/10.1177/0272989X14547740

Boden, M. A. (2017) *Inteligencia Artificial.* Turner Noema

Bogert, E., Lauharatanahirun, N., & Schecter, A. (2022). Human preferences toward algorithmic advice in a word association task. *Scientific Reports, 12*, Article 14501. https://doi.org/10.1038/s41598-022-18638-2

Bogert, E., Schecter, A., & Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific Reports, 11*, Article 8028. https://doi.org/10.1038/s41598-021-87480-9

Böhnke, J., Varghese, J., Karch, A., Rübsamen, N., Bode, L., Mast, M., Marschollek, M., Schamer, S., Rathert, H., Jack, T., Beerbaum, P., Rübsamen, N., Böhnke, J., Karch, A., Das, P. P., Wiese, L., Groszweski-Anders, C., Haller, A., Frank, T., & Wulff, A. (2022). Systematic review identifies deficiencies in reporting of diagnostic test accuracy among clinical decision support systems. *Journal of Clinical Epidemiology, 151*, 171–184. https://doi.org/10.1016/j.jclinepi.2022.08.003

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to computer programmer as woman is to homemaker? debiasing word embeddings.* ArXiv. https://doi.org/10.48550/arXiv.1607.06520

Bonezzi, A., & Ostinelli, M. (2021). Can algorithms legitimize discrimination? *Journal of Experimental Psychology: Applied, 27*(2), 447–459. https://doi.org/10.1037/xap0000294

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2021). Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics, 47*(12), e3. https://doi.org/10.1136/medethics-2019-105860

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of*

the ACM on Human-Computer Interaction, 5, Article 188. https://doi.org/10.1145/3449287

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81,* 77–91. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making, 33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Cabitza, F., Campagner, A., Angius, R., Natali, C., & Reverberi, C. (2023). AI Shall Have No Dominion: How to measure technology dominance in AI-supported human decision-making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3544548.3581095

Cabitza, F., Campagner, A., Ronzio, L., Cameli, M., Mandoli, G. E., Pastore, M. C., Sconfienza, L. M., Folgado, D., Barandas, M., & Gamboa, H. (2023). Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis. *Artificial Intelligence in Medicine, 138,* Article 102506. https://doi.org/10.1016/j.artmed.2023.102506

Cabitza, F., Campagner, A., & Sconfienza, L. M. (2021). Studying human-AI collaboration protocols: the case of the Kasparov's law in radiological double reading. *Health Information Science and Systems*, *9*(1), 1–20. https://doi.org/10.1007/s13755-021-00138-8

Cabitza, F., Ciucci, D., & Rasoini, R. (2019). A Giant with Feet of Clay: On the Validity of the Data that Feed Machine Learning in Medicine. *Lecture Notes in Information Systems and Organisation*, *28*, 121–136. https://doi.org/10.1007/978-3-319-90503-7_10

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA, 318*(6), 517–518. https://doi.org/10.1001/jama.2017.7797

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour, 5*(12), 1636–1642. https://doi.org/10.1038/s41562-021-01146-0

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Carlucci Aiello, L. (2016). The multifaceted impact of Ada Lovelace in the digital age. *Artificial Intelligence, 235,* 58–62. https://doi.org/10.1016/j.artint.2016.02.003

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

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety, 28*(3), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

Chamizo, V. D., Aznar-Casanova, J. A., & Artigas, A. A. (2003). Human overshadowing in a virtual pool: Simple guidance is a good competitor against locale learning. *Learning and Motivation, 34*(3), 262–281. https://doi.org/10.1016/S0023-9690(03)00020-1

Chen, I., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory? *Advances in Neural Information Processing Systems, Canada,* 3543–3554. https://doi.org/10.48550/arXiv.1805.12002

Chen, I. Y., Joshi, S., & Ghassemi, M. (2020). Treating health disparities with artificial intelligence. *Nature Medicine, 26*(1), 16–17. https://doi.org/10.1038/s41591-019-0649-2

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science, 4*(1), 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757

Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care? *AMA Journal of Ethics, 21*(2), 167–179. https://doi.org/10.1001/amajethics.2019.167

Chen, V., Liao, Q. V., Wortman Vaughan, J., & Bansal, G. (2023). Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction*, *7*, 1–32. https://doi.org/10.1145/3610219

Claudy, M. C., Aquino, K., & Graso, M. (2022). Artificial intelligence can't be charmed: The effects of impartiality on laypeople's algorithmic preferences. *Frontiers in Psychology, 13,* Article 898027. https://doi.org/10.3389/fpsyg.2022.898027

Courtland, R. (2018). Bias detectives: the researchers striving to make algorithms fair. *Nature, 558*(7710), 357–360. https://doi.org/10.1038/d41586-018-05469-3

Croskerry, P., Campbell, S. G., & Petrie, D. A. (2023). The challenge of cognitive science for medical diagnosis. *Cognitive Research: Principles and Implications, 8,* Article 13. https://doi.org/10.1186/s41235-022-00460-z

Crowley, R. S., Legowski, E., Medvedeva, O., Reitmeyer, K., Tseytlin, E., Castine, M., Jukic, D., & Mello-Thoms, C. (2013). Automated detection of heuristics and biases among pathologists in a computer-based system. *Advances in Health Sciences Education, 18*(3), 343–363. https://doi.org/10.1007/s10459-012-9374-z

Dai, L., Sheng, B., Chen, T., Wu, Q., Liu, R., Cai, C., Wu, L., Yang, D., Hamzah, H., Liu, Y., Wang, X., Guan, Z., Yu, S., Li, T., Tang, Z., Ran, A., Che, H., Chen, H., Zheng, Y., … Jia, W. (2024). A

deep learning system for predicting time to progression of diabetic retinopathy. *Nature Medicine.* https://doi.org/10.1038/s41591-023-02702-z

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Australia,* 4691–4697. https://doi.org/10.24963/ijcai.2017/654

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

Dawood, M., Vu, Q. D., Young, L. S., Branson, K., Jones, L., Rajpoot, N., & Minhas, F. ul A. A. (2024). Cancer drug sensitivity prediction from routine histology images. *npj Precision Oncology, 8,* Article 5. https://doi.org/10.1038/s41698-023-00491-9

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, USA,* 1–12. https://doi.org/10.1145/3313831.3376638

de Miguel, I. (2020). Should we have a right to refuse diagnostics and treatment planning by artificial intelligence? *Medicine, Health Care and Philosophy*, *23*(2), 247–252. https://doi.org/10.1007/s11019-020-09939-2

de Miguel, I., Sanz, B., & Lazcoz, G. (2020). Machine learning in the EU health care context: exploring the ethical, legal and social issues. *Information, Communication & Society*, *23*(8), 1139–1153. https://doi.org/10.1080/1369118X.2020.1719185

DeCamp, M., & Lindvall, C. (2020). Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association, 27*(12), 2020–2023. https://doi.org/10.1093/jamia/ocaa094

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering, 61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dobler, C. C., Morrow, A. S., & Kamath, C. C. (2019). Clinicians' cognitive biases: A potential barrier to implementation of evidence-based clinical practice. *BMJ Evidence-Based Medicine, 24*(4), 137–140. https://doi.org/10.1136/bmjebm-2018-111074

Domjan, M. (2016) *Principios de aprendizaje y conducta* (7ª ed.) Cengage Learning Editores.

