LA INFLUENCIA DE LOS ALGORITMOS EN LAS DECISIONES Y JUICIOS HUMANOS

Experimentos en contextos de política, citas y arte

PROGRAMA DE DOCTORADO: PSICOLOGÍA

Tesis doctoral realizada por Ujué Agudo Díaz

Dirigida por la Dra. Helena Matute Greño

Deusto
Universidad de Deusto

Deusto
Universidad de Deusto

LA INFLUENCIA DE LOS ALGORITMOS EN LAS DECISIONES Y JUICIOS HUMANOS

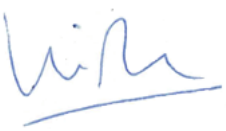
Experimentos en contextos de política, citas y arte

PROGRAMA DE DOCTORADO: PSICOLOGÍA

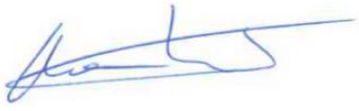
Tesis doctoral realizada por Ujué Agudo Díaz

Dirigida por la Dra. Helena Matute Greño

La doctoranda



La directora



Bilbao, 21 de septiembre de 2021

Esta tesis fue realizada en el marco de los proyectos PSI2016-78818-R e IT955-16 concedidos a la Dra. Helena Matute por la Agencia Estatal de Investigación del Gobierno de España y por el Departamento de Educación, Universidades e Investigación del Gobierno Vasco respectivamente.

Agradecimientos

Supongo que todos los trabajos de doctorando se sobrentienden como un proceso de colaboración. Este es, desde luego, ejemplo de trabajo de equipo. Un trabajo reflejado en cientos de conversaciones y discusiones sobre métodos, tecnología y comportamiento humano. Por eso esta tesis está redactada en primera persona del plural. Porque si algo se aprende en un proceso de doctorado es la importancia de mostrar reconocimiento hacia las fuentes. Por ello, mi agradecimiento infinito a mi directora Helena, por permitirme formar parte de Labpsico cuando aún no era ni su doctoranda, aceptar mi solicitud para emprender este proceso juntas y enseñarme tanto a diario. Mi agradecimiento también a mis compañeros del lab, los actuales y los que ya no están, por responder siempre a mis dudas a pesar de la barrera de la distancia. A mi gente de Biko, por ayudarme a compaginar ambos mundos, y especialmente a Miren por participar en los dos últimos experimentos de esta tesis, a Amaya por su asesoramiento a nivel visual y, sobre todo, a mi compañero de Bikolabs, Karlos, por enriquecer mis reflexiones, hacer suya esta tesis y sus experimentos, y facilitarme tanto el camino. Y, por último, mi agradecimiento a mi familia por… todo.

Índice

Aspectos Éticos y Ciencia Abierta

El proyecto de investigación de esta tesis fue revisado y aprobado por el Consejo de Revisión Ética de la Universidad de Deusto (Ref: ETK-7/18-19) al considerar que se ajustaba a los principios metodológicos, éticos y jurídicos exigidos y no observarse ningún riesgo para los participantes. La participación de los participantes en los experimentos fue voluntaria y sus respuestas fueron recogidas online de forma anónima con su permiso explícito, al aceptar su envío al final de cada experimento. Ninguna información personal fue recopilada.

Por consideraciones éticas, los contextos de decisión planteados y los candidatos mostrados en los experimentos de esta tesis fueron ficticios. Además, aunque simulamos que los participantes interactuaban con un algoritmo de IA, en realidad este algoritmo no existía.

Los datos en bruto de todos los experimentos se encuentran disponibles en abierto y pueden ser descargados desde Open Science Framework: (https://osf.io/t7dbv/?view_only=21d75e0817984300bea0340bc64d7539). Además, algunos de los experimentos (Experimentos 6, 7 y 9) fueron pre-registrados en AsPredicted.org, tal y como se indica en el detalle de estos experimentos.

Resumen

La inteligencia artificial forma ya parte de nuestro día a día, y muchas veces no somos conscientes de ello. Son algoritmos de inteligencia artificial los que nos recomiendan qué libro leer, qué productos adquirir, qué nueva serie ver, dónde alojarnos o comer, o con quién salir. Su amplia penetración ha generado un debate sobre hasta qué punto su presencia puede influyendo en nuestras decisiones. Este debate no ha tenido, por el momento, un reflejo muy amplio en la investigación empírica. Por ello, en este trabajo comprobamos si los algoritmos pueden influir las decisiones con diferentes tipos de recomendaciones (explícitas y encubiertas), en contextos de decisión de impacto para las personas como la política y las citas románticas. Además, exploramos cómo las personas juzgan el desempeño de los algoritmos en un terreno donde no es tan común la interacción con ellos: el campo del arte. Nuestros resultados, a lo largo de nueve experimentos, muestran que la mera recomendación de un supuesto algoritmo puede influir en las decisiones humanas y que el desempeño de la inteligencia artificial en el terreno artístico resulta minusvalorado cuando el público conoce su autoría. Comprender mejor cómo los juicios y decisiones humanas se ven afectados en la interacción con sistemas algorítmicos resulta esencial para evitar subestimar el efecto de la recomendación y la presencia del algoritmo en nuestras vidas.

Parte I. Introducción

Capítulo 1. **La creciente presencia de los algoritmos en las decisiones**

Cuando se habla del impacto de los algoritmos de inteligencia artificial (IA) en nuestras vidas, tendemos a recrear mentalmente escenas de películas de ciencia ficción y a asumir que estamos hablando de un futuro lejano. Sin embargo, su presencia en el día a día es ya una realidad.

Por un lado, encontramos algoritmos en decisiones que nos afectan como ciudadanos, pero ante las que no podemos intervenir. Instituciones, empresas y otros tipos de organismos utilizan hoy algoritmos de IA para evaluarnos y asignarnos un puntaje con el que determinar si concedernos o no un préstamo bancario (Cheney, 2016), una hipoteca (Greenawalt, 2018) o un seguro de vida (A. Chen, 2019); decidir a qué recursos de salud podemos acceder (Obermeyer y cols., 2019), qué precios pagaremos online (L. Chen y cols., 2016) o qué ofertas de empleo se nos mostrarán (Angwin, Scheiber y cols., 2017); si se nos debe considerar en un proceso de selección de personal (Dastin, 2018; Harwell, 2018, 2019), si merecemos ser depositarios de confianza en nuestro trabajo (Tucker, 2019) o si se nos puede considerar ciudadanos ejemplares (Botsman, 2017); decidir a qué plazas escolares podrán optar nuestros hijos (Marsh, 2019) o qué calificaciones recibirán (Duncan y cols., 2020); si el barrio en el que vivimos es susceptible de vigilancia policial preventiva (Lapowsky, 2018); cuáles deberían ser las condiciones de nuestra libertad condicional si acabamos presos

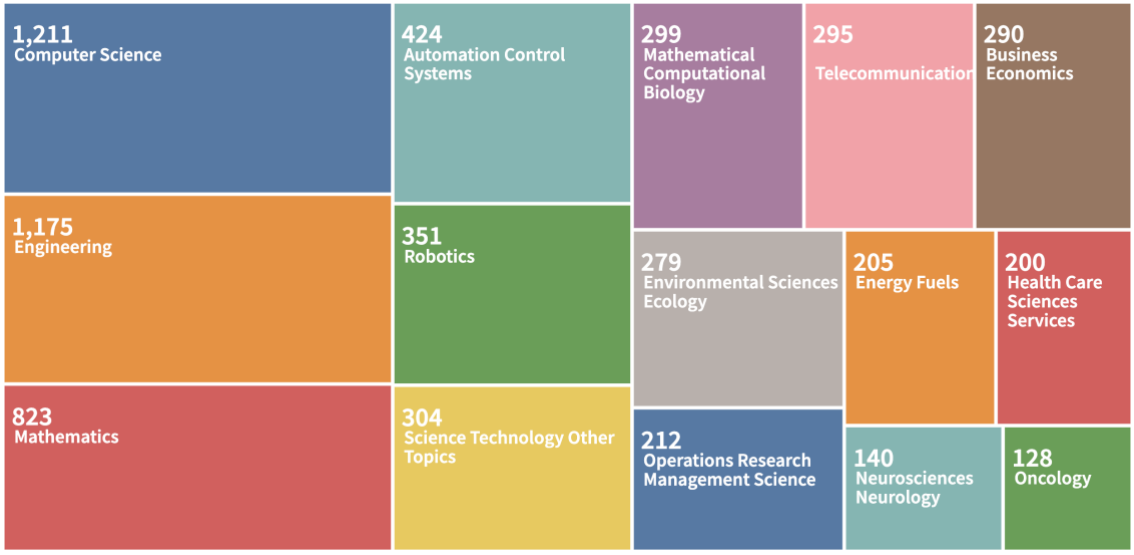
(Angwin y cols., 2016; Casacuberta y Guersenzvaig, 2018) o, incluso, cuál debería ser nuestra condena en un delito menor (Niiler, 2019).

Por otro lado, a nivel personal, también recibimos las recomendaciones de los algoritmos en decisiones cotidianas de forma más o menos explícita. Consejos sobre qué comprar (por ejemplo, Amazon; L. Chen y cols., 2016), qué película o serie ver (Netflix; Hosanagar, 2019), a qué información de actualidad prestar atención (Twitter, Facebook; Diakopoulos y Koliska, 2017; Pariser, 2011), qué música escuchar (Spotify; Bovenkamp, 2017), qué cuidador contratar (Harwell, 2018), o con quién tener una cita (Tinder u OKCupid; Duportail, 2019; Hern, 2014; Tiffany, 2019).

Aunque es abundante la investigación dedicada al impacto de los algoritmos en las decisiones, y a pesar de que se trata de un área claramente interdisciplinar, el foco de esta investigación tiende a centrarse en disciplinas relacionadas con los aspectos técnicos del algoritmo, otorgándose un menor peso a las disciplinas relacionadas con el comportamiento humano (véase Figura 1).

Figura 1

Búsqueda Realizada en Web of Science, a día 21 de septiembre de 2021, para Artículos Publicados en 2021 con los Términos "Algorithm" y "Decision Making"



Nota. El área "Psychology" no aparece en la figura por encontrarse fuera de los 15 primeros puestos, concretamente en el puesto 21 con 105 artículos. Tampoco se muestra "Behavioral Sciences", en el puesto 16 con 117 artículos. Neurociencias, en el puesto 14, sería la categoría más cercana a la disciplina de Psicología.

También existe un importante esfuerzo de investigación para detectar los potenciales efectos (y riesgos) de los algoritmos de IA en nuestras vidas a medio y largo plazo, aunque gran parte del debate generado se centra en situaciones hipotéticas y lejanas en el tiempo, como es el caso de la conducción autónoma que ninguna compañía automovilística estima que llegue a alcanzarse completamente antes de varias décadas (Sabatini, 2017). Mientras, sin embargo, existe poca investigación psicológica dedicada a analizar qué influencia tienen los algoritmos de IA en los juicios humanos y las decisiones cotidianas. Este es el tema central de esta tesis.

Pero antes de entrar en materia, resulta necesario aclarar previamente qué se entiende por algoritmo de IA y cuáles son sus rasgos distintivos.

Definición de algoritmo

El término *algoritmo* comenzó a popularizarse dentro de la jerga informática en el arranque de la década de 1960 (Miyzaki, 2012) y hacía referencia al conjunto de reglas e instrucciones de un ordenador para actuar sobre un conjunto de datos con el objetivo de resolver un problema (Barocas y Selbst, 2016; Lee, 2018). En sus inicios, estas reglas eran dictaminadas por expertos humanos. Sin embargo, en los últimos años se entiende por algoritmo de IA a la tecnología que, dados unos datos y un resultado deseado, tiene la capacidad de identificar cuáles deberían ser las reglas más adecuadas para resolver el problema (Duan y cols., 2019) y, en algunos casos, es capaz de aprender sobre el desempeño de su cometido (Fry y Krzywinski, 2019). Se trata de tecnología que cuenta con la capacidad para procesar grandes cantidades de datos (Castelo, 2019; Spiesel, 2020) y con ciertas habilidades cognitivas más o menos desarrolladas, como la percepción, el aprendizaje, el razonamiento o la toma de decisiones (Eiband y cols., 2019).

A menudo los algoritmos poseen también la facultad de mejorar su rendimiento mediante aprendizaje automático (*machine learning*), ofrecen predicciones probabilísticas, detectan patrones, automatizan tareas (Elish y Boyd, 2018), personalizan su respuesta o toman decisiones sin intervención humana (Lee, 2018; Spiesel, 2020). Cuando en esta tesis se utiliza el término algoritmo nos referimos

a esta tecnología que incorpora IA y cuenta con capacidades cognitivas flexibles y la habilidad de aprender de la experiencia.

Estereotipos sobre los algoritmos

Más allá de su definición, lo que resulta especialmente relevante para el propósito de este trabajo es el imaginario que personas e instituciones han generado respecto a las capacidades de los algoritmos (Bucher, 2017). Según Sundar (2008), las personas asocian a los algoritmos ciertos estereotipos, proceso que el autor ha acuñado como *heurístico de la máquina* (*machine heuristic*). Los heurísticos son atajos de pensamiento que utilizamos las personas de forma automática para simplificar los procesos de decisión. Se trata de estrategias mentales sencillas, conscientes o inconscientes, que permiten responder rápidamente a la demanda del contexto de decisión sin tener que razonar de forma profunda, recopilar mucha información o emplear excesivo tiempo o energía (Gigerenzer, 2015; Tversky y Kahneman, 1974). En lo que se refiere a los algoritmos, el heurístico de la máquina desencadenaría el estereotipo de que los algoritmos de IA se encuentran libres de sesgo (Araujo y cols., 2020; Sundar, 2008), son más eficaces, objetivos y precisos que los humanos en ciertas tareas (Sundar, 2008; Sundar y Kim, 2019), al tiempo que más inflexibles, fríos y carentes de emoción (Sundar, 2020). El heurístico de la máquina ha sido abordado empíricamente por diferentes autores. Por ejemplo, Waddell (2019), en un experimento sobre el consumo de noticias, encontró que las personas atribuyen más credibilidad e imparcialidad a la información proveniente de fuentes automatizadas que a la firmada por periodistas humanos. También Pan y colaboradores (2007)

investigaron esta credibilidad de los algoritmos en un experimento donde se mostraban diferentes listados de resultados de Google a los participantes, variando la posición en la que se ubicaba la entrada más relevante. Los participantes depositaron su confianza en el buscador al optar habitualmente por escoger el primer resultado mostrado por Google, a pesar de que el primer enlace no siempre resultaba el más pertinente para su propósito.