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE*, *18*(3), Article e0279720. https://doi.org/10.1371/journal.pone.0279720

Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., & Pinto dos Santos, D. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology, 307*(4), Article e222176. https://doi.org/10.1148/radiol.222176

Drew, T., Võ, M. L. H., & Wolfe, J. M. (2013). The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological Science, 24*(9), 1848–1853. https://doi.org/10.1177/0956797613479386

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

Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences of the United States of America, 112*(33), E4512–E4521. https://doi.org/10.1073/pnas.1419828112

Epstein, R., & Robertson, R. E. (2017). Suppressing the Search Engine Manipulation Effect (SEME). *Proceedings of the ACM on Human-Computer Interaction*, *1*, 1–22. https://doi.org/10.1145/3134677

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. https://doi.org/10.1038/nature21056

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

European Commission. (2021). *Proposal for a regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. COM (2021) 206 final.* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process theories of higher cognition. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fahnenstich, H., Rieger, T., & Roesler, E. (2024). Trusting under risk – comparing human to AI decision support agents. *Computers in Human Behavior*, *153*, Article 108107. https://doi.org/10.1016/j.chb.2023.108107

Famiglini, L., Campagner, A., Barandas, M., La Maida, G. A., Gallazzi, E., & Cabitza, F. (2024). Evidence-based XAI: An empirical approach to design more effective and explainable decision support systems. *Computers in Biology and Medicine*, *170*, Article 108042. https://doi.org/10.1016/j.compbiomed.2024.108042

Fawver, B., Thomas, J. L., Drew, T., Mills, M. K., Auffermann, W. F., Lohse, K. R., & Williams, A. M. (2020). Seeing isn't necessarily believing: Misleading contextual information influences perceptual-cognitive bias in radiologists. *Journal of Experimental Psychology: Applied, 26*(4), 579–592. https://doi.org/10.1037/xap0000274

Fernández-Aguilar, C., Martín-Martín, J. J., Minué Lorenzo, S., & Fernández Ajuria, A. (2022). Use of heuristics during the clinical decision process from family care physicians in real conditions. *Journal of Evaluation in Clinical Practice, 28*(1), 135–141. https://doi.org/10.1111/jep.13608

Fletcher, R. R., Nakeshimana, A., & Olubeko, O. (2021). Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Frontiers in Artificial Intelligence, 15*(3), Article 561802. https://doi.org/10.3389/frai.2020.561802

Franklin, M., Awad, E., & Lagnado, D. (2021). Blaming automated vehicles in difficult situations. *IScience*, *24*(4), Article 102252. https://doi.org/10.1016/j.isci.2021.102252

Fry, H. (2018). *Hello world: Being human in the age of algorithms.* W. W. Norton & Company.

Garcia-Vidal, C., Sanjuan, G., Puerta-Alcalde, P., Moreno-García, E., & Soriano, A. (2019). Artificial intelligence to support clinical decision-making processes. *eBioMedicine, 46,* 27–29. https://doi.org/10.1016/j.ebiom.2019.07.019

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI say: susceptibility in deployment of

clinical decision-aids. *npj Digital Medicine, 4*(1), Article 31. https://doi.org/10.1038/s41746-021-00385-9

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Gigerenzer, G. (2001). The Adaptive Toolbox. In Gigerenzer G. & Selten R. (Eds.), *Bounded Rationality: The Adaptive Toolbox* (pp. 36–50). The MIT Press.

Glocker, B., Jones, C., Roschewitz, M., & Winzeck, S. (2023). Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence*, *5*(6), Article e230060 https://doi.org/10.1148/ryai.230060

Goddard, K., Roudsari, A., & Wyatt, J. (2011). Automation bias - A hidden issue for clinical decision support system use. *Studies in Health Technology and Informatics, 164*, 17–22. https://doi.org/10.3233/978-1-60750-709-3-17

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics, 83*(5), 368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine, 165*(13), 1493–1499. https://doi.org/10.1001/archinte.165.13.1493

Graber, M. L., Kissam, S., Payne, V. L., Meyer, A. N. D., Sorensen, A., Lenfestey, N., Tant, E., Henriksen, K., LaBresh, K., & Singh, H. (2012). Cognitive interventions to reduce

diagnostic error: A narrative review. *BMJ Quality and Safety, 21*(7), 535–557. https://doi.org/10.1136/bmjqs-2011-000149

Grote, T., & Keeling, G. (2022). On algorithmic fairness in medical practice. *Cambridge Quarterly of Healthcare Ethics, 31*(1), 83–94. https://doi.org/10.1017/S0963180121000839

Han, R., Acosta, N. J., Shakeri, Z., Ioannidis, J. P. A., Topol, E. J., & Rajpurkar, P. (2023). *Randomized controlled trials evaluating AI in clinical practice: a scoping evaluation.* medRxiv. https://doi.org/10.1101/2023.09.12.23295381

Hastings, J. (2024). Preventing harm from non-conscious bias in medical generative AI. *The Lancet Digital Health, 6*(1), e2–e3. https://doi.org/10.1016/S2589-7500(23)00246-7

Harbarth, L., Gößwein, E., Bodemer, D., & Schnaubert, L. (2024). (Over)trusting AI recommendations: How system and person variables affect dimensions of complacency. *International Journal of Human–Computer Interaction*, 1–20. https://doi.org/10.1080/10447318.2023.2301250

High-Level Expert Group on AI (2019a). *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

High-Level Expert Group on AI (2019b). *A definition of AI: Main capabilities and scientific disciplines*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hinton, G. (2018). Deep learning—A technology with the potential to transform health care. *JAMA, 320*(11), 1101–1102. https://doi.org/10.1001/jama.2018.11100

Hollister, B., & Bonham, V. L. (2018). Should electronic health record-derived social and behavioral data be used in precision medicine research? *AMA Journal of Ethics, 20*(9), 873–880. https://doi.org/10.1001/amajethics.2018.873

Howard, J. J., Rabbitt, L. R., & Sirotin, Y. B. (2020). Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making. *PloS ONE, 15*(8), Article e0237855. https://doi.org/10.1371/journal.pone.0237855

Hudson, S., & Franklin, M. (2023). *Science Communications for Explainable Artificial Intelligence*. ArXiv https://doi.org/https://doi.org/10.48550/arXiv.2308.16377

Jabbour, S., Fouhey, D., Shepard, S., Valley, T. S., Kazerooni, E. A., Banovic, N., Wiens, J., & Sjoding, M. W. (2023). Measuring the impact of AI in the diagnosis of hospitalized patients. *JAMA, 330*(23), 2275–2284. https://doi.org/10.1001/jama.2023.22295

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry, 11*, Article 108. https://doi.org/10.1038/s41398-021-01224-x

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*(9), 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Kahneman, D. (2012). *Thinking, fast and slow*. Penguin.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgment*. William Collins.

Keller, N., Jenny, M. A., Spies, C. A., & Herzog, S. M. (2020). Augmenting decision competence in healthcare using AI-based cognitive models. *2020 IEEE International Conference on Healthcare Informatics, Germany*, 1–4. https://doi.org/10.1109/ICHI48887.2020.9374376

Keshavarzi Arshadi, A., Salem, M., Karner, H., Garcia, K., Arab, A., Yuan, J. S., & Goodarzi, H. (2024). Functional microRNA-targeting drug discovery by graph-based deep learning. *Patterns, 5*(1), Article 100909. https://doi.org/10.1016/j.patter.2023.100909

Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. *Science, 380*(6651), 1222–1223. https://doi.org/10.1126/science.adi0248

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2020). Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences, 117*(48), 30096–30100. https://doi.org/10.1073/pnas.1912790117

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior, 114*, Article 106553. https://doi.org/10.1016/j.chb.2020.106553

Kool, W., & Botvinick, M. (2018). Mental labour. *Nature Human Behaviour, 2*(12), 899–908. https://doi.org/10.1038/s41562-018-0401-9

Korteling, J. E. H., van de Boer-Visschedijk, G. C., Blankendaal, R. A. M., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human versus artificial intelligence. *Frontiers in Artificial Intelligence, 4*, Article 622364. https://doi.org/10.3389/frai.2021.622364

Kupfer, C., Prassl, R., Fleiß, J., Malin, C., Thalmann, S., & Kubicek, B. (2023). Check the box! How to deal with automation bias in AI-based personnel selection. *Frontiers in Psychology, 14*, Article 1118723. https://doi.org/10.3389/fpsyg.2023.1118723

Kwong, J. C. C., Nguyen, D., Khondker, A., Kim, J. K., Johnson, A. E. W., McCradden, M. M., Kulkarni, G. S., Lorenzo, A., Erdman, L., & Rickard, M. (2024). When the model trains you: Induced belief revision and its implications on artificial intelligence research and

patient care — A case study on predicting obstructive hydronephrosis in children. *NEJM AI*, *1*(2), 1–7. https://doi.org/10.1056/AIcs2300004

L'Imperio, V., Wulczyn, E., Plass, M., Müller, H., Tamini, N., Gianotti, L., Zucchini, N., Reihs, R., Corrado, G. S., Webster, D. R., Peng, L. H., Chen, P.-H. C., Lavitrano, M., Liu, Y., Steiner, D. F., Zatloukal, K., & Pagni, F. (2023). Pathologist validation of a machine learning–derived feature for colon cancer risk stratification. *JAMA Network Open, 6*(3), Article e2254891. https://doi.org/10.1001/jamanetworkopen.2022.54891

Lacroux, A., & Martin-Lacroux, C. (2022). Should I trust the artificial intelligence to recruit? Recruiters' perceptions and behavior when faced with algorithm-based recommendation systems during resume screening. *Frontiers in Psychology, 13,* Article 895997. https://doi.org/10.3389/fpsyg.2022.895997

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models. *Proceedings of the Conference on Fairness, Accountability, and Transparency, USA,* 29–38. https://doi.org/10.1145/3287560.3287590

Lai, Y., Kankanhalli, A., & Ong, D. C. (2021). Human-AI collaboration in healthcare: A review and research agenda. *Proceedings of the 54th Hawaii International Conference on System Sciences, USA*, 390–399. https://scholarspace.manoa.hawaii.edu/items/48162f4c-8c99-47dc-8fb5-fe74e9cbb51d

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences, 117*(23), 12592–12594. https://doi.org/10.1073/pnas.1919012117

Lazcoz, G., & de Hert, P. (2023). Humans in the GDPR and AIA governance of automated and algorithmic systems. Essential pre-requisites against abdicating responsibilities.

Computer Law & Security Review, *50*, Article 105833. https://doi.org/10.1016/j.clsr.2023.105833

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

Linder, J. A., Doctor, J. N., Friedberg, M. W., Reyes Nieva, H., Birks, C., Meeker, D., & Fox, C. R. (2014). Time of day and the decision to prescribe antibiotics. *JAMA Internal Medicine, 174*(12), 2029–2031. https://doi.org/10.1001/jamainternmed.2014.5225

Liu, M., Wu, J., Wang, N., Zhang, X., Bai, Y., Guo, J., Zhang, L., Liu, S., & Tao, K. (2023). The value of artificial intelligence in the diagnosis of lung cancer: A systematic review and meta-analysis. *PLoS ONE, 18*(3), Article e0273445. https://doi.org/10.1371/journal.pone.0273445

Loftus, T. J., Tighe, P. J., Filiberto, A. C., Efron, P. A., Brakenridge, S. C., Mohr, A. M., Rashidi, P., Upchurch, G. R., & Bihorac, A. (2020). Artificial intelligence and surgical decision-making. *JAMA Surgery, 155*(2), 148–158. https://doi.org/10.1001/jamasurg.2019.4917

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

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association, 24*(2), 423–431. https://doi.org/10.1093/jamia/ocw105

Lyell, D., Coiera, E., Chen, J., Shah, P., & Magrabi, F. (2021). How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical

devices. *BMJ Health & Care Informatics, 28*, Article e100301. https://doi.org/10.1136/bmjhci-2020-100301

Lyell, D., Magrabi, F., & Coiera, E. (2018). The effect of cognitive load and task complexity on automation bias in electronic prescribing. *Human Factors, 60*(7), 1008–1021. https://doi.org/10.1177/0018720818781224

Lyell, D., Magrabi, F., Raban, M. Z., Pont, L. G., Baysari, M. T., Day, R. O., & Coiera, E. (2017). Automation bias in electronic prescribing. *BMC Medical Informatics and Decision Making, 17*(1), 1–10. https://doi.org/10.1186/s12911-017-0425-5

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

McNeil, B. J., Pauker, S. G., Sox, H. C., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine, 306*(21), 1259–1262. https://doi.org/10.1056/NEJM198205273062103

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), Article 115. https://doi.org/10.1145/3457607

Mittermaier, M., Raza, M. M., & Kvedar, J. C. (2023). Bias in AI-based models for medical applications: challenges and mitigation strategies. *npj Digital Medicine, 6*(1), Article 113. https://doi.org/10.1038/s41746-023-00858-z

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature, 616*(7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4

Morewedge, C. K. (2022). Preference for human, not algorithm aversion. *Trends in Cognitive Sciences, 26*(10), 824–826. https://doi.org/10.1016/j.tics.2022.07.007