Esta idea de que los algoritmos son capaces de lograr un desempeño superior al de los humanos en muchos dominios y tareas (Dietvorst y cols., 2015) ha sido popularizada y acrecentada por empresas y medios de comunicación que a diario publican nuevos titulares sobre el tema. Tareas como la selección de los candidatos más adecuados para un puesto profesional (Cappelli y cols., 2018), la detección de un paro cardíaco a través de una llamada telefónica (Peters, 2018), la predicción de los resultados en terapia de pareja (Nasir y cols., 2017), la identificación de cánceres en radiografías (Sample, 2020) o la reducción de errores de diagnóstico médico (Wang y cols., 2016). Un ejemplo es el de la compañía Babylon Health que afirma que su tecnología de IA ha superado a médicos experimentados en la evaluación de los síntomas de diferentes enfermedades, logrando diagnosticar correctamente el 80% de las enfermedades, frente al diagnóstico de los profesionales médicos cuyo porcentaje de acierto varía entre el 64% y el 97% (Sandle, 2018).

Sesgos en los algoritmos

Como hemos mencionado, según Sundar (2008), las personas considerarían a los algoritmos, además de superiores en eficacia (Beer, 2017; Castelo, 2019; Taddeo y

Floridi, 2018), objetivos en sus decisiones (Beer, 2017; Collins, 2018; Kahneman y cols., 2016; Miller, 2018; Sundar, 2008). Pero, ¿realmente son objetivos? Como apunta Willson (2017, p.5), aunque los algoritmos se presentan como mecanismos de decisión neutral, "no existe un algoritmo en bruto que pueda considerarse una instrucción sencilla y objetiva". La objetividad de los algoritmos de IA puede ser puesta en cuestión desde el momento en el que no son entes aislados, sino que se encuentran insertos en sistemas políticos, técnicos, culturales y sociales más amplios (Birhane, 2019), con el consecuente riesgo que esto supone de impactar de forma considerable en las vidas de las personas.

Además, a pesar de que hablemos de sistemas automatizados, lo habitual es que se produzca intervención humana en alguna de las etapas del ciclo de vida de estos algoritmos de IA, por lo que, aun pudiendo llegar a aislarse, son susceptibles de desarrollar sesgo en cualquiera de estas etapas (Xu y Doshi, 2019). Por ejemplo, resulta muy conocido el caso de sesgo en los datos del algoritmo COMPAS, utilizado en algunos estados norteamericanos para calcular la probabilidad de reincidencia de un acusado (Angwin y cols., 2016). Se trata de una puntuación de riesgo que ha sido señalada por poseer un claro sesgo racista en sus evaluaciones. Los acusados de raza negra tienen el doble de probabilidad que los de raza blanca de ser clasificados erróneamente como de alto riesgo de reincidencia violenta, mientras que los acusados de raza blanca se clasifican como de bajo riesgo un 63.2% más a menudo que los acusados de raza negra, también erróneamente (Larson y cols., 2016).

Obermeyer y colaboradores (2019) muestran otro ejemplo similar de sesgo en el algoritmo. Los investigadores encontraron que el algoritmo utilizado para decidir la

atención médica de alrededor de 200 millones de estadounidenses perjudicaba a las personas de raza negra. El motivo era que el algoritmo usaba el coste de la atención médica como variable para determinar las necesidades de salud de los pacientes y predecir quién necesitaría atención adicional. Los datos de entrenamiento de la IA estaban sesgados, dado que históricamente se ha invertido menos en el cuidado de pacientes negros que de blancos. Al basarse en ellos para decidir, el algoritmo reducía a más de la mitad el número de pacientes negros susceptibles de poder acogerse a un programa de atención médica. Otro caso de sesgo, que además causó una gran repercusión mediática, fue el del agente conversacional de Microsoft en Twitter (Victor, 2016), que tuvo que ser retirado por desarrollar actitudes racistas al aprender de su interacción con usuarios después de ser publicado. Este sería un sesgo en la retroalimentación del algoritmo, es decir, un sesgo desarrollado en la interacción con usuarios reales tras su puesta en marcha y no durante su entrenamiento inicial.

Los algoritmos además pueden mostrar otros sesgos, no solo de carácter racista como los ejemplos mencionados. El algoritmo de publicidad de Facebook, por ejemplo, ha sido acusado de comportamiento discriminatorio al permitir que sus anunciantes puedan seleccionar a la audiencia destino no solo por raza, sino también por sexo o edad (Angwin, 2017; Angwin, Scheiber y cols., 2017; Angwin, Tobin y cols., 2017).

Independientemente de la etapa del ciclo de vida de la IA en la que se desarrolle el sesgo, el proceso de aprendizaje continuo de la IA puede contribuir a perpetuarlo, como señala O'Neil (2018), al retroalimentarse con resultados sesgados. Además, la IA puede provocar que el sesgo se amplifique, dado que en muchas

ocasiones se utilizan modelos pre-entrenados como base para desarrollar otros nuevos modelos o se reutilizan aplicándose a otros contextos. Si un modelo pre-entrenado presenta sesgo, los siguientes algoritmos que se basan en él pueden heredar su sesgo, extendiendo y amplificando el impacto negativo de este. Es el caso que padeció en persona la investigadora y activista Buolamwini (2016) cuando aún era estudiante de informática en Georgia Tech. Por aquel entonces, el robot social con el que realizaba sus prácticas no era capaz de identificar su cara debido al color de piel de Boulamwini. Tiempo después, mientras participaba en una demostración de otro robot social en Hong Kong, de nuevo su rostro fue el único que la IA no pudo reconocer. El robot utilizaba el mismo algoritmo de reconocimiento facial que el de la investigadora en Georgia. En palabras de la propia Buolamwini (2016, 3:00): "el sesgo algorítmico puede viajar tan rápido como el tiempo que lleva descargar archivos de Internet".

La posible introducción de sesgo en alguna de las etapas del ciclo de vida de la IA o en varias de ellas puede deberse a un acto totalmente accidental, aunque también puede responder intencionalmente al interés del ente propietario del algoritmo. Por ejemplo, Eslami y colaboradores (2017) encontraron que el algoritmo de calificación de Booking.com incrementaba al alza las puntuaciones de los hoteles en su plataforma con el objetivo intencionado de mejorar la percepción de estos ante sus usuarios. Por su parte, la empresa Uber fue acusada de uso interesado de su algoritmo para empujar a trabajar en exceso a sus conductores (Scheiber, 2017). Para ello, utilizaba una funcionalidad similar a la auto-reproducción de vídeo de Netflix o Youtube en su aplicación interna, que tramitaba automáticamente la siguiente solicitud de traslado antes de que el conductor terminara el traslado anterior. De forma semejante actuaba

el algoritmo de varias aerolíneas comerciales separando a los pasajeros que viajaban juntos para forzarles a pagar por sentarse unos al lado de los otros (Coffey, 2018). Esta mala praxis no implica necesariamente que la empresa haya programado de forma explícita esta funcionalidad. Un algoritmo cuyo objetivo sea maximizar beneficios podría aprender por sí mismo que una vía económica rentable sea separar a los pasajeros.

La opacidad de los algoritmos

En muchas de las situaciones mencionadas, auditar externamente el algoritmo antes de su implantación para garantizar su objetividad no hubiera sido posible debido, fundamentalmente, a dos razones. Por un lado, muchos de los algoritmos que se utilizan hoy en día son de propiedad privada y, por tanto, inaccesibles al escrutinio externo (Willson, 2017). Es el caso del algoritmo de COMPAS, mencionado previamente, que determina las condiciones de la libertad condicional de los acusados en algunos juzgados de EE. UU. (Angwin y cols., 2016). Para poder analizar la objetividad del algoritmo, los investigadores tuvieron que llevar a cabo un proceso de ingeniería inversa. A partir de las puntuaciones de dos años antes generadas por el sistema COMPAS, junto con los antecedentes penales de los acusados y su nivel de reincidencia, los investigadores consiguieron deducir cuáles habían sido los factores ponderados por el algoritmo para tomar sus decisiones. Un modelo de auditoría que no siempre resulta posible por falta de acceso a los datos, recursos necesarios, etc.

Por otro lado, muchos de los algoritmos de IA resultan ser cajas negras que no permiten conocer la lógica de su funcionamiento por su propia complejidad técnica. Es

el caso de los algoritmos de aprendizaje automático que pueden abordar problemas complejos en situaciones donde aplicar reglas simples y lógicas no derivaría en buenos resultados. Como contrapartida, estos algoritmos realizan predicciones que resultan inescrutables debido a la propia abstracción del sistema (Challen y cols., 2019), el cual, a medida que se sofistica, reduce al máximo las posibilidades de comprensión humana (Merino, 2018). En estas situaciones, si bien las piezas que componen el algoritmo pueden resultar entendibles individualmente, no así su conjunto. Los algoritmos de aprendizaje automático se basan en lo que se conoce como *redes neuronales artificiales*, compuestas por una serie de capas con multitud de "neuronas" artificiales interconectadas entre sí (Castelvecchi, 2016). El algoritmo modifica el peso que otorga a estas conexiones durante su aprendizaje, hasta determinar cuál es el valor más adecuado para lograr el resultado con mayor precisión (Burrell, 2016). La complejidad de relaciones en este sistema, que divide el problema en miles o millones de neuronas según la red empleada, provoca que la toma de decisiones resulte opaca (Kenyon, 2018). Así, un algoritmo puede hoy resolver un problema hasta el momento irresoluble sin que sea posible comprender cómo lo ha conseguido. Es el caso del algoritmo Deep Patient, que resultó sorprendentemente eficaz a la hora de predecir la esquizofrenia, un trastorno difícil de pronosticar por los médicos, sin que ningún experto pudiera deducir qué patrones, datos o señales habían sido los utilizados por el algoritmo para determinar el diagnóstico (Knight, 2017).

Sin embargo, la naturaleza inescrutable de los algoritmos se convierte en un grave problema cuando su desempeño resulta inadecuado y deriva en consecuencias perjudiciales para los ciudadanos. Es el caso de la falta de precisión de los algoritmos

de reconocimiento facial o de predicción de crímenes usados por cuerpos policiales en todo el mundo (Burgess, 2020; Greig, 2018; Koebler, 2020), por los que una persona puede ser acusada de un crimen que no ha cometido a causa de un fallo en el reconocimiento facial del algoritmo (Hill, 2020). Un error que pueden deberse a un sesgo muestral, por falta de representación igualitaria de todo tipo de personas en el entrenamiento de la IA, y a la alta dependencia que estos sistemas tienen de elementos ajenos a la persona a identificar, como la iluminación del lugar o el entorno en el que se encuentra el individuo (Harlan y Schnuck, 2021).

Otro ejemplo de desempeño erróneo con consecuencias para los ciudadanos se ha producido recientemente en la distribución de las vacunas contra la COVID a los sanitarios de la Stanford Medicine. El algoritmo priorizó a médicos de alto rango sobre médicos residentes cuando eran a menudo estos últimos quienes corrían el mayor riesgo debido a su mayor interacción con los pacientes (C. Chen, 2020).

Además, en ocasiones las situaciones problemáticas se producen porque los algoritmos basan su lógica en criterios pseudocientíficos. Es el caso de la IA de reconocimiento de emociones mediante la detección de posturas o expresiones faciales (Varghese, 2019; Wiggers, 2019), una tecnología que no cuenta con base científica que la avale (Barrett y cols., 2019).

Cada vez son más las voces que alertan sobre el riesgo de la toma automatizada de decisiones y las consecuencias de aceptar la opacidad de los algoritmos a corto, medio y largo plazo (Diakopoulos, 2016; Gillespie, 2014; Zittrain, 2019). Por ejemplo, Zittrain (2019), profesor de derecho en la Harvard Law School, ha señalado que el número de pruebas que son necesarias para descubrir las

interacciones adversas de la IA aumenta exponencialmente a medida que proliferan los algoritmos opacos.

Resulta reseñable que, sin embargo, los algoritmos y sus recomendaciones ya se encuentren insertos en muchos de los sistemas que determinan nuestras decisiones y juicios diarios. Y, por ello, no resulta recomendable concebir el algoritmo como una simple pieza tecnológica a estudiar de forma aislada al contexto psicológico y social donde se implementa (Beer, 2017), dado que se encuentra involucrado en multitud de procesos sociales (Kitchin, 2017; Neyland y Möllers, 2016; Willson, 2017) y psicológicos (Eslami y cols., 2015; Sundar, 2020).

Influencia de los algoritmos en las decisiones y juicios humanos

Por todo lo mencionado anteriormente, nuestro propósito con esta tesis es comprobar de forma empírica si los algoritmos influyen en las decisiones y juicios humanos en los que ya están presentes. Consideramos que este resulta un campo de investigación crítico y muy interdisciplinar que, sin embargo, ha sido muy poco abordado desde la Psicología hasta el momento, por lo que contribuir en esta línea es la aspiración de esta tesis. Y lo hacemos desde dos perspectivas que se reflejan en las dos series experimentales de este trabajo. Una primera serie sobre cómo las recomendaciones algorítmicas pueden estar siendo usadas como herramientas persuasivas. Y una segunda serie, derivada de los resultados obtenidos en la primera serie experimental, sobre cómo las percepciones y estereotipos respecto a las capacidades de los algoritmos de IA pueden estar impactando en las decisiones y juicios humanos.