Moulton, C. E., Regehr, G., Mylopoulos, M., & MacRae, H. M. (2007). Slowing down when you should: A new model of expert judgment. *Academic Medicine, 82*(10), S109–S116. https://doi.org/10.1097/ACM.0b013e3181405a76

Mullainathan, S., & Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review, 107*(5), 476–480. https://doi.org/10.1257/aer.p20171084

Mullainathan, S., & Obermeyer, Z. (2021). On the Inequity of Predicting A While Hoping for B. *AEA Papers and Proceedings, 111*, 37–42. https://doi.org/10.1257/pandp.20211078

Navarro, D. J., & Foxcroft, D. R. (2022). *Learning statistics with Jamovi: A tutorial for psychology students and other beginners*. https://doi.org/10.24384/hgc3-7p15

Neprash, H. T., & Barnett, M. L. (2019). Association of primary care clinic appointment time with opioid prescribing. *JAMA Network Open, 2*(8), Article e1910373. https://doi.org/10.1001/jamanetworkopen.2019.10373

Ng, A. Y., Oberije, C. J. G., Ambrózay, É., Szabó, E., Serfőző, O., Karpati, E., Fox, G., Glocker, B., Morris, E. A., Forrai, G., & Kecskemethy, P. D. (2023). Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. *Nature Medicine, 29*(12), 3044–3049. https://doi.org/10.1038/s41591-023-02625-9

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns, 2*(10), Article 100347. https://doi.org/10.1016/j.patter.2021.100347

Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E., & Gogate, V. (2021). Anchoring bias affects mental model formation and user reliance in explainable AI systems. *26th International Conference on Intelligent User Interfaces, USA*, 340–350. https://doi.org/10.1145/3397481.3450639

O'Neil, C. (2016). *Weapons of Math Destruction.* Crown Publishers.

Obermeyer, Z., & Lee, T. H. (2017). Lost in thought — The limits of the human mind and the future of medicine. *New England Journal of Medicine, 377*(13), 1209–1211. https://doi.org/10.1056/NEJMp1705348

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Obermeyer, Z., & Topol, E. J. (2021). Digital medicine, artificial intelligence, bias, and patients' perspectives. *The Lancet, 397*(10289), 2038. https://doi.org/10.1016/S0140-6736(21)01152-1

Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing bias in artificial intelligence in health care. *JAMA, 322*(24), 2377–2378. https://doi.org/10.1001/jama.2019.18058

Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2023). Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence, 5*, 1076–1086. https://doi.org/10.1038/s42256-023-00720-7

Patel, B. N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., Rajpurkar, P., Amrhein, T., Gupta, R., Halabi, S., Langlotz, C., Lo, E., Mammarappallil, J., Mariano, A. J., Riley, G., Seekins, J., Shen, L., Zucker, E., & Lungren, M. P. (2019). Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digital Medicine, 2*, Article 111. https://doi.org/10.1038/s41746-019-0189-7

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J.-C., Castillo, C. D., Chellappa, R., White, D., & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences, 115*(24), 6171–6176. https://doi.org/10.1073/pnas.1721355115

Pierce, R. L., Van Biesen, W., Van Cauwenberge, D., Decruyenaere, J., & Sterckx, S. (2022). Explainability in medicine in an era of AI-based clinical decision support systems. *Frontiers in Genetics*, *13*, Article 903600. https://doi.org/10.3389/fgene.2022.903600

Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S., & Obermeyer, Z. (2021). An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine, 27*(1), 136–140. https://doi.org/10.1038/s41591-020-01192-7

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting, 36*(6), 691–702. https://doi.org/10.1002/for.2464

Quigley, M., & Haselgrove, M. (2023). Overshadowing, but not relative validity, between the elements of an outcome during human associative learning. *Neurobiology of Learning and Memory, 203,* Article 107790. https://doi.org/10.1016/j.nlm.2023.107790

Rago, A., Cocarascu, O., Bechlivanidis, C., Lagnado, D., & Toni, F. (2021). Argumentative explanations for interactive recommendations. *Artificial Intelligence*, *296*, Article 103506. https://doi.org/10.1016/j.artint.2021.103506

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine, 28*(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 6*, Article 83. https://doi.org/10.1145/3512930

Rebelo, A. D., Verboom, D. E., dos Santos, N. R., & de Graaf, J. W. (2023). The impact of artificial intelligence on the tasks of mental healthcare workers: A scoping review. *Computers in Human Behavior: Artificial Humans, 1*(2), Article 100008. https://doi.org/10.1016/j.chbah.2023.100008

Rebitschek, F. G., Gigerenzer, G., & Wagner, G. G. (2021). People underestimate the errors made by algorithms for credit scoring and recidivism prediction but accept even fewer errors. *Scientific Reports*, *11*, Article 20171 https://doi.org/10.1038/s41598-021-99802-y

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., Antonelli, G., Awadie, H., Bernhofer, S., Carballal, S., Dinis-Ribeiro, M., Fernández-Clotett, A., Esparrach, G. F., Gralnek, I., Higasa, Y., Hirabayashi, T., Hirai, T., Iwatate, M., Kawano, M., Mader, M., … Cherubini, A. (2022). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports, 12*, Article 14952. https://doi.org/10.1038/s41598-022-18751-2

Rezazade Mehrizi, M. H., Mol, F., Peter, M., Ranschaert, E., Dos Santos, D. P., Shahidi, R., Fatehi, M., & Dratsch, T. (2023). The impact of AI suggestions on radiologists' decisions:

a pilot study of explainability and attitudinal priming interventions in mammography examination. *Scientific Reports, 13*, Article 9230. https://doi.org/10.1038/s41598-023-36435-3

Rieger, T., Roesler, E., & Manzey, D. (2022). Challenging presumed technological superiority when working with (artificial) colleagues. *Scientific Reports*, *12*, Article 3768. https://doi.org/10.1038/s41598-022-07808-x

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., Helbich, T. H., Chevalier, M., Tan, T., Mertelmeier, T., Wallis, M. G., Andersson, I., Zackrisson, S., Mann, R. M., & Sechopoulos, I. (2019). Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists. *Journal of the National Cancer Institute, 111*(9), 916–922. https://doi.org/10.1093/jnci/djy222

Russell, S. & Norvig, P. (2010) *Artificial Intelligence: A Modern Approach* (3rd ed.) Prentice Hall.

Sanz-Urquijo, B., Fosch-Villaronga, E., & Lopez-Belloso, M. (2022). The disconnect between the goals of trustworthy AI for law enforcement and the EU research agenda. *AI and Ethics, 3,* 1283–1294. https://doi.org/10.1007/s43681-022-00235-8

Saposnik, G., Redelmeier, D., Ruff, C. C., & Tobler, P. N. (2016). Cognitive biases associated with medical decisions: a systematic review. *BMC Medical Informatics and Decision Making, 16*, Article 138. https://doi.org/10.1186/s12911-016-0377-1

Schultebraucks, K., Shalev, A. Y., Michopoulos, V., Grudzen, C. R., Shin, S.-M., Stevens, J. S., Maples-Keller, J. L., Jovanovic, T., Bonanno, G. A., Rothbaum, B. O., Marmar, C. R., Nemeroff, C. B., Ressler, K. J., & Galatzer-Levy, I. R. (2020). A validated predictive algorithm of post-traumatic stress course following emergency department admission

after a traumatic stressor. *Nature Medicine, 26*(7), 1084–1088. https://doi.org/10.1038/s41591-020-0951-z

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine, 27*(12), 2176–2182. https://doi.org/10.1038/s41591-021-01595-0

Shin, M., Kim, J., van Opheusden, B., & Griffiths, T. L. (2023). Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences, 120*(12), 1–6. https://doi.org/10.1073/pnas.2214840120

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA, 320*(21), 2199–2200. https://doi.org/10.1001/jama.2018.17163

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human Computer Studies, 51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). Rational actors or rational fools: Implications of the affect heuristic for behavioral economics. *Journal of Socio-Economics, 31*(4), 329–342. https://doi.org/10.1016/S1053-5357(02)00174-9

Solans, D., Beretta, A., Portela, M., Castillo, C., & Monreale, A. (2022). *Human response to an AI-based decision support system: A user study on the effects of accuracy and bias*. ArXiv http://arxiv.org/abs/2203.15514.

Somashekhar, S. P., Sepúlveda, M.-J., Puglielli, S., Norden, A. D., Shortliffe, E. H., Rohit Kumar, C., Rauthan, A., Arun Kumar, N., Patil, P., Rhee, K., & Ramya, Y. (2018). Watson for Oncology and breast cancer treatment recommendations: agreement with an expert

multidisciplinary tumor board. *Annals of Oncology, 29*(2), 418–423. https://doi.org/10.1093/annonc/mdx781

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., White, S., Habli, I., & Reynolds, N. (2019). Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health and Care Informatics, 26*, Article e100081. https://doi.org/10.1136/bmjhci-2019-100081

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

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, Scotland UK*, Article 538. https://doi.org/10.1145/3290605.3300768

Suresh, H., Lao, N., & Liccardi, I. (2020). Misplaced trust: Measuring the interference of machine learning in human decision-making. *Proceedings of the 12th ACM Conference on Web Science, UK*, 315–324. https://doi.org/10.1145/3394231.3397922

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine, 3*, Article 17. https://doi.org/10.1038/s41746-020-0221-y

Thomasian, N. M., Eickhoff, C., & Adashi, E. Y. (2021). Advancing health equity with artificial intelligence. *Journal of Public Health Policy, 42*(4), 602–611. https://doi.org/10.1057/s41271-021-00319-5

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Topol, E. J. (2023). As artificial intelligence goes multimodal, medical applications multiply. *Science, 381*(6663). https://doi.org/10.1126/science.adk6139

Trayanova, N. A., & Topol, E. J. (2022). Deep learning a person's risk of sudden cardiac death. *The Lancet, 399*(10339), 1933. https://doi.org/10.1016/S0140-6736(22)00881-9

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., & Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine, 26*(8), 1229–1234. https://doi.org/10.1038/s41591-020-0942-0

Turing, A. M. (1950) Computing machinery and intelligence. *Mind, 59*(246), 433-460. https://www.jstor.org/stable/2251299

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(30), 453–458. https://doi.org/10.1007/978-1-4613-2391-4_2

Valdivia, A., Hyde-Vaamonde, C., & García-Marcos, J. (2022). *Judging the algorithm: A case study on the risk assessment tool for gender-based violence implemented in the Basque country.* ArXiv https://arxiv.org/abs/2203.03723v2.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction, 7,* Article 129. https://doi.org/10.1145/3579605

Véliz, C. (2019). Three things digital ethics can learn from medical ethics. *Nature Electronics*, *2*(8), 316–318. https://doi.org/10.1038/s41928-019-0294-2

Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI & SOCIETY*, *36*(2), 487–497. https://doi.org/10.1007/s00146-021-01189-x

Véliz, C. (2023). Chatbots shouldn't use emojis. *Nature*, *615*(7952), 375–375. https://doi.org/10.1038/d41586-023-00758-y

Vered, M., Livni, T., Howe, P. D. L., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. *Artificial Intelligence, 322*, Article 103952. https://doi.org/10.1016/j.artint.2023.103952

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

Walsh, C. G., Chaudhry, B., Dua, P., Goodman, K. W., Kaplan, B., Kavuluru, R., Solomonides, A., & Subbian, V. (2021). Stigma, biomarkers, and algorithmic bias: Recommendations for precision behavioral health with artificial intelligence. *JAMIA Open, 3*(1), 9–15. https://doi.org/10.1093/JAMIAOPEN/OOZ054

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine, 25*(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Wulczyn, E., Nagpal, K., Symonds, M., Moran, M., Plass, M., Reihs, R., Nader, F., Tan, F., Cai, Y., Brown, T., Flament-Auvigne, I., Amin, M. B., Stumpe, M. C., Müller, H., Regitnig, P., Holzinger, A., Corrado, G. S., Peng, L. H., Chen, P.-H. C., … Mermel, C. H. (2021). Predicting prostate cancer specific-mortality with artificial intelligence-based Gleason grading. *Communications Medicine, 1*, Article 10. https://doi.org/10.1038/s43856-021-00005-3

Yang, Y., Yuan, Y., Zhang, G., Wang, H., Chen, Y. C., Liu, Y., Tarolli, C. G., Crepeau, D., Bukartyk, J., Junna, M. R., Videnovic, A., Ellis, T. D., Lipford, M. C., Dorsey, R., & Katabi, D. (2022). Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nature Medicine, 28*(10), 2207–2215. https://doi.org/10.1038/s41591-022-01932-x

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making, 32*(4), 403–414. https://doi.org/10.1002/bdm.2118

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Scotland UK*, Article 279. https://doi.org/10.1145/3290605.3300509

Yu, K.-H., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019). Do I trust my machine teammate? an investigation from perception to decision. *Proceedings of the 24th International Conference on Intelligent User Interfaces, USA,* 460–468. https://doi.org/10.1145/3301275.3302277

Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., & Snyder, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic

pathology image features. *Nature Communications, 7*, Article 12474. https://doi.org/10.1038/ncomms12474

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering, 2*(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., Butte, A. J., & Alsentzer, E. (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health, 6*(1), e12–e22. https://doi.org/10.1016/S2589-7500(23)00225-X

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Spain,* 295–305. https://doi.org/10.1145/3351095.3372852

Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., Liu, T., Xu, M., Lozano, M. G., Woodward-Court, P., Kihara, Y., Allen, N., Gallacher, J. E. J., Littlejohns, T., Aslam, T., Bishop, P., Black, G., Sergouniotis, P., Atan, D., … Keane, P. A. (2023). A foundation model for generalizable disease detection from retinal images. *Nature, 622*(7981), 156–163. https://doi.org/10.1038/s41586-023-06555-x