Parte II. La influencia de la recomendación algorítmica en las decisiones

Capítulo 2. **La recomendación algorítmica como herramienta persuasiva**

¿A qué nos referimos cuando hablamos de la influencia de los algoritmos y de su uso de la recomendación como herramienta persuasiva? Desde la Psicología Social se considera el concepto de influencia como un término más amplio que abarca al término de persuasión. Mientras que la persuasión aborda el estudio del cambio de actitudes, perseguido de forma intencional, la influencia contempla también la variación de percepciones, opiniones, actitudes y comportamientos que pueden darse tanto de forma intencional como no intencional en multitud de contextos (Morales y cols., 2007). Sin embargo, los términos influencia y persuasión se utilizan habitualmente como sinónimos, y así lo haremos a lo largo de esta tesis. En el ámbito tecnológico, se utiliza el concepto de *tecnología persuasiva* (Al-slaity, 2021; Fogg, 2002) para referirse a aquella capaz de influir y provocar un cambio de comportamiento o de actitud en las personas (Harris y cols., 2017; Oduor y cols., 2014), motivo por el que, en principio, hablaremos de los algoritmos como herramientas de persuasión, pero teniendo en cuenta que, desde el punto de vista psicológico, está aún pendiente demostrar que verdaderamente los algoritmos sean capaces de influir en las personas.

Cuando pretendemos estudiar el efecto de la llamada tecnología persuasiva (en nuestro caso, la influencia de los algoritmos de IA) en los juicios y decisiones humanos, encontramos que no existe una teoría unificada sobre la persuasión, y menos aún

sobre la persuasión en tecnología. Además, como ya comentamos, la gran mayoría de las investigaciones realizadas hasta la fecha provienen del ámbito tecnológico, por lo que a menudo nos encontramos sin referentes teóricos para abordar nuestra investigación. En cualquier caso, y a nivel general, podemos destacar que, mientras que algunos investigadores abordan la persuasión desde modelos relacionados con la aceptación de la tecnología, como la Teoría Unificada de Uso y Aceptación de la Tecnología (Venkatesh y cols., 2003) o el Modelo de Aceptación de la Tecnología (Davis, 1985), otros autores se apoyan en modelos clásicos de la Psicología Social, como los modelos de cambio de actitudes a través de la comunicación, como el Modelo de Probabilidad de Elaboración de la Persuasión (Petty y Cacioppo, 1986), o el Modelo Heurístico-Sistemático (Chaiken y cols., 1989). Estos modelos postulan que la eficacia de una comunicación persuasiva dependerá tanto del procesamiento motivado del receptor para analizar los argumentos del mensaje persuasivo (a través de la llamada *ruta central*), como de señales o claves heurísticas que contribuyan a la aceptación de ese mensaje persuasivo de forma menos reflexiva y más automática, como el atractivo o credibilidad del emisor, o la repetición del mensaje (a través de la llamada *ruta periférica*) (Morales y cols., 2007).

Otros modelos de la Psicología Social mencionados en los trabajos sobre persuasión en tecnología son los que abordan la influencia desde la perspectiva de la relación interpersonal, es decir, cuando existe una intencionalidad por parte del agente de influencia de modificar las actitudes o el comportamiento de las personas. El marco teórico de referencia en esta área de la Psicología Social (Cialdini, 1993) propone una serie de principios de influencia (principio de reciprocidad, de validación

social, de coherencia, de simpatía, de escasez o de autoridad) que actúan como reglas de comportamiento o heurísticos para facilitar el ejercicio persuasivo en las relaciones sociales. En tecnología, este marco teórico es referenciado cuando se analiza la influencia que la tecnología ejerce al ostentar el papel de agente persuasivo que interactúa con personas (Al-slaity, 2021) en formato de robot social o de agente conversacional (Yoo y cols., 2013).

Sin embargo, existen otras muchas situaciones en las que los algoritmos de IA pueden ejercer influencia en las decisiones y que son de gran interés, pero que los modelos teóricos de persuasión comentados no abordan, ni realizan predicciones sobre ellos que nos ayuden en su estudio. Se trata de situaciones donde:

1. El algoritmo toma la decisión de forma completamente autónoma (Araujo y cols., 2020) sin intervención humana en el ajuste o la validación de esta decisión (Elish y Boyd, 2018; Fry y Krzywinski, 2019). Se trata de una situación donde la persona solo puede detener la decisión. Un ejemplo de este tipo de influencia es la función de auto-reproducción de Youtube y de otras plataformas de visualización de vídeo bajo demanda. Al terminar de consumir un contenido en ellas, el siguiente se reproduce de forma automática y el usuario solo dispone de unos segundos para interrumpir el proceso o debe configurar la plataforma del servicio para evitar que esta función esté activa.

2. El algoritmo selecciona, filtra e incluso censura la información a mostrar (M. P. Burden, 2012; Gillespie, 2014), por lo que psicológicamente pueden

impactar en las decisiones y opiniones de los usuarios. Conforme el volumen de información crece en nuestra sociedad, más peso adquieren los algoritmos que clasifican, ordenan y filtran esa información (Alvarado y cols., 2019). En este proceso de selección personalizada del contenido para cada usuario, los algoritmos pueden modificar las creencias de estos. Por ejemplo, Instagram o Facebook han sido acusados de "encerrar" a sus usuarios en *cámaras de eco digitales* que reducen su exposición a otros puntos de vista al sugerirles únicamente contenido que refuerza sus creencias previas (Pariser, 2011).

3. El algoritmo puede lograr la conformidad del usuario mediante su recomendación explícita (Susser y cols., 2019). Es el caso de Netflix, cuyo sistema de recomendación supone el 80% del consumo de vídeo en la plataforma (Thurman y cols., 2019).

4. El algoritmo puede influir a las personas de forma encubierta, aplicando para ello principios de persuasión interpersonal (Cialdini y Sagarin, 2005) o explotando los sesgos cognitivos de las personas de forma no transparente para el usuario (Susser y cols., 2019). Por ejemplo, según Epstein y Robertson (2015), los algoritmos podrían influir en los votantes indecisos en unas elecciones políticas al manipular el orden de los resultados de un buscador de contenidos, de forma que la información sobre ciertos candidatos políticos apareciese en las primeras posiciones (efecto de primacía).

Dado que nuestro interés, desde el área de la Psicología, se encuentra en estudiar aquellas interacciones en las que la decisión final descansa en las personas, y dado que, como ya comentamos en la Introducción, no conocemos aún cuál es el efecto persuasivo de la presencia de los algoritmos en la vida de las personas, creemos que es prioritario investigar si realmente se produce (y en qué condiciones) la supuesta influencia de los algoritmos sobre los juicios y la toma de decisiones, en lugar de evaluar los méritos y posibles aplicaciones en el ámbito de la tecnología de las diferentes propuestas teóricas sobre persuasión.

Por tanto, profundizaremos, en esta serie experimental, en los dos últimos tipos de influencia mencionados, la recomendación explícita y la encubierta, que son aquellas en las que el peso de la decisión recae en la persona. Además, lo haremos desde una perspectiva empírica, lo más neutral posible desde el punto de vista teórico, ya que creemos que solo conociendo bien el fenómeno en primer lugar será posible posteriormente desarrollar las teorías más adecuadas para poder predecirlo y controlarlo.

De hecho, es interesante destacar que existen en la actualidad dos líneas de investigación en relación a esta posible influencia de los algoritmos que mantienen visiones contradictorias sobre el efecto del consejo algorítmico en las personas: la literatura de la *apreciación al algoritmo* y la de la *aversión al algoritmo*.

Por un lado, la literatura de apreciación del algoritmo (*algorithm appreciation*) considera que se atribuye un mayor valor y una mayor confianza a la recomendación algorítmica que al consejo humano. Por ejemplo, en el trabajo de Logg y colaboradores (2019) se muestra cómo las personas prefieren las estimaciones de un algoritmo frente

a la de otras personas, en tareas tan variadas como la estimación del peso de un individuo a partir de una fotografía o de la popularidad de una canción en una lista de éxitos.

La apreciación del algoritmo ha resultado ser un fenómeno complejo que varía dependiendo de diversos factores: el agente de decisión con quien se compare el algoritmo (Logg y cols., 2019), las capacidades requeridas en la tarea de decisión (Lee, 2018), la fuente de datos que alimenta al algoritmo (Thurman y cols., 2019), el impacto de la decisión, o el contexto de esta (Araujo y cols., 2020). Por ejemplo, según indican Araujo y colaboradores (2020), se produce apreciación algorítmica cuando la recomendación se compara con el juicio de otras personas o incluso el juicio propio, en tareas que requieren habilidades mecánicas, en decisiones con alto impacto en la vida de las personas, y en contextos como la justicia o la salud.

Este fenómeno de la apreciación, relativamente reciente, podría a su vez hallarse relacionado con un fenómeno previo, con mayor recorrido en la literatura científica, conocido como *sesgo de automatización* (*automation bias*). El sesgo de automatización es la inclinación de las personas a aceptar las recomendaciones explícitas de los sistemas informatizados de ayuda, habitualmente en contextos profesionales, hasta el punto de llegar a delegar completamente en ellos la toma y la responsabilidad de la decisión, sin cuestionar si esta es correcta, o al menos óptima, ni buscar información que la ponga en entredicho (Challen y cols., 2019; Cummings, 2004; Geslevich, 2019).

El sesgo de automatización ha sido ampliamente documentado en la tres últimas décadas en variedad de dominios, tales como la aviación, la salud o el contexto

militar (Sundar y Kim, 2019) con bastante robustez en todos ellos (Goddard y cols., 2012). Por ejemplo, Johnson y colaboradores (2002) señalaron el riesgo del sesgo de automatización en un experimento con pilotos comerciales e instrumentos de planificación de vuelos. En él, el 40% de los pilotos mostraron un exceso de confianza en el sistema automático, aceptando decisiones de la máquina que resultaban considerablemente peores que el nivel óptimo.

En una posición contraria a la de la literatura de la apreciación al algoritmo se sitúa la literatura de aversión al algoritmo (*algorithm aversión*). Desde esta se afirma que las personas muestran desconfianza hacia el consejo algorítmico explícito, en comparación con el de un humano, y que son reticentes a utilizar la recomendación del algoritmo incluso cuando su consejo es mejor (Dietvorst, 2016). Así lo respalda, por ejemplo, la investigación de Castelo y colaboradores (2019), quienes encontraron que proporcionar pruebas del rendimiento superior de los algoritmos mejora la confianza en su recomendación, pero no es suficiente para que la gente prefiera el consejo algorítmico al de un experto humano.

Esta desconfianza hacia los algoritmos ha sido demostrada empíricamente por un buen número de estudios (véase Burton y cols., 2019; o Schwienbacher, 2020, para una revisión). Por ejemplo, Yeomans y colaboradores (2019) reportaron que las personas confiaban más en otras personas que en la sugerencia de un algoritmo en la recomendación de chistes. Por su parte, Dietvorst y colaboradores (2015) señalaron que se juzga más duramente a los algoritmos que a las personas cuando se les ve fallar.

Sin embargo, al igual que ocurría con la apreciación del algoritmo, la aversión al algoritmo tampoco parece ser un fenómeno sencillo, contándose en mayor número los moduladores propuestos sobre el efecto (por ejemplo, la subjetividad de la tarea, la experiencia de las personas que toman las decisiones, la familiaridad con el algoritmo, la expectativa de rendimiento de este, o la incertidumbre propia del contexto de decisión) que la evidencia empírica encontrada sobre cada uno de ellos (Dietvorst y cols., 2015).

El volumen de investigaciones realizadas bajo el marco teórico de la aversión al algoritmo es amplio, con trabajos que cuentan con más de dos décadas (por ejemplo, Arkes y cols., 1986; Dawes, 1979; Dijkstra, 1999; Dijkstra y cols., 1998; Dzindolet y cols., 2002; Einhorn, 1986; Grove y Meehl, 1996; Highhouse, 2008; Önkal y cols., 2009; Promberger y Baron, 2006; Sieck y Arkes, 2005; Sinha y Swearingen, 2001; o Wærn y Ramberg, 1996). Sin embargo, hay que señalar que mucha de la evidencia existente podría no ajustarse a la realidad de este fenómeno hoy en día dado que la tecnología evoluciona a gran velocidad y el paso del tiempo en este campo puede implicar un cambio total de paradigma.

De hecho, son pocos los estudios sobre aversión al algoritmo que utilizan el término *algoritmo* en sus experimentos. En estos trabajos, los sistemas de recomendación algorítmica son presentados a los participantes con términos como “modelo” (Dietvorst y cols., 2015, 2018; Önkal y cols., 2009), “fórmula” (Eastwood y cols., 2012), “ecuación estadística” (Sieck y Arkes, 2005), “programa” (Arkes y cols., 2007), “ordenador” (Longoni y cols., 2019), “software de predicción” (Berger y cols., 2020) o “sistema informático” (Prahl y Van Swol, 2017); términos que no son

representativos de la imagen que personas e instituciones han construido alrededor del concepto de algoritmo en los últimos años. Los algoritmos han evolucionado enormemente, pasando de ser simples fórmulas matemáticas en sus primeros años, a incorporar inteligencia artificial. Al mismo tiempo, la interacción con algoritmos de IA resulta cada vez más familiar a las personas al aumentar la presencia de sus recomendaciones y servicios en hábitos y decisiones cotidianas (Schwienbacher, 2020).

Tanto en las investigaciones sobre aversión como en las de apreciación no existe demasiada uniformidad en los métodos empleados, lo que complica obtener una clara conclusión sobre ambos fenómenos. Por ejemplo, mientras algunos estudios recogen si los participantes solicitan activamente la recomendación (Alexander y cols., 2018; Castelo y cols., 2019; Longoni y cols., 2019; Sieck y Arkes, 2005) o si declaran en autoinformes su disposición a aceptarla, muchos otros utilizan lo que se conoce como *sistema juez-asesor* (*judge-advisor system, JAS*). Según este procedimiento, primero los participantes elaboran un juicio bajo incertidumbre, para después recibir la recomendación algorítmica. Lo que se mide es si los participantes ajustan su juicio inicial tras revisar el consejo del algoritmo (Berger y cols., 2020; Logg y cols., 2019; Önkal y cols., 2009; Prahl y Van Swol, 2017).

Por otro lado, usar o no la recomendación explícita del algoritmo no supone el mismo impacto para los participantes en los estudios sobre apreciación y aversión. En algunos casos el acierto de los participantes se vincula a su retribución económica en el experimento, determinando su importe o el acceso a una bonificación extra (Alexander y cols., 2018; Berger y cols., 2020; Dietvorst y cols., 2015, 2018; Logg y cols.,

2019; Prahl y Van Swol, 2017), mientras que en otros el comportamiento de los participantes no es incentivado económicamente.

La complejidad tanto del fenómeno de la apreciación como del fenómeno de la aversión queda evidenciada en los resultados de algunos de los experimentos mencionados, en los que los investigadores reportan aversión al algoritmo en algunas de las condiciones experimentales, pero apreciación en otras (Alexander y cols., 2018; Berger y cols., 2020; Dietvorst y cols., 2015, 2018). Por ejemplo, en Dietvorst y colaboradores (2015), los participantes mostraban aversión cuando debían utilizar las previsiones de un algoritmo sin poder ajustarlas, pero apreciación cuando podían ajustarlas.

Si algo podemos concluir de estas investigaciones es que aún falta mucho para poder conocer en qué casos se produce apreciación y en qué casos aversión. En cualquier caso, tampoco esta cuestión parece prioritaria desde nuestro punto de vista en el momento actual. Es decir, saber si existe más aversión hacia los algoritmos o más apreciación no nos lleva a contestar a la pregunta que nos planteábamos al inicio de la tesis, es decir, si los algoritmos pueden o no influir en las decisiones humanas. De hecho, quizá la recomendación encubierta (o incluso una explícita) de un algoritmo podría influir en una decisión importante sin necesidad de que la persona llegue a apreciar necesariamente la recomendación algorítmica o al propio algoritmo; y podría suceder también que no se produjera influencia y que ello no implicara la existencia de una aversión hacia el algoritmo.

Por todo ello, una vez más llegamos a la conclusión de la necesidad de una investigación previa, mucho más básica, encaminada a mostrar si las personas se dejan

influir o no por las recomendaciones de los algoritmos en decisiones importantes. Por lo que hemos podido comprobar, existe muy poca investigación académica publicada sobre este punto. Esta primera serie experimental pretende contribuir en esta línea, comprobando la influencia de recomendaciones algorítmicas explícitas y encubiertas en dos contextos importantes para el bienestar de las personas: política y citas.

La recomendación explícita

Utilizamos recomendadores algorítmicos casi a diario: para realizar compras online, reservar vacaciones, descubrir nuevos cantantes, ver películas, buscar trabajo, encontrar pareja, informarnos, decidir dónde invertimos nuestro dinero o interactuar con nuestros conocidos (por ejemplo, Amazon, Booking, Spotify, Netflix, Linkedin, Tinder, Twitter, Wealthfront o Facebook respectivamente). Estas recomendaciones algorítmicas son la pieza clave en un buen número negocios. Según detalla Hosanagar (2019), un 80% del consumo de Netflix se origina en base a sus recomendaciones, el 35% de las ventas de Amazon o la mayor parte de las coincidencias en la aplicación de citas Tinder tienen como base la recomendación algorítmica.

Los sistemas de recomendación surgieron como herramientas para ayudar a los usuarios a consultar catálogos con grandes volúmenes de información. Ante un exceso de alternativas, la elección puede convertirse en un acto poco satisfactorio (Iyengar, 2011; Schwartz, 2005). Por ello, los recomendadores resultan herramientas útiles para identificar los artículos de mayor interés para los usuarios a partir de sus preferencias y comportamientos previos (Knijnenburg y cols., 2012; Seaver, 2019).

Aunque inicialmente estos sistemas de recomendación inferían las preferencias de los usuarios mediante su retroalimentación directa (los usuarios calificaban su satisfacción con los artículos del catálogo recomendados y consumidos), en los últimos años estos sistemas han incorporado algoritmos de IA que ponderan los datos de la interacción de los usuarios con la plataforma por encima de sus gustos declarados (Hallinan y Striphas, 2016). Por ejemplo, la aplicación de recomendación de candidatos de citas Tinder recoge de sus usuarios datos como la frecuencia y el momento de conexión, la etnia de los emparejamientos logrados, las palabras más utilizadas o el tiempo de visualización de las fotografías para determinar qué candidatos mostrar a cada usuario (Duportail, 2017).

A partir de datos como estos, los algoritmos predicen qué recomendaciones personalizadas tienen más probabilidad de generar retención, es decir, una mayor cantidad de consumo del usuario en la plataforma (que es el verdadero objetivo perseguido por los algoritmos), quedando la satisfacción del usuario relegada en importancia. Prácticamente toda la oferta mostrada en estos sistemas está mostrando recomendaciones personalizadas.

Por ello, los recomendadores pueden actuar como tecnología persuasiva en su objetivo por incrementar el consumo de los usuarios, o al recomendar artículos concretos de forma interesadas (Alslaity y Tran, 2019; Cosley y cols., 2003; Gretzel y Fesenmaier, 2006). Banker y Khetani (2019), por ejemplo, mostraron cómo la recomendación de un algoritmo puede influir en las decisiones de compra. En el primero de sus experimentos, los autores mostraron a los participantes seis baterías portátiles para teléfonos móviles. Según la condición, el algoritmo recomendaba a los

participantes la mejor opción de las presentadas, o una claramente inferior, o no sugería nada. Mientras que, en la condición de control donde no se recomendaba ningún producto, la opción inferior solo era elegida en el 9% de los casos, un 49% de los participantes optaban por ella cuando el algoritmo la mostraba como la recomendada.

La presencia de los algoritmos de recomendación es cada vez más habitual en nuestras vidas por lo que consideramos importante abordar cómo responderán las personas a ella en situaciones cotidianas, como plantearon Banker y Khetani (2019), pero también relevantes (no solo con elecciones sencillas como comprar una batería portátil para un teléfono móvil).

Como en un sistema de recomendación algorítmica el usuario siempre tiene la última palabra, se le presupone libertad de elección, al contrario que en un sistema de decisión autónomo (Araujo y cols., 2020). Sin embargo, resulta necesario comprobar empíricamente qué poder de influencia real tiene la recomendación algorítmica explícita en situaciones relevantes. Pero antes de hacerlo, describimos a continuación el otro tipo de influencia mencionado, la recomendación encubierta.

La recomendación encubierta

Como se mencionaba al comienzo de este capítulo, los algoritmos pueden tratar de influir sobre el comportamiento o la actitud de las personas mediante diferentes estrategias. Además de utilizar recomendaciones explícitas para persuadir, los algoritmos de IA también pueden usar técnicas encubiertas de manipulación al aprovecharse de los heurísticos y los sesgos humanos. Como mencionamos

anteriormente, los heurísticos son atajos mentales que permiten emitir respuestas rápidas a las demandas del entorno sin exigir un razonamiento profundo, una exhaustiva recopilación de datos, o un consumo excesivo de tiempo o energía. Se trata de reacciones predeterminadas, profundamente arraigadas en la mente humana, que permiten una toma de decisiones altamente eficiente en la mayoría de las ocasiones, pero que pueden conducir a error y derivar en sesgo cuando se aplican en decisiones que no son seguras o apropiadas (Kahneman, 2012).

El uso de heurísticos para condicionar los entornos de decisión e influenciar actitudes y comportamientos ha ganado popularidad en los últimos años, no solo en el campo de la tecnología sino en otras muchas áreas, en gran parte debido al éxito del libro *Un pequeño empujón* de Thaler y Sunstein (2009). En él, sus autores proponen el uso de *nudges* (traducido como *empujones* y entendidos como aspectos del contexto de decisión que desencadenan el uso de heurísticos) para orientar de forma encubierta el comportamiento de las personas en la dirección deseada, sin necesidad de coacción, prohibiciones o incentivos económicos (Thaler y Sunstein, 2009).

El concepto de empujón se apoya en la literatura previa sobre heurísticos y sesgos en los juicios y toma de decisiones (Gigerenzer, 2018; Tversky y Kahneman, 1974) y especialmente en las teorías de proceso dual (Thaler y Sunstein, 2009, p. 35), las cuales cuentan con un amplio recorrido (y debate) en la literatura psicológica (De Neys, 2021; Dewey, 2021; Evans, 2010; Evans y Stanovich, 2013; Stanovich y West, 2000). Según estas teorías, las personas poseerían dos formas de procesamiento cognitivo: por un lado, las personas contarían con un tipo de procesamiento reflexivo, consciente y de razonamiento deliberado (Sistema 2 o Tipo 2), que requeriría un

considerable volumen de recursos cognitivos (Banerjee y John, 2019; Hertwig y Grüne-Yanoff, 2017; Kahneman, 2012); y por otro lado, una forma de procesamiento automático e intuitivo (con diferente nombre según los autores, siendo los más comunes Sistema 1 o Tipo 1), que utilizaría heurísticos para tomar decisiones de forma rápida, aunque sería susceptible a sufrir sesgos sistemáticos.

Dado que en muchas ocasiones nuestras decisiones estarían involuntariamente guiadas por el Sistema 1, los empujones se plantean como mecanismos para dirigirlas. Y aunque estos empujones pueden diseñarse para apelar a cualquiera de los dos tipos o sistemas de procesamiento, la mayoría se diseñarían para aprovecharse de las carencias y heurísticos del Sistema 1 (Caraban y cols., 2019; Hertwig y Grüne-Yanoff, 2017), de forma no-transparente o encubierta, es decir, de manera que las intenciones que motivan el empujón no resulten salientes (Caraban y cols., 2019; Hortal, 2019).

No existe un inventario completo de empujones ni una categorización de estos que haya sido ampliamente aceptada (véase como ejemplo, las diversas categorías propuestas por Hansen y Jespersen, 2013; o Meske y Potthoff, 2017). Aun así, describimos algunos de estos empujones con el objetivo de clarificar más el concepto:

- Opción por defecto (*default*): Establecer opciones de decisión seleccionadas de forma predeterminada para que adquieran vigencia a menos que la persona especifique lo contrario de forma activa. Este tipo de empujón habría permitido, por ejemplo, aumentar el porcentaje de personas que consienten la donación de órganos en algunos países al establecerse en

ellos la donación como opción predeterminada (E. J. Johnson y Goldstein, 2003).

- Prueba social (*social proof*): Mostrar el comportamiento mayoritario del grupo social al que la persona pertenece para provocar que esta lo asuma como normativo y lo imite. Bond y colaboradores (2012) utilizaron este empujón para movilizar al votante estadounidense a través de Facebook, al mostrar a sus usuarios cuáles de sus conocidos más cercanos ya habían ido a votar.

- Enmarcado (*framing*): Formular las opciones de decisión de forma que se resalte una característica concreta de estas. Por ejemplo, presentando a pacientes que se van a someter a una operación las probabilidades de supervivencia a cinco años en términos de ganancia (el 90% siguen vivos) o en términos de pérdida (el 10% mueren; McNeil y cols., 1982).

Aunque el uso intencionado de heurísticos, sesgos y diversos mecanismos psicológicos para influir sobre la conducta no es algo nuevo (Gigerenzer, 2015b), la propuesta de Thaler y Sunstein (2009) ha adquirido tanta popularidad en los últimos años que muchas de las intervenciones conductuales y de las estrategias persuasivas utilizadas en la literatura psicológica han pasado a ser acuñadas como empujones. Además, estos empujones han sido aplicadas por gobiernos y organizaciones internacionales con el objetivo de diseñar políticas supuestamente más efectivas (Gigerenzer, 2015b, 2018), llegando Thaler y Sunstein a asesorar en asuntos

regulatorios al presidente de EE.UU., Barack Obama, o al Primer Ministro de Reino Unido, David Cameron (Hansen y Jespersen, 2013). Sin embargo, son muchas las voces críticas que han alertado de su potencial de manipulación, de limitación de autonomía (Hortal, 2019; Sætra, 2019; Sugden, 2017; Susser y cols., 2019) y de su cuestionable eficacia en todo tipo de contextos (Caraban y cols., 2019).

Esta idea del uso de heurísticos y sesgos para influir en el comportamiento se ha trasladado de forma muy natural al ámbito de la tecnología, un terreno en el que se considera que los empujones podrían alcanzar su máxima potencialidad al poder ser usados por los algoritmos para alterar el entorno de decisión de forma dinámica, escogiendo el algoritmo incluso el heurístico personalizado con el que influir a cada individuo a partir de los datos recopilados sobre él (Caraban y cols., 2019; Karlsen y Andersen, 2019; Suh, 2019; Susser y cols., 2019; Weinmann y cols., 2016).