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist — it's time to make it fair. *Nature, 559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8

Apéndice A

Instrucciones utilizadas en la tarea de clasificación

Consentimiento informado

Tu participación es voluntaria y anónima.	**Your participation is voluntary and anonymous.**
Antes de nada, queremos agradecer tu participación en este experimento, ya que sin la colaboración de personas como tú no sería posible esta investigación. Debes saber que en esta tarea no hay respuestas buenas ni malas. Lo que queremos estudiar son los mecanismos psicológicos básicos que se dan en todas las personas. Para ello necesitamos que, **si deseas participar, lo hagas con el mayor interés**. La realización de la tarea te llevará unos 10 minutos. No tienes que identificarte, y los datos que nos aportes se unirán a los del total del grupo y serán analizados estadísticamente. Tu participación es voluntaria y anónima. Si tras haber leído el mensaje deseas continuar, pulsa el botón.	First of all, we'd like to thank you for your participation in this experiment. It would be impossible to carry out this investigation without the cooperation of people like you. You should know that there are no good or bad answers to this task. What we want to study are the basic psychological mechanisms inherent in everybody. In order to do this, **we need you to participate taking as much interest as possible.** The task will take about 10 minutes to complete. You don't have to identify yourself and any collected data will be added to the rest of the group's data and analysed statistically. Your participation is voluntary and anonymous. If after having read the message you wish to continue, please press the button.

Pulsa F11 para ver la tarea en pantalla completa.	Press F11 to view the task in full screen.

Datos demográficos

Necesitamos algo de información básica sobre ti.	**We need some basic information about you.**
Trataremos la información de **forma anónima**, considerando el total del grupo y no de forma individual. **No guardamos información personal** sobre ti y los datos solo serán utilizados para esta investigación. Edad: Género: Masculino/Femenino/No binario/Prefiero no contestar	We remind you that the data is stored and always treated in **an anonymous manner**, only considering the group of all participants in the experiment not in an individual way. **No sensitive data is kept** and the rest of the data will only be used for scientific purposes. Age: Gender: Male/Female/Non-binary/I prefer not to answer

Introducción a la tarea de clasificación

Imagina que eres un médico especialista en una enfermedad llamada **"Síndrome de Lindsay"**. Los tejidos sobre los que actúa esta enfermedad tienen **dos tipos de células:** unas se ven de color claro y otras de color oscuro**.** El Síndrome de Lindsay provoca que en **la muestra de tejido afectado** extraído a una persona haya un **mayor número de células de color [oscuro/claro]** que de color [claro/oscuro]. En cambio, en una muestra de tejido **no afectado** por el síndrome hay un **mayor número de células de color [claro/oscuro]** que de color [oscuro/claro].	Imagine that you are a doctor who specialised in treating a very rare condition called **"Lindsay Syndrome"** The human tissues affected by this disease have **two types of cells**: some are light-coloured and others are dark-coloured. Lindsay syndrome causes the **affected tissue sample** to have a **greater number of [dark/dark] coloured cells** than [light/dark] coloured cells. In contrast, a **sample of tissue not affected** by the syndrome has **a greater number of [light/dark]** than [dark/light] coloured cells.

Fase de práctica

Es posible saber si una muestra de tejido humano está afectada por el síndrome de Lindsay o no, **mirando la proporción de células de color oscuro y de color claro en el tejido**. Vamos a asegurarnos de que puedes distinguirlas realizando una breve práctica. A continuación, vas a **ver una serie de muestras de tejidos**, y deberás indicar **si están afectados por el síndrome de Lindsay o no.** El objetivo de esta práctica es que aprendas el criterio de clasificación.	It is possible to tell whether a sample of tissue is affected by Lindsay Syndrome or not **by looking at the proportion of dark-coloured and light-coloured cells in the tissue.** Let's make sure you can tell the difference by doing a short practice run. You will then look at **a series of tissue samples**, and you will have to indicate whether they are **affected by Lindsay Syndrome or not**. The aim of this practice is that you learn the classification criteria.

Recuerda, **el criterio para realizar la tarea** es: Si hay más células de **color [oscuro/claro]** que de color [claro/oscuro], el tejido está afectado por el síndrome de Lindsay y la muestra de tejido se clasifica como un **POSITIVO.** Si hay más células de **color [claro/oscuro]** que de color [oscuro/claro], el tejido NO está afectado por el síndrome de Lindsay y la muestra de tejido se clasifica como un **NEGATIVO.** Cuando estés preparado para realizar la tarea**, pulsa el botón "Siguiente".**	Remember, the **classification criteria** are: If there are **more [dark/light] coloured cells** than [light/dark] coloured cells, the tissue is affected by Lindsay Syndrome, so the tissue sample is classified as **POSITIVE**. If there are **more [light/dark] coloured cells** than [dark/light] coloured cells, the tissue is NOT affected by Lindsay Syndrome, so the tissue sample is classified as **NEGATIVE**. When you are ready to perform the task, **press the "Next" button**.

Segundo bloque de ensayos de la fase de práctica

Ahora continuarás practicando la clasificación con algunas muestras más. Recibirás información sobre como lo has hecho al finalizar.	You will now continue practising the classification with some more samples. You will receive feedback on your performance at the end.

Mensaje al finalizar la fase de práctica

¡Muy bien! Ahora eres capaz de clasificar los tejidos de forma que puedes determinar cuales están afectados por el síndrome de Lindsay y cuáles no.	**Very good!** Now, you can classify tissue samples indicating which ones are affected by Lindsay Syndrome and which ones are not.
AVISO Parece que no has entendido correctamente en qué consiste el síndrome de Lindsay. Por favor, **vuelve a leer las instrucciones**.	**NOTICE** You seem to have misunderstood what Lindsay Syndrome is. Please, **read the instructions again.**

Presentación del sistema de inteligencia artificial ficticio

El diagnóstico del **síndrome de Lindsay** mediante la observación de muestras de tejido permite **detectar la enfermedad** en fases muy tempranas, mejorando las posibilidades de tratamiento y recuperación de los pacientes. Recientemente, se ha creado un **sistema de asistencia al diagnóstico** basado en un **algoritmo de inteligencia artificial (IA)** que puede ayudar a los médicos en la detección temprana del síndrome de Lindsay. Este sistema analiza cada muestra de tejido y **ofrece recomendaciones durante la tarea** de clasificación de las muestras.	Observing tissue samples can help us diagnose **Lindsay Syndrome** at an **early stage**, improving patients' chances for treatment and recovery. Recently, a **diagnostic assistance system based** on **an artificial intelligence (AI) algorithm** has been developed to help doctors in the early detection of Lindsay Syndrome. This system analyses each tissue sample and **provides recommendations during its classification.**

Nos han llegado muchas muestras de tejido de un grupo de personas. Algunas están sufriendo el síndrome de Lindsay y otras no. Te pedimos **que nos ayudes a clasificar estas muestras**. Durante la tarea de clasificación, **es posible que la inteligencia artificial esté conectada**. Si es así, te ofrecerá su recomendación. Si la **IA está conectada**, **verás su recomendación** durante toda la tarea en la parte superior de la pantalla, encima de la muestra, como se indica	We have received many tissue samples from a group of people. Some are suffering from Lindsay Syndrome, while others are not. **We are asking you to help us classify these samples.** During the classification task, the **artificial intelligence system may be connected**. If so, it will offer you its recommendation. **If the AI is on it will offer you a recommendation** that will appear at the top of the screen above the sample, as you can see in the image. But if the **AI is off**, you will have to perform the whole task **without assistance**.

en la imagen. En **caso de que la IA no esté conectada** tendrás que realizar toda la tarea **sin ayuda**.	