Diversos estudios, con foco en el campo de la tecnología, han abordado cómo el uso de heurísticos podría implicar manipulación encubierta a través de heurísticos. Un ejemplo de esto es el mencionado experimento de Facebook sobre el comportamiento de los votantes durante las elecciones al Congreso de los EE.UU. en 2010 (Bond y cols., 2012). Con una muestra de casi 61 millones de sus usuarios, investigadores de la Universidad de California en colaboración con trabajadores de Facebook realizaron un experimento para comprobar si podían movilizar al voto a través de la plataforma. Para ello, los investigadores mostraron a los usuarios un mensaje en su perfil de Facebook para que indicaran si ya habían ido a votar y, según la condición experimental, los usuarios visualizaron o no las fotografías de seis amigos íntimos que ya habían votado (o al menos declaraban haberlo hecho). Al mostrar estas

fotografías, los investigadores explotaron el heurístico prueba social (Cialdini, 1993, Chapter 4) con el objetivo de empujar a los usuarios a imitar el comportamiento de sus amigos. En palabras de los autores, esta manipulación experimental logró la participación directa de unos 60.000 votantes e indirectamente de 280.000. Unas cifras que pueden condicionar el resultado de cualquier elección democrática. El estudio fue posteriormente replicado con similares resultados, utilizando como escenario las elecciones presidenciales de 2012 en EE.UU (Jones y cols., 2017). Un caso que muestra como los heurísticos y sesgos pueden ser utilizados para manipular el pensamiento y el comportamiento, a veces en interés de terceros.

Otro ejemplo de persuasión encubierta en tecnología mediante heurísticos y empujones es el abordado por Epstein y Robertson (2015) durante las elecciones Lok Sabha de 2014 en la India. Con el objetivo de comprobar si podían manipular el voto indeciso, los investigadores presentaron a los participantes información sobre los dos candidatos a las elecciones en una herramienta de búsqueda similar al buscador de Google. El orden de los contenidos fue manipulado para que las primeras páginas de resultados favorecieran a uno u otro de los candidatos en función del grupo al que fueron asignados. Según los autores, el *efecto de primacía* (*primacy effect*), por el cual los participantes consideraron más relevante la información de los primeros resultados de la búsqueda, inclinó el voto hacia uno y otro candidato en un 20% de los casos, con un éxito de hasta un 72.7% en algunos grupos demográficos.

En otro contexto distinto pero con proceder similar se encuentra un criticado experimento de “contagio emocional” de Facebook (Kramer y cols., 2014). En él se manipuló la visibilidad de casi 700.000 publicaciones de los usuarios de Facebook para

investigar si al reducir la exposición de los usuarios a contenido emocional negativo o positivo se podía condicionar el tono de sus publicaciones posteriores. Según los autores, los usuarios de la plataforma se contagiaron emocionalmente en la dirección de la manipulación sufrida, publicando con el mismo tono negativo o positivo al que habían sido expuestos.

Pero esta influencia no solo se ha probado a nivel experimental. Hace poco tiempo, la empresa Cambridge Analytica fue acusada por utilizar información de más 50 millones de usuarios de Facebook para perfilar e influir a los votantes indecisos en diversas campañas políticas (Cadwalladr y Graham-Harrison, 2018; Susser, 2019), incluidas la del presidente de los Estados Unidos Donald Trump en 2016, las elecciones presidenciales de Nigeria de 2015 o las controvertidas elecciones de 2017 en Kenia en las que Uhuru Kenyatta salió victorioso (Al Jazeera, 2018). Aunque no se han publicado estudios empíricos que demuestren la eficacia de esta influencia (Del Castillo, 2018), la mera posibilidad de que esta afectara a las elecciones presidenciales de EE.UU provocó un alto impacto mediático y reacciones a nivel institucional sin precedente, como la petición de que el director ejecutivo de Facebook, Mark Zuckerberg, testificara sobre lo sucedido en el Congreso Estadounidense (Watson, 2018).

Como hemos visto, los algoritmos se hallan involucrados en muchas de nuestras actividades y decisiones diarias (Willson, 2017) a pesar de no ser una tecnología neutral, ni de estar libre de sesgos o de intencionalidad en sus recomendaciones. Su naturaleza inescrutable y su capacidad para personalizar su respuesta a cada individuo mediante el procesamiento de grandes volúmenes de datos (Alvarado y cols., 2019; Oduor y cols., 2014) convierten a los algoritmos de IA en

tecnología con alta capacidad persuasiva. Por este motivo, consideramos importante abordar cómo responderán las personas a ella en situaciones de decisiones cotidianas pero relevantes.

Capítulo 3. **Serie Experimental 1**

Abordamos una primera serie de seis experimentos para comprobar si la recomendación algorítmica podía influir en las preferencias y el comportamiento de las personas en contextos de decisión relevantes, como el voto político o las citas románticas. Se trata de contextos de decisión donde los algoritmos comienzan a ofrecer recomendaciones y donde su consejo puede impactar profundamente en la vida de las personas (los resultados de unas elecciones democráticas o la elección de una pareja).

Aun sabiendo que hay numerosas variables que pueden influir en las decisiones y que serían susceptibles de manipularse, para los experimentos de esta serie determinamos que los participantes tomaran sus decisiones basándose únicamente en las fotografías de los candidatos políticos o de citas mostrados, en aras de la simplicidad y porque el aspecto físico es una de las características que más afecta a los votantes en política (Palmer y Peterson, 2015; White y cols., 2013) y a la interacción en las plataformas de citas (Hern, 2014).

Experimento 1. Recomendación explícita en contexto político

El objetivo de este experimento fue evaluar si la recomendación explícita de candidatos políticos por parte de un supuesto algoritmo de IA podía influir en las preferencias de voto de los participantes. Como se ha mencionado en la sección de Aspectos Éticos y Ciencia Abierta, los contextos de decisión y los candidatos fueron ficticios. Además, el algoritmo de IA en realidad no existía. Nuestra hipótesis era que

los participantes valorarían mejor a los candidatos políticos recomendados por el algoritmo que a los candidatos control.

Método

Participantes y Materiales

Reclutamos a 441 participantes (46.7% mujeres, 2.5% no reportado; edades 18-71, M = 39.3, SD = 10.7) a través de Twitter mediante el procedimiento de bola de nieve. Para ello, publicamos en esta red social una invitación para participar en un experimento sobre el papel de los procesos psicológicos en las decisiones políticas, solicitando a los usuarios de la plataforma que difundieran la invitación. El mensaje estaba redactado en español y contenía un enlace al sitio web donde realizamos el experimento, que estaba redactado también en español.

Dado que no teníamos conocimiento de experimentos previos similares a este, no pudimos realizar un análisis de potencia a priori para determinar el tamaño de la muestra necesaria para el experimento, pero sí realizamos a posteriori un análisis de sensibilidad que indicó que, con el tamaño de muestra obtenido y una potencia del 90%, nuestro experimento era capaz de detectar efectos de tamaño pequeño (η^2_p = 0.009).

De forma automatizada, se asignó aleatoriamente a todos los participantes a uno de los dos grupos del experimento: recomendación explícita (n = 219) y sin-recomendación (n = 222). Utilizamos fotografías de la base de datos normalizada de Bainbridge y colaboradores (2013) para las imágenes de los candidatos políticos. Estas

fotografías fueron calibradas en atractivo en un experimento previo (véase Apéndice A).

Diseño y Procedimiento

El diseño de este experimento, que constaba de tres fases para los dos grupos, se resume en la Tabla 1, que además permitirá más adelante revisar también el diseño de los siguientes experimentos de esta serie.

Tabla 1

Resumen del Diseño de los Experimentos 1, 2, 3, 4 y 5

Experimento	Contexto	Grupo	Fase 1	Fase 2
Experimento 1	Política	Explícito	E1-E32	D1-D4 (*) C1-C4
		Sin-recomendación	E1-E32	D1-D4 C1-C4
Experimento 2	Política	Encubierto	E1-E16 D1-D4 (x4)	D1-D4 C1-C4
Experimento 3	Citas	Explícito	E1-E40	D1-D4 (*) C1-C4
		Encubierto	E1-E20 D1-D4 (x5)	D1-D4 C1-C4
		Sin-recomendación	E1-E40	D1-D4 C1-C4
Experimento 4	Citas	Explícito	E1-E40	D1-D4 (*) C1-C4
		Encubierto	E1-E20 D1-D4 (x5)	D1-D4 C1-C4
		Sin-recomendación	E1-E40	D1-D4 C1-C4
Experimento 5	Política	Explícito	E1-E40	D1-D4 (*) C1-C4
		Encubierto	E1-E20 D1-D4 (x5)	D1-D4 C1-C4
	Citas	Explícito	E1-E40	D1-D4 (*) C1-C4
		Encubierto	E1-E20 D1-D4 (x5)	D1-D4 C1-C4

Nota. Las imágenes pueden ser D = Diana, recomendado; C = Control, no recomendado; o E = Extras, de relleno; (x4) (x5) = pre-expuesto 4 o 5 veces, respectivamente; (*) = distintivo "+90% compatibilidad". El papel de los estímulos como candidatos diana (D1-D4) y control (C1-C4) fue contrabalanceado. No se señala en la tabla la Fase 0 de confiabilidad por ser idéntica en todos los casos. Los Experimentos 2, 3, 4 y 5 se describirán en detalle más adelante.

La primera parte del experimento (Fase 0) consistía en lo que puede describirse como una fase de confiabilidad, idéntica para todos los participantes de los cinco experimentos. El propósito era generar confianza hacia el falso algoritmo, de forma que los participantes entendieran como personalizadas las recomendaciones aleatorias que recibirían durante el experimento. Para ello, tras aceptar el consentimiento informado online e indicar su género, edad y orientación política en una escala del 1 (*extrema izquierda*) al 9 (*extrema derecha*), se solicitó a los participantes que respondieran a algunas preguntas "pertenecientes a los estudios más importantes en este campo" para que nuestro supuesto algoritmo pudiera inferir su personalidad. Utilizamos aquí una versión del efecto Forer (1949), similar a la implementada por Barberia y colaboradores (2018). En nuestro caso, mostrábamos a los participantes unas preguntas de personalidad ficticias, extraídas de un test de compatibilidad ideológica (Soulié, 2015; por ejemplo, “Las ayudas sociales deberían reducirse porque hacen que la gente deje de buscar trabajo”), mezcladas con ítems de una antigua página web[1] que ya no está activa y que afirmaba inferir la personalidad y gustos de sus usuarios en base a extrañas preguntas del tipo “¿Qué tipo de patatas fritas disfrutas más?” o “¿Te gustan los autos de choque?” (véase Figura 2 y Figura 3).

[1] Hunch.com, con 1.2 millones de visitantes únicos en 2010. "Hunch (Website)" (2021)

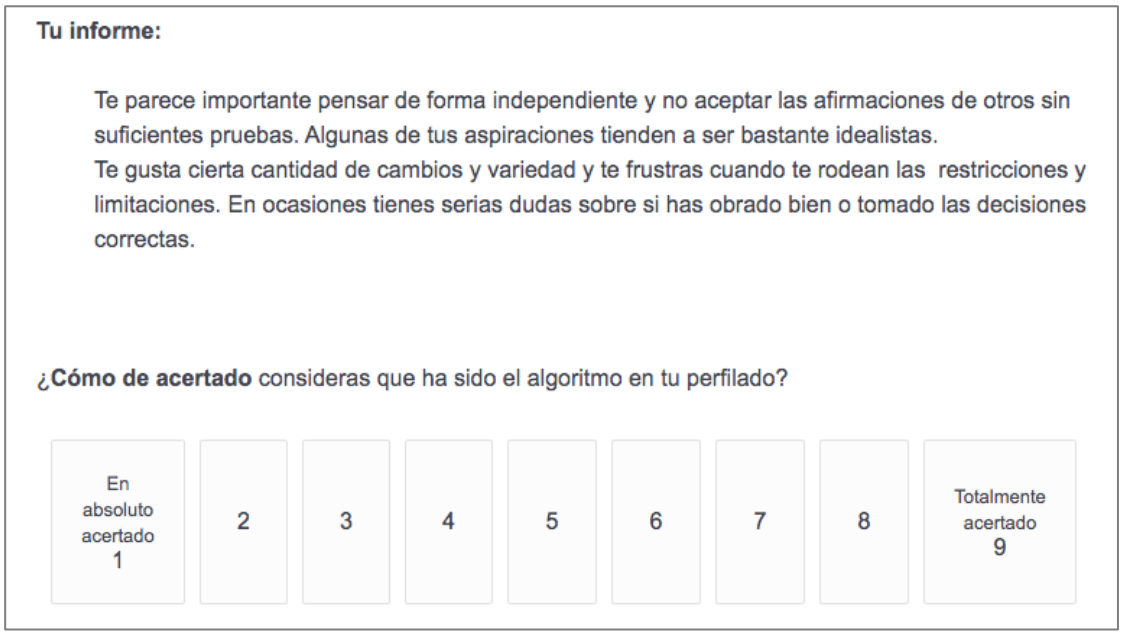
Figura 2

Ejemplo de Pregunta del Test Ficticio de Personalidad

Tras las preguntas del ficticio test de personalidad, los participantes recibieron un informe supuestamente personalizado que, sin que lo supieran, en realidad era idéntico para todos ellos. Al igual que en el experimento de Forer (1949), el informe estaba redactado vagamente para hacer creer a los participantes que el algoritmo había adivinado su personalidad. La mayoría de los participantes calificaron el informe falso del algoritmo de moderada a altamente acertado (M = 6.71, SD = 1.65, en una escala de 1 a 9).

Figura 3

Informe Basado en el Efecto Forer, Supuestamente Personalizado por Participante



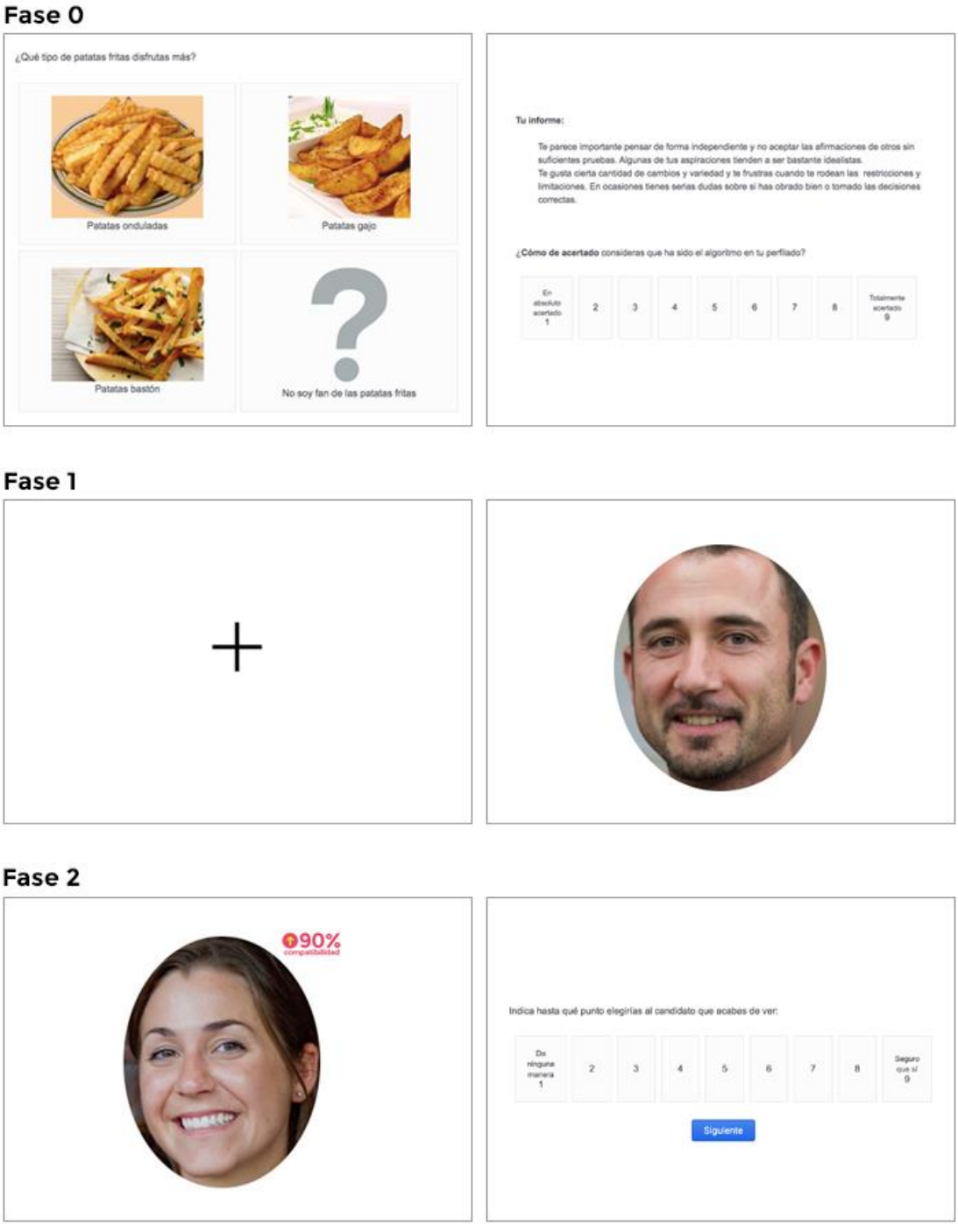
Durante la Fase 1 del experimento, todos los participantes observaron 32 fotografías de representantes políticos ficticios de otro país (50% mujeres) que habían sido previamente calibradas en atractivo en un experimento previo (véase Apéndice A). Las fotografías fueron mostradas durante 1 segundo cada una y la tarea solicitada a los participantes era observar las imágenes porque se les harían algunas preguntas más adelante y porque supuestamente esta visualización era necesaria para que el sistema analizara sus preferencias y pudiera identificar los candidatos políticos más compatibles con ellos.

El verdadero objetivo de esta visualización de fotografías extras era reforzar la confiabilidad del algoritmo antes de su recomendación. El orden de presentación de

las fotografías fue aleatorio para cada participante tanto en esta fase como en la siguiente. La Figura 4 muestra ejemplos de las pantallas visualizadas en cada fase.

Figura 4

Ejemplos de Pantallas Visualizadas en Cada Fase



En la Fase 2 se explicó a todos los participantes que el algoritmo mostraría ocho nuevos candidatos políticos de otro país en función de su perfil y que tendrían que indicar, tras un rápido vistazo, hasta qué punto les votarían, utilizando una escala del 1 al 9. Las ocho fotografías de los nuevos políticos ficticios (de nuevo 50% mujeres) se presentaron, igual que en la fase anterior, de una en una y en formato de pantalla completa. En el grupo explícito, cuatro de los candidatos (los estímulos *diana*) se mostraron como los más compatibles con la personalidad y las preferencias del participante, señalándose como candidatos recomendados mediante un distintivo con el texto "+90% de compatibilidad". Las otras cuatro fotografías fueron candidatos control y no mostraron el distintivo (véase Figura 5). Dado que ahora se solicitaba a los participantes que no solo observaran cada fotografía, sino que también las puntuaran tras su visualización, el tiempo de exposición aumentó a 2 segundos por imagen durante esta fase. Se contrabalancearon las imágenes en su papel como candidatos diana o candidatos control (véase Tabla 2).

Tabla 2

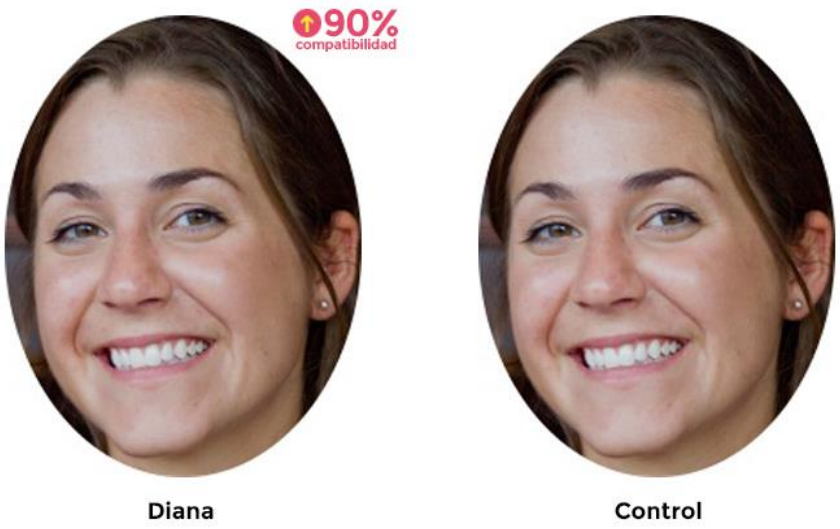
Contrabalanceos de los Estímulos Diana y Control en los Experimentos 1, 2, 3 y 4

	D1	D2	D3	D4	C1	C2	C3	C4
Balanceo 1	A	B	C	D	E	F	G	H
Balanceo 2	E	F	G	H	A	B	C	D

Nota. D = Estímulo Diana, recomendado; C = Estímulo Control, no recomendado; A-H = Fotografías de los candidatos.

Figura 5

Ejemplo de Presentación de Candidatos



Nota. El papel de los estímulos como candidatos diana (con distintivo "+90% compatibilidad" en el grupo explícito) y candidatos control (sin distintivo) estaba contrabalanceado. Fotografía con permisos para mostrar públicamente de la base de datos pública de Bainbridge y colaboradores (2013).

Las puntuaciones medias de los cuatro candidatos diana y de los cuatro candidatos control fueron nuestras variables dependientes. En el grupo sin-recomendación, los participantes no recibieron ninguna sugerencia del algoritmo, es decir, ni los candidatos diana ni los candidatos control mostraron distintivo, por lo que no había diferencias entre candidatos diana y control en este grupo. Así, esperábamos que los candidatos diana atrajeran más votos que los candidatos control en el grupo explícito y que los participantes no mostraran ninguna preferencia entre candidatos en el grupo sin-recomendación.

El experimento concluyó con una pregunta para comprobar si los participantes del grupo explícito recordaban haber visto el distintivo de compatibilidad ("¿Percibiste el distintivo de compatibilidad en cuatro de los candidatos mostrados?"), una pregunta, a aquellos que afirmaban haber visto el distintivo, sobre si consideraban que les había podido influir en sus valoraciones ("¿Hasta qué punto consideras que el porcentaje de compatibilidad en algunos candidatos ha podido influir en tus puntuaciones, mejorando tu opinión sobre ellos? ") y una breve explicación sobre el propósito del experimento tras aceptar el envío de datos.

Resultados y Discusión

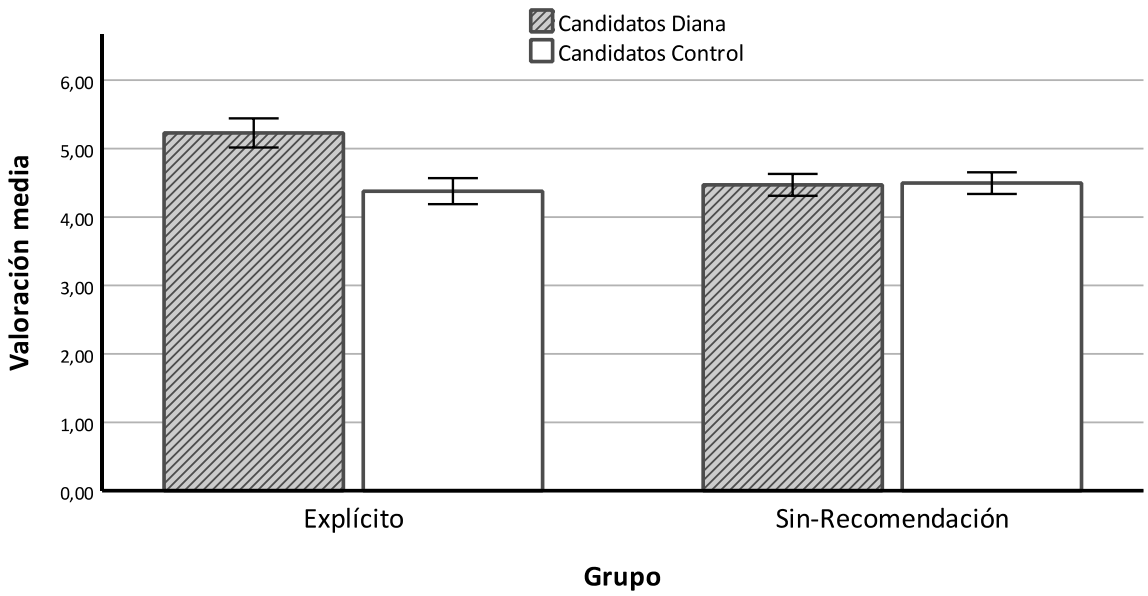
La Figura 6 resume los resultados de este experimento. Realizamos un ANOVA mixto[2] 2 (candidato: diana *vs*. control) x 2 (grupo: explícito *vs*. sin-recomendación), que mostró un efecto principal para el grupo, $F(1, 439) = 8.15$, $p = .005$, $\eta^2_p = 0.018$, efecto principal para el candidato, $F(1, 439) = 37.6$, $p < .001$, $\eta^2_p = 0.079$, así como interacción Candidato x Grupo, $F(1, 439) = 42.5$, $p < .001$, $\eta^2_p = 0.088$. Las comparaciones a posteriori mediante la prueba de Tukey indicaron que, tal y como habíamos predicho, los participantes en el grupo explícito tenían una mayor disposición a votar por los candidatos diana respecto a los candidatos control, $t(439) = 8.913$, $p < .001$, $d = 0.56$.

[2] Dado el tamaño de la muestra de este experimento y que el ANOVA se considera un test robusto a las desviaciones de normalidad (Field, 2013), no contemplamos utilizar análisis no-paramétricos. No obstante, reportamos adicionalmente los resultados del test Wilcoxon para las medias de candidatos diana y control por grupo, puesto que las puntuaciones a los candidatos fueron recogidas con una escala Likert. Explícito: $Z = 6.587$, $p = .000$. Sin-Recomendación: $Z = -0.404$, $p = .686$.

Como también esperábamos, no existían diferencias entre candidatos diana y candidatos control dentro del grupo sin-recomendación, $t(439) = -0.273$, $p = .993$, $d = -0.02$. La edad y el género no afectaron a los resultados de este experimento, ni a ninguno de los siguientes (todos $p > .05$), por lo que fueron colapsados en todos ellos.

Figura 6

Promedio de la Disposición a Votar a los Candidatos Diana y Control por Grupo en el Experimento 1



Nota. Las barras de error representan un 95% de CI.

Los resultados de este primer experimento parecen indicar que, al menos en ausencia de información adicional sobre los candidatos políticos, un algoritmo es capaz de influir en las preferencias de voto de los participantes a través de la recomendación explícita. Tal y como esperábamos, los participantes del grupo sin-recomendación no

mostraron ninguna preferencia por los candidatos diana, mientras que estos sí resultaron mejor valorados en el grupo explícito, donde habían sido recomendados por el algoritmo mediante el distintivo “+90% compatibilidad”.

Consideramos que estos resultados contribuyen a incrementar las escasas investigaciones publicadas sobre el potencial de la IA para influir en las decisiones de las personas, en contraste con la enorme cantidad de investigaciones privadas y no publicadas que realizan a diario las empresas de IA (Villareal, 2019). Nuestros resultados aportan evidencia de que la aceptación de la recomendación explícita del algoritmo puede llegar a influir a votantes indecisos y, con ello, afectar a los resultados de unas elecciones democráticas.

A partir de estos resultados, nos preguntamos si, en este mismo contexto de decisión, los algoritmos serían capaces de ejercer su influencia de otro modo. ¿Se aceptaría la recomendación del algoritmo si esta no se presentara de forma explícita sino de forma encubierta? Con esta pregunta en mente, diseñamos el experimento que sigue a continuación.