Recuerda, deberás clasificar cada muestra de células indicando **POSITIVO** si el tejido está afectado por el síndrome o **NEGATIVO** si el tejido no está afectado. En un **tejido afectado** por el **síndrome de Lindsay,** hay más células de **color [oscuro/claro],** mientras que en un tejido **no afectado** por el síndrome hay más células de **color [claro/oscuro].** ¡Buena suerte!	Remember, you should classify each tissue sample as **POSITIVE** if the tissue **is affected** by the syndrome or **NEGATIVE** if the tissue **is not affected.** A human tissue **affected** by Lindsay Syndrome has **more [dark/light] coloured cells.** A tissue **not affected** by the syndrome has **more [light/dark] coloured cells**. Good luck!
Mensajes para indicar a los participantes si verán las recomendaciones o no.	The **AI is on**, so it will offer you a recommendation.
	The **IA is off**, so you will have to perform the task without assistance.

Introducción a la Fase 2 de la tarea de clasificación

Casi hemos terminado, pero antes… Han llegado unas pocas muestras de tejido adicionales. Nos gustaría que nos **ayudases a clasificarlas tal como has hecho antes**, siguiendo el mismo criterio. ¡Muchas gracias!	We are almost finished, but first... A few additional tissue samples have arrived. We would like you **to help us classify them as you did before**, following the same criteria. Thank you very much!

Envío de datos

A continuación, **podrás enviar los resultados para que se incluyan en nuestro estudio**. Los datos que nos aportes se unirán a los del grupo y serán analizados estadísticamente de manera anónima. Para hacerlo, **haz click en el botón "Enviar". Si por alguna razón no deseas enviárnoslos, sal de la página o cierra el navegador sin enviar tus respuestas y estas no se guardarán.**	Now, **you can send the results to be included in our study**. All data will be added to the rest of the group and analysed statistically and anonymously. **If you want to send the data click on "Send"**. If there is a reason **you do not want to send us your answers, please exit the page or close the browser without submitting your answers**, and they will not be saved

Antes de terminar queremos darte las gracias por tu colaboración en este experimento. **Nuestra labor de investigación sería imposible si no fuera por la participación voluntaria de personas como tú** que deciden dedicarnos parte de su tiempo. Suponemos que ahora que has terminado el experimento te gustaría recibir una explicación sobre lo que tratábamos de estudiar. Nuestra intención es estudiar cómo afecta al comportamiento de las personas la recomendación de algoritmos de inteligencia artificial en diversos ámbitos. Algunos de los voluntarios veíais una inteligencia artificial que os ayudaba en la clasificación de las muestras de tejidos, otros, sin embargo, lo hacíais solos. Lo que tratamos de investigar es si clasificáis de la misma o diferente forma cuando sois ayudados por la IA. Lo sabremos cuando analicemos los resultados. **¡Muchas gracias por tu participación!**	Before concluding, we would like to thank you for your cooperation in this experiment. **Our research work would be impossible without the voluntary participation of people like you** who decide to dedicate some of their time to us. Now that you have finished the experiment, you probably would welcome an explanation of what we were trying to study. We intend to study how people's behaviour is affected by the recommendation of artificial intelligence algorithms in various areas. Some of you viewed the recommendations made by an artificial intelligence during the classification task, while others did the task without assistance. We were investigating if there are differences in the classification task as a function of whether you are assisted or not by an AI. We will know when we analyse the results. Please do not discuss the details of the study with your acquaintances or on social networks until we have finished so as not to influence them. **Thank you very much for your participation.** Don’t forget to click the button to confirm that you’ve completed the experiment.

Apéndice B

Índices de error en cada tipo de estímulo

Tabla 1B

Ratio Medio de Clasificación Errónea para Todos los Estímulos de Diferentes Proporciones de Células Oscuras/Claras Utilizados en la Tarea de Clasificación en los Experimentos de la Serie 1.

Experimento	EE	Grupo	Fase 1			Fase 2		
			Ratio error		Dif.	Ratio error		Dif.
			M	*SD*		*M*	*SD*	
1	80/20	Con IA	.00	.01	$p = .993$			
		Sin IA	.00	.01				
	70/30	Con IA	.00	.00	$p = .099$			
		Sin IA	.00	.02				
	60/40	Con IA	.01	.06	$p = .138$			
		Sin IA	.03	.11				
	40/60	Con IA	.19	.28	$p < .001$			
		Sin IA	.06	.16				
	30/70	Con IA	.00	.00	$p = .070$			
		Sin IA	.00	.04				
	20/80	Con IA	.00	.03	$p = .410$			
		Sin IA	.00	.01				
2	80/20	Con IA	.00	.00	-			
		Sin IA	.00	.00				
	70/30	Con IA	.00	.02	$p = .245$	.01	.04	$p = .256$
		Sin IA	.00	.22		.00	.02	
	60/40	Con IA	.00	.03	$p = .008$	.02	.08	$p = .298$
		Sin IA	.02	.07		.03	.10	
	40/60	Con IA	.35	.39	$p < .001$	.18	.31	$p < .001$
		Sin IA	.06	.15		.05	.15	
	30/70	Con IA	.00	.02	$p = .432$	.02	.10	$p = .103$
		Sin IA	.00	.03		.00	.03	
	20/80	Con IA	.00	.01	$p = .644$			
		Sin IA	.00	.01				
3	80/20	Con IA→Sin IA	.00	.05	$p = .129$	.00	.02	$p = .316$
		Sin IA→Con IA	.00	.00		.00	.00	
	70/30	Con IA→Sin IA	.00	.01	$p = .904$	.00	.00	$p = .182$
		Sin IA→Con IA	.00	.02		.00	.03	
	60/40	Con IA→Sin IA	.01	.04	$p = .034$	.01	.04	$p = .223$
		Sin IA→Con IA	.04	.11		.00	.03	
	40/60	Con IA→Sin IA	.37	.37	$p < .001$	.33	.38	$p = .403$
		Sin IA→Con IA	.07	.18		.37	.41	
	30/70	Con IA→Sin IA	.01	.07	$p = .137$	.05	.13	$p = .025$
		Sin IA→Con IA	.00	.02		.01	.07	
	20/80	Con IA→Sin IA	.00	.01	$p = .316$	.00	.00	-
		Sin IA→Con IA	.00	.00		.00	.00	