Experimento 2. Recomendación encubierta en contexto político

Dado que en el experimento anterior la recomendación explícita de un supuesto algoritmo había influido en las preferencias de voto de los participantes, nuestro objetivo con este nuevo experimento era comprobar si también podría lograrse esta persuasión de forma encubierta utilizando un heurístico para empujar la decisión. Así, en este Experimento 2, utilizamos el heurístico de familiaridad. Este atajo mental, muy relacionado con las preferencias de gusto y elección (Abakoumkin, 2011), favorece que las personas prefieran los estímulos familiares frente a los estímulos novedosos (Bailenson y cols., 2008; Cameron, 2009; Montoya y cols., 2017; Peskin y Newell, 2004; Zajonc, 1968). En nuestro experimento, para generar familiaridad utilizamos el efecto conocido como *efecto de mera exposición* (Zajonc, 1968). Diversas investigaciones han mostrado que la pre-exposición repetida de ciertos estímulos (Bornstein, 1989), además de aumentar la familiaridad y la preferencia por estos (Abakoumkin, 2011; Bornstein, 1989; Montoya y cols., 2017; Peskin y Newell, 2004), incrementa una respuesta positiva hacia ellos posteriormente (Zajonc, 1968), sin que el reconocimiento sea necesariamente un requisito previo (Montoya y cols., 2017). A pesar del amplio número de investigaciones existentes acerca del efecto de mera exposición y sobre la familiaridad, su papel en el contexto político no ha sido, que nos conste, muy abordado. La escasa literatura existente sobre el tema apunta, sin embargo, a resultados en línea con nuestra hipótesis. B. C. Burden (2002), por ejemplo, encontró que los candidatos políticos que resultaban más familiares a los electores contaban con mayor probabilidad de ser elegidos, incluso cuando su familiaridad se

había conseguido a base de publicidad negativa hacia ellos. Por todo esto, nuestra hipótesis era que algunos candidatos políticos resultarían más familiares a los participantes si eran pre-expuestos repetidas veces, y que esta familiaridad, a su vez, empujaría las preferencias de voto hacia dichos candidatos de forma encubierta.

Método

Participantes y Materiales

Reclutamos a 218 participantes (48.2% mujeres, 0.7% no reportado; edades 30-49, *M* = 35.80, *SD* = 4.46) a través de Twitter, como en el Experimento 1, utilizando de nuevo el procedimiento de bola de nieve, al solicitar a los usuarios de la plataforma que difundieran la invitación a participar en el experimento. De nuevo, el mensaje públicamente distribuido se encontraba redactado en español y contenía un enlace al sitio web del experimento en el mismo idioma. En este caso, el análisis de sensibilidad indicó que, con este tamaño de muestra, nuestro experimento contaba con una potencia del 90% para detectar un efecto de tamaño pequeño (*d* = 0.20). El conjunto de imágenes utilizadas pertenecía, al igual que en el Experimento 1, a Bainbridge y colaboradores (2013) y habían sido calibradas en atractivo en un experimento previo (véase Apéndice A).

Diseño y Procedimiento

El Experimento 1 ya había demostrado que no existían diferencias en la preferencia de voto entre candidatos diana y control cuando los participantes no eran expuestos a la recomendación explícita del falso algoritmo, es decir, cuando no llevaban el distintivo de compatibilidad en el grupo sin-recomendación (véase Tabla 1 y

Figura 5). Por ello, en este experimento decidimos utilizar un diseño intra-sujeto: un solo grupo en el que se comparaban las puntuaciones dadas a los candidatos pre-expuestos, frente a los candidatos no pre-expuestos. El diseño (véase Tabla 1) y las fases del procedimiento fueron muy similares a las del Experimento 1. Tras aceptar el consentimiento informado e indicar sus datos demográficos, los participantes completaron el test ficticio de personalidad y recibieron el informe de personalidad presumiblemente individualizado durante la Fase 0. A continuación, en la Fase 1, de nuevo observaron imágenes de candidatos políticos (ficticios) pero, esta vez, las cuatro de las imágenes de los candidatos diana fueron pre-expuestas cuatro veces (16 ensayos) para producir familiaridad por el efecto de mera exposición. Para determinar el número de pre-exposiciones que eran necesarias, seguimos el consejo de Bornstein (1989), quien indica que se logra un mayor tamaño del efecto cuando se usa un número relativamente pequeño de repeticiones (entre una y nueve). Las cuatro imágenes utilizadas como candidatos diana (pre-expuestas) y las cuatro que fueron usadas como candidatos control en la Fase 2 (véase Tabla 1) fueron contrabalanceadas como en el Experimento 1 (véase Tabla 2).

Durante la Fase 1, cada participante también observó otras 16 imágenes como estímulos extras, o de relleno, para completar los 32 ensayos utilizados durante la Fase 1 del Experimento 1. Las 32 imágenes de esta fase se presentaron en orden aleatorio, aunque con la condición de que no se produjeran más de dos repeticiones seguidas del mismo estímulo. Esta disposición de imágenes estaba motivada por la preocupación de que una repetición demasiado evidente de los ítems pudiera provocar rechazo o aburrimiento en los participantes, aunque no es necesario que la repetición pase

inadvertida para que se produzca el efecto de la mera exposición (Bornstein, 1989; Montoya y cols., 2017). Como en el Experimento 1, los candidatos (diana y extras) se pre-expusieron durante 1 segundo cada uno. La literatura anterior sobre el efecto de mera exposición resuelve que los tiempos de pre-exposición entre 1 y 4 segundos son los que provocan un mayor efecto en las preferencias posteriores (Montoya y cols., 2017).

Durante la Fase 2, los participantes observaron las cuatro imágenes de candidatos diana además de cuatro nuevas imágenes de candidatos control durante 2 segundos cada una antes de indicar su disposición a votarles. Aunque no hay ningún estándar en la literatura de este efecto respecto a la duración del tiempo de evaluación, es habitual que se limite con exposiciones que varían entre los 2 segundos (Pheterson y Horai, 1976; Zajonc, 1968), 1 segundo (Raft y Zajonc, 1980), u 8 segundos (Peskin y Newell, 2004). Los ocho candidatos utilizados en esta fase de puntuación fueron contrabalanceados en su papel como candidatos diana o control (véase Tabla 2) y se presentaron en orden aleatorio a cada participante, quien indicaba su disposición a votar por cada uno de ellos en una escala de 1 a 9. Las medias de estas calificaciones para los candidatos diana y para los candidatos control fueron nuestras variables dependientes. Tras la valoración de los candidatos, los participantes indicaron si habían detectado la repetición de los candidatos y si, en caso de haberla detectado, consideraban que la repetición había influido en sus puntuaciones. Por último, los participantes recibieron una explicación sobre el verdadero propósito del experimento.

Resultados y Discusión

Una prueba T-Student de muestras relacionadas reveló que, en contra de nuestras expectativas, no había una mayor disposición a votar a los candidatos diana (M = 4.45, SD = 1.45) frente a los candidatos control (M = 4.34, SD = 1.28; $t(217) = 1.58$, $p = .058$, $d = 0.11$). Por tanto, según nuestros resultados, el falso algoritmo no fue capaz de influir de forma encubierta en las preferencias de voto de los participantes en este experimento.

Dado que no encontramos influencia en la preferencia de voto es preciso evaluar las posibles causas de esta ausencia de efecto. Una posibilidad es que los participantes contaran en esta ocasión con fuertes preferencias individuales hacia algunos candidatos, siendo nuestra manipulación encubierta incapaz de cambiarlas. Sin embargo, dado que los estímulos diana y control estaban contrabalanceados, esta explicación parece implausible. Otra posible explicación, quizá más plausible, es que no conseguimos captar la atención de los participantes durante la fase de pre-exposición, lo que habría evitado que se generase familiaridad hacia los candidatos diana. Por último, estos resultados pudieron también deberse a la posible falibilidad de los empujones (Sunstein, 2017). Caraban y colaboradores (2019), en una revisión sistemática sobre los empujones en entornos tecnológicos, encontraron que solo el 66% de los empujones estudiados habían afectado de forma significativa el comportamiento o las actitudes de los participantes. De hecho, los autores indicaron que un empujón ineficaz podría provocar el efecto contrario al buscado, desencadenando en las personas comportamientos compensatorios. Dado que en este experimento tuvimos que elegir los parámetros de la manipulación mediante ensayo y

error (el número de repeticiones de los estímulos diana y el tiempo de pre-exposición), quizá era necesario probar un conjunto diferente de parámetros en siguientes experimentos antes de sacar conclusiones de nuestros resultados. Por ejemplo, en su meta-análisis del efecto de mera exposición, Bornstein (1989) recomienda, como mencionamos, un número de repeticiones de entre uno y nueve para lograr un efecto de moderado a fuerte. Basándonos en esta propuesta y en la advertencia de Montoya y colaboradores (2017) sobre el efecto negativo de la duración del estudio en los resultados, decidimos usar solo cuatro repeticiones. Sin embargo, y dado que nuestra elección de cuatro pre-exposiciones no produjo el efecto deseado, decidimos introducir cambios en el procedimiento del siguiente experimento, además de aprovechar el nuevo experimento para tratar de replicar la influencia de la recomendación explícita observada en el Experimento 1 en un contexto de decisión muy diferente: la búsqueda de pareja.

Experimento 3. Recomendación explícita y encubierta en contexto de citas

Con el propósito de poner a prueba la potencial generalización de los resultados observados en el Experimento 1, en este nuevo experimento cambiamos el contexto político de los dos experimentos anteriores por un nuevo contexto: el de las citas online. El objetivo de este siguiente experimento era replicar la influencia explícita del Experimento 1 en un nuevo contexto, además de poner a prueba si ciertas modificaciones en el procedimiento pudieran mejorar la influencia de la recomendación algorítmica encubierta que no resultó efectiva en el Experimento 2.

Nuestra principal hipótesis era que el algoritmo influiría en las preferencias de los participantes a través de su recomendación explícita, igual que había ocurrido en el Experimento 1 pero ahora en el contexto de una aplicación de citas. Asimismo, esperábamos también que los cambios introducidos en el procedimiento del Experimento 2 sirvieran para poder verificar la influencia de la recomendación encubierta en este nuevo contexto.

Este entorno de decisión resulta muy interesante para la investigación empírica dado que la mayoría de las páginas web de citas argumentan que sus sofisticados algoritmos de compatibilidad logran emparejamientos más compatibles que las citas tradicionales (Finkel y cols., 2012), y que sus resultados son mejores a corto y largo plazo al recomendar candidatos compatibles desde el inicio. Aunque no hay suficientes investigaciones que apoyen esta afirmación y estas aplicaciones de citas tampoco han publicado pruebas rigurosas sobre su supuesta superioridad respecto a la búsqueda de

citas tradicional, el número de usuarios en estas plataformas ha aumentado en los últimos años (Duportail, 2019; Finkel y cols., 2012).

Método

Participantes y Materiales

Reclutamos 280 participantes (48.2% mujeres, 0.7% no reportado; edades 30-49, *M* = 35.80, *SD* = 4.46) a través de la plataforma en línea Prolific Academic. Utilizamos los filtros de selección de muestra que proporciona la plataforma, de modo que la invitación para participar en el experimento se dirigió a usuarios de edades comprendidas entre los 30 y los 45 años y de etnia blanca/caucásica, para que coincidieran con la edad y la etnia de los candidatos de la base de datos fotográfica utilizada y evitar así un posible efecto de la edad o la etnia en los resultados (véase Apéndice A). La invitación en esta ocasión estaba redactada en inglés y el experimento se llevó a cabo también en inglés. El programa asignó al azar a los participantes a uno de los tres grupos del experimento: recomendación explícita (*n* = 94), recomendación encubierta (*n* = 90), y sin-recomendación, como grupo control (*n* = 96). El análisis de sensibilidad indicó que, con este tamaño de muestra, contábamos con una potencia del 90% para detectar un efecto de tamaño pequeño (η^2_p = 0.021).

Diseño y Procedimiento

En este Experimento 3, tras el consentimiento informado online y responder sobre su género y edad, los participantes indicaron su estado sentimental (con pareja o sin ella) y si preferían indicar su preferencia hacia los candidatos hombres o mujeres, por si su experiencia previa pudiera afectar a sus respuestas. A continuación, como en

los experimentos anteriores, los participantes rellenaron el ficticio test de personalidad durante la Fase 0. En esta ocasión, las preguntas relacionadas con la política utilizadas previamente fueron sustituidas por cuatro ítems del test de personalidad romántica de Fisher (2018) (por ejemplo, “Encuentro estimulantes las situaciones impredecibles” / “I find unpredictable situations exhilarating”) y dos preguntas de la página web de citas eDarling (por ejemplo, “Tu plan perfecto para una primera cita: Música en directo; Senderismo; Cine; Cenar fuera” / “Your perfect first date plan: Live music; Trekking; A movie; Dining out”).

Seguidamente, los participantes pasaron a la Fase 1. La Tabla 1 muestra un resumen del diseño experimental, comparándolo con el resto de experimentos de esta serie. Durante la Fase 1, los participantes fueron expuestos a 40 fotografías de candidatos ficticios de citas (mujeres u hombres, según la preferencia que los participantes habían indicado). Cada fotografía se presentó durante 1 segundo como en los experimentos previos.

En el grupo explícito y en el grupo sin-recomendación, las 40 fotografías fueron de relleno en esta fase. En el grupo encubierto solo 20 fotografías eran de relleno, además de cuatro fotografías diana que fueron pre-expuestas cinco veces cada una en los otros 20 ensayos a fin de generar familiaridad. Como en el Experimento 2 utilizamos cuatro repeticiones para inducir la familiaridad y no logramos influencia con la recomendación algorítmica, decidimos ampliar a cinco repeticiones por cada fotografía diana siguiendo la sugerencia de Rhodes y colaboradores (2001). Los autores, en su experimento sobre el efecto de mera exposición con rostros compuestos promediados, utilizaron también cuatro repeticiones como en nuestro

Experimento 2 sin encontrar tampoco efecto de la repetición en el atractivo. Los investigadores infirieron que ampliar el número de pre-exposiciones para ciertos estímulos complejos, como es el caso de los rostros, podría haber sido más efectivo. Recogiendo su hipótesis, decidimos comprobarlo. Como en anteriores experimentos, el orden de presentación de cada imagen fue aleatorio para cada participante, con la regla de evitar más de dos repeticiones seguidas de un mismo candidato.