Experimento	EE	Grupo	Fase 1			Fase 2		
			Ratio error		Dif.	Ratio error		Dif.
			M	*SD*		*M*	*SD*	
4	80/20	Con IA	.00	.00	-			
		Sin IA	.00	.00				
	70/30	Con IA	.00	.01	$p = .315$	.00	.00	$p = .156$
		Sin IA	.00	.04		.00	.04	
	60/40	Con IA	.00	.01	$p = .152$	.00	.03	$p = .067$
		Sin IA	.01	.06		.04	.12	
	40/60	Con IA	.35	.42	$p < .001$	.26	.36	$p < .001$
		Sin IA	.01	.04		.00	.04	
	30/70	Con IA	.00	.01	$p = .656$	.00	.03	$p = .320$
		Sin IA	.00	.03		.00	.00	
	20/80	Con IA	.00	.00	$p = .320$			
		Sin IA	.00	.04				

Nota. La razón de clasificación errónea por tipo de estímulo es el número total de errores de clasificación cometidos en estímulos de un tipo determinado dividido por el número total de ensayos con dichos estímulos en cada fase de la tarea. La columna Dif. representa el valor *p* de la prueba t de Student para la diferencia entre las ratios medias de clasificación errónea de los grupos con asistencia de la IA (con IA) y sin asistencia (sin IA).

Tabla 2B

Ratio Medio de Clasificación Errónea para Todos los Estímulos de Diferentes Proporciones de Células Oscuras/Claras Utilizados en la Tarea de Clasificación en los Experimentos 5 y 6 de la Serie 2.

Experimento	EE	Grupo	Fase 1			Fase 2		
			Ratio error		Dif.	Ratio error		Dif.
			M	*SD*		*M*	*SD*	
5	80/20	Info-error	.00	.04	p = .226			
		No-info	.00	.00				
	70/30	Info-error	.00	.01	p = .410	.01	.06	p = .082
		No-info	.00	.01		.00	.00	
	60/40	Info-error	.01	.05	p = .690	.05	.17	p = .016
		No-info	.01	.08		.01	.05	
	40/60	Info-error	.29	.36	p = .254	.22	.35	p = .171
		No-info	.35	.40		.29	.37	
	30/70	Info-error	.02	.11	p = .336	.03	.11	p = .542
		No-info	.01	.05		.04	.15	
	20/80	Info-error	.00	.01	p = .653			
		No-info	.00	.01				
6	80/20	Info-error	.00	.04	-			
		Info-acierto	.00	.03				
		No info	.00	.00				
	70/30	Info-error	.01	.09	p = .781	.01	.12	
		Info-acierto	.00	.03		.00	.00	
		No info	.00	.02		.00	.02	
	60/40	Info-error	.11	.16	p = .095	.04	.15	p = .730
		Info-acierto	.09	.12		.03	.11	
		No info	.06	.09		.03	.12	
	40/60	Info-error	.12	.27	p = .011	.07	.19	p = .073
		Info-acierto	.21	.34		.13	.29	
		No info	.30	.37		.18	.33	
	30/70	Info-error	.00	.02	p = .697	.00	.04	p = .752
		Info-acierto	.00	.01		.01	.11	
		No info	.00	.00		.01	.06	
	20/80	Info-error	.00	.01	-			
		Info-acierto	.00	.00				
		No info	.00	.00				

Nota. La columna Dif. representa el valor *p* de la prueba t de Student para la diferencia entre grupos en las ratios medias de clasificación errónea excepto en el caso del Experimento 6, donde la columna Dif. representa el valor *p* asociado al factor grupo (info error, info acierto, no info) en el análisis de varianza con la ratio de error en cada tipo de estímulo como variable dependiente. Dado que el Experimento 7 incluyó una variable adicional, el tipo de error, lo cual aumentó la complejidad de su diseño, decidimos excluirlo de la Tabla 2B.

Tabla 3B

Ratio Medio de Clasificación Errónea para los Estímulos de Proporciones de Células Oscuras/Claras 60/40 y 40/60 Utilizados en la Tarea de Clasificación en los Experimentos de la Serie 2 en Función de la Información sobre el Error y el Tipo de Error que Podía Cometerse en la Clasificación Según los Criterios de Clasificación B1 y B2.

Experimento	EE	Grupo	Fase 1			Fase 2		
			Ratio error		Dif.	Ratio error		Dif.
			M	*SD*		*M*	*SD*	
5	60/40	Con IA – info error FP	.02	.07		.08	.22	Info.
		Con IA – info error FN	.00	.02	Tipo de	.02	.07	error
		Con IA – no info FP	.03	.11	error	.02	.07	Tipo de
		Con IA – no info FN	.00	.01		.00	.00	error
	40/60	Con IA – info error FP	.37	.38		.29	.40	
		Con IA – info error FN	.20	.31	Tipo de	.14	.14	Tipo de
		Con IA – no info FP	.49	.43	error	.41	.41	error
		Con IA – no info FN	.22	.33		.16	.16	
6	60/40	Con IA – info error FP	.17	.16		.06	.19	
		Con IA – info error FN	.04	.12		.02	.08	
		Con IA – info acierto FP	.13	.08	Tipo de	.00	.03	-
		Con IA – info acierto FN	.05	.14	error	.05	.15	
		Con IA – no info FP	.14	.10		.04	.16	
		Con IA – no info FN	.00	.02		.02	.08	
	40/60	Con IA – info error FP	.22	.35		.13	.25	
		Con IA – info error FN	.04	.16	Info	.02	.11	
		Con IA – info acierto FP	.24	.38	error.	.15	.31	Tipo de
		Con IA – info acierto FN	.18	.31	Tipo de	.10	.29	error
		Con IA – no info FP	.37	.37	error	.22	.34	
		Con IA – no info FN	.21	.36		.13	.31	
7	60/40	Con IA – info error FP	.15	.27		.14	.28	
		Con IA – info error FN	.19	.34	Info.	.13	.26	Info.
		Con IA – no info FP	.35	.40	error	.31	.37	error
		Con IA – no info FN	.24	.34		.20	.34	
	40/60	Con IA – info error FP	.09	.20		.10	.22	
		Con IA – info error FN	.03	.09	Tipo de	.02	.07	Tipo de
		Con IA – no info FP	.10	.21	error	.08	.19	error
		Con IA – no info FN	.02	.09		.00	.02	

Nota. FP, indica falso positivo, FN, falso negativo. La columna Dif. indica que factor, información sobre el error o tipo de error, demostró un efecto principal en el análisis de varianza con la ratio de clasificaciones erróneas en cada tipo de estímulo, 60/40 o 40/60, como variable dependiente.