Durante la Fase 2, al igual que en los experimentos anteriores, todos los participantes utilizaron una escala del 1 al 9 para valorar a los candidatos finales. En esta ocasión indicaban su disposición a enviar un mensaje a cuatro candidatos diana y cuatro candidatos control mediante la plataforma de citas. Los ocho candidatos utilizados en esta fase fueron los cuatro candidatos diana que habían sido mostrados al grupo encubierto durante la Fase 1, además de cuatro nuevos candidatos control. Las ocho fotografías fueron de nuevo contrabalanceadas para utilizarse como candidatos diana o candidatos control (véase Tabla 2). En el grupo explícito, las imágenes de los candidatos diana mostraron el distintivo de recomendación con el texto "+90% de compatibilidad" utilizado en el Experimento 1. No hubo manipulación en el grupo sin-recomendación en ninguna de las dos fases, por lo que no esperábamos ninguna diferencia entre las imágenes diana y las imágenes control en este grupo, que fueron contrabalanceadas. Al igual que en los experimentos anteriores, las ocho imágenes utilizadas en esta fase se presentaron en orden aleatorio para cada participante. Sin embargo, a diferencia de estos, en esta ocasión no se utilizaron restricciones de tiempo en la visualización de los candidatos durante la Fase 2 para emular un contexto de búsqueda da pareja realista. Además, la escala de valoración se encabezó con los

símbolos de una x y un corazón en ambos extremos, siguiendo el diseño de la famosa aplicación de citas Tinder.

Como en el Experimento 1 y 2, para terminar, los participantes del grupo explícito indicaron si habían visto el distintivo de compatibilidad en los candidatos diana y los del grupo encubierto si habían percibido la repetición de candidatos, y, además, si en ambos grupos consideraban que esto (el distinto o la repetición de candidatos) había influido en sus puntuaciones. Además, se facilitó a todos los participantes una breve explicación sobre el propósito del experimento además del acceso al pago por su colaboración.

Resultados y Discusión

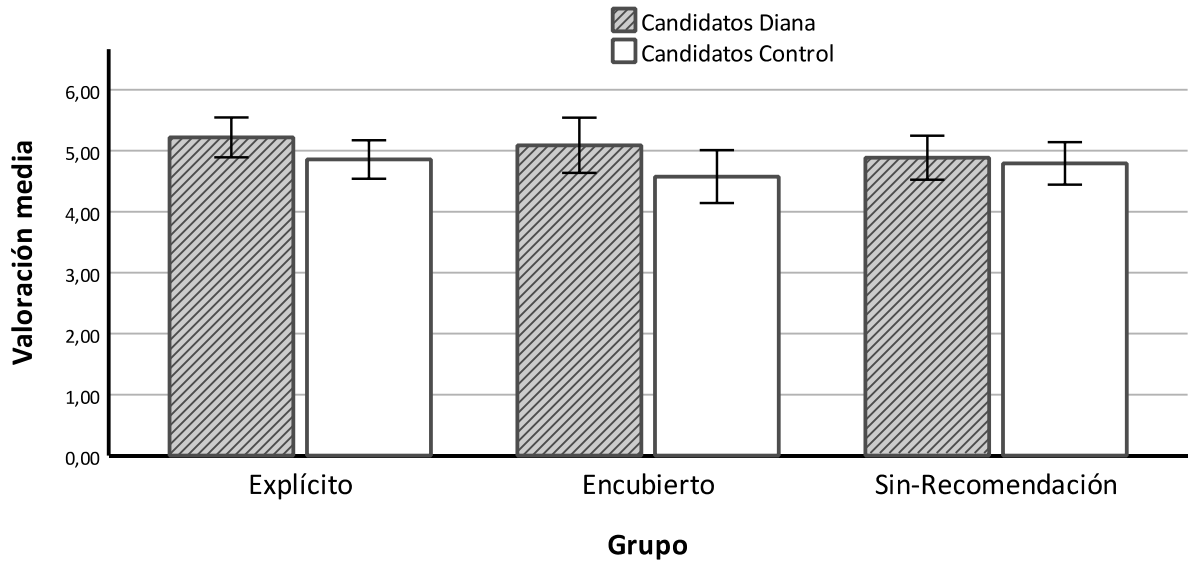
Los resultados se muestran en la Figura 7. Un ANOVA mixto[3] 2 (candidato: diana, control) x 3 (grupo: explícito, encubierto, sin-recomendación) mostró un efecto principal del candidato, $F(1, 277) = 17.03$, $p < .001$, $\eta^2_p = 0.058$, pero no un efecto principal del grupo, $F(2, 277) = 0.44$, $p = .644$, $\eta^2_p = 0.003$, ni interacción Candidato x Grupo, $F(2, 277) = 2.47$, $p = .087$, $\eta^2_p = 0.017$. A pesar de la falta de significación estadística en la interacción, consideramos necesario realizar las comparaciones a posteriori mediante la prueba de Tukey para asegurarnos de que ninguno de los grupos había mostrado diferencias en las valoraciones a los candidatos. Al contar con controles en todos los grupos (candidatos control), además de tener un grupo control (sin-recomendación), estimamos que quizá el posible efecto dentro de alguno de los

[3] Al igual que en el Experimento 1, dado que las puntuaciones a los candidatos fueron recogidas con una escala Likert, reportamos adicionalmente los resultados del test Wilcoxon para las medias de candidatos diana y control por grupo. Explícito: $Z = 2.554$, $p = .011$. Encubierto: $Z = 3.529$, $p = .000$. Sin-Recomendación: $Z = 0.552$, $p = .581$.

grupos podría estar quedando oculto. Tal y como esperábamos, encontramos que la preferencia de los participantes hacia los candidatos diana fue significativamente mayor que hacia los candidatos control en el grupo encubierto, $t(277) = 3.72$, $p = .003$, $d = 0.24$. Sin embargo, los participantes no mostraron una disposición significativamente mayor a enviar un mensaje de citas a los candidatos diana frente a los candidatos control en el grupo explícito, $t(277) = 2.68$, $p = .083$, $d = 0.24$. Como esperábamos, no hubo diferencia entre las puntuaciones de candidatos diana y control en el grupo sin-recomendación, $t(277) = 0.70$, $p = .982$, $d = 0.06$.

Figura 7

Valoración Media de los Candidatos Diana y Control, en los Grupos Explícito, Encubierto y Sin-Recomendación en el Experimento 3



Nota. Las barras de error representan un 95% de CI.

El efecto principal encontrado para el candidato, aunque lo reportamos por transparencia, no resulta relevante para el propósito de este experimento. Interpretamos que probablemente se trata de un artefacto producido por las diferencias de valoración entre candidatos diana y control en todos los grupos (en algunos casos menores) y una interacción atenuada en la que un grupo mostró el efecto de candidato esperado y los otros no.

Los resultados de este experimento, en contraste con los del Experimento 2, parecen indicar que resulta relativamente fácil influir de forma encubierta en las preferencias de las personas, al menos en el contexto de las citas online. Para ello, solo parece ser necesario aumentar la familiaridad de los candidatos. Sin embargo, estos resultados se contradicen con los encontrados en los Experimentos 1 y 2. Por una parte, en el Experimento 1 el algoritmo explícito influyó en la disposición de las personas a votar por los candidatos políticos recomendados, pero este resultado de la recomendación explícita no se pudo replicar en el presente experimento con un contexto diferente, el de las citas. Por otro lado, el algoritmo encubierto no fue eficaz en el Experimento 2 en política, pero sí en el experimento actual con el contexto de búsqueda de pareja.

Quizá las modificaciones realizadas en el procedimiento del Experimento 3 respecto a los experimentos anteriores pudieran haber sido responsables, en conjunto o por separado, de las diferencias en los resultados observados. Modificaciones que fueron variadas, como el aumento en el número de pre-exposiciones, el idioma del experimento (inglés vs. español), el canal de reclutamiento de la muestra (Prolific Academic vs. Twitter), el número de repeticiones para lograr el efecto de mera

exposición (cinco vs. cuatro), las restricciones de tiempo en la Fase 2 (sin límite vs. 2 segundos), los símbolos en la escala (presentes vs. ausentes) o el diferente contexto en el que se enmarcaron los experimentos (citas vs. política). Al no contar con referencias de procedimiento similares en la literatura, muchas de estas modificaciones respondían a la necesidad de ajustar el nuevo procedimiento al propósito del experimento mediante ensayo y error. Esto nos llevaría a considerar la posibilidad de que el algoritmo encubierto pudiera ser eficaz también en el contexto político con los nuevos parámetros utilizados en el presente experimento.

Además de todos los cambios realizados, existía un elemento en nuestros experimentos que también pudiera estar influyendo negativamente los resultados pese a mantenerse uniforme en todos ellos. Se trataba del banco de imágenes utilizado. Tal y como puede observarse en la Figura 5, la base de datos fotográfica de Bainbridge y colaboradores (2013) mostraba las fotografías de los candidatos con un formato que dista mucho del usado en contextos de búsqueda de pareja reales. Por ello, decidimos realizar un nuevo experimento para comprobar si se replicaban los resultados encontrados en el Experimento 3 al utilizar, como última modificación de nuestro procedimiento, un banco de imágenes más actual.

Experimento 4. Recomendación explícita y encubierta en contexto de citas. Réplica

Como hemos mencionado antes, este nuevo experimento perseguía dos propósitos. Por un lado, aportar solidez a los resultados del Experimento 3 al replicar el experimento y asentar el procedimiento con los cambios realizados: inglés como idioma del experimento, reclutamiento a través de Prolific Academic, cinco repeticiones para lograr el efecto de mera exposición, sin límite de tiempo en las puntuaciones de la Fase 2, símbolos presentes en la escala y contexto de citas. Por otro lado, sustituir la base de datos fotográfica de anteriores experimentos por una más actual.

Método

Participantes y Materiales

Reclutamos 180 participantes (56.7% mujeres, con edades entre 29 y 45 años (M = 36.1, SD = 4.52) y de etnia blanca/caucásica, mediante la plataforma online de Prolific Academic, para que coincidieran con la edad y etnia de los candidatos de la nueva base de datos del experimento, de Karras y colaboradores (2018). Los participantes fueron repartidos en tres grupos, como en el experimento anterior: grupo con recomendación explícita, n = 59; grupo con recomendación encubierta, n = 60; y grupo sin-recomendación, n = 61. El análisis de sensibilidad indicó que contábamos con una potencia del 90% para detectar un efecto de tamaño pequeño (η^2_p = 0.033) para este tamaño de muestra.

Antes de utilizar la nueva base de datos fotográfica, de Karras y colaboradores (2018), calibramos las imágenes necesarias en atractivo en un nuevo experimento previo que puede consultarse en detalle en el Apéndice B (véase también la comparación de imágenes respecto a la base de datos de los experimentos previos en la Figura 8).

Figura 8

Comparación de Imágenes de las Dos Bases de Datos Fotográficas Utilizadas

Nota. A = Ejemplo de fotografía de la base de datos de Bainbridge y colaboradores (2013), utilizada en los Experimentos 1, 2 y 3. Imagen con permisos para mostrarse. B = Ejemplo de fotografía de la base de Karras y colaboradores (2018), utilizada en los Experimentos 4, 5, 6 y 7. Imagen con licencia CC BY 2.0, nombre Untitled, autor McKinnon de Kuyper y disponible en https://www.flickr.com/photos/98491013@N02/12648556524/

Diseño y Procedimiento

Como en el resto de experimentos, la Tabla 1 muestra un resumen del diseño experimental. El procedimiento fue idéntico al del Experimento 3, con la única diferencia del cambio de base de datos fotográfica.

Tras el consentimiento informado, las preguntas sobre demografía, estado sentimental y preferencia por hombres o mujeres, los participantes pasaron a rellenar el test ficticio de personalidad de la Fase 0 y a leer y valorar el informe de personalidad. A continuación, durante la Fase 1, los participantes contemplaron 40 rostros de hombres o mujeres según su preferencia indicada. En el grupo encubierto, cuatro de los candidatos fueron expuestos repetidamente en cinco ocasiones; los otros 20 rostros de relleno hasta completar el total de 40 ensayos fueron fotografías extras. Por su parte, el grupo explícito y el grupo sin-recomendación visualizaron los 20 rostros de relleno del grupo encubierto además de otros 20 rostros extras más. En todos los grupos, y como en anteriores experimentos, el orden de las imágenes fue aleatorio para cada participante, evitando que la fotografía del mismo candidato se repitiera más de dos veces seguidas.

A continuación, durante la Fase 2, los participantes indicaron cómo de dispuestos estarían a enviar un mensaje a ocho candidatos por la página de citas, cuatro de ellos candidatos diana y cuatro candidatos control, presentados en orden aleatorio. En el grupo encubierto, los candidatos diana se correspondían con los pre-expuestos en la Fase 1. En el grupo explícito, los candidatos diana eran los mismos que en el grupo encubierto pero, en lugar de haber sido pre-expuestos en la Fase 1, mostraban en esta Fase 2 el distintivo de compatibilidad. También los candidatos

control eran los mismos que en el grupo encubierto. Por último, en el grupo sin-recomendación, los candidatos diana y control mostrados eran también los mismos que en los otros grupos, pero sin manipulación de ninguno de los candidatos (ni pre-exposición ni distintivo). En todos los grupos las ocho fotografías se contrabalancearon para ser utilizadas como dianas y controles (véase Tabla 2).

Como en experimentos anteriores, para finalizar se preguntó a los participantes por la visibilidad del distintivo de compatibilidad o la percepción de la repetición de los candidatos, y por la posible influencia del distintivo o la repetición en sus puntuaciones. Tras estas preguntas, los participantes fueron informados del verdadero objetivo del experimento y se les facilitó el acceso al pago.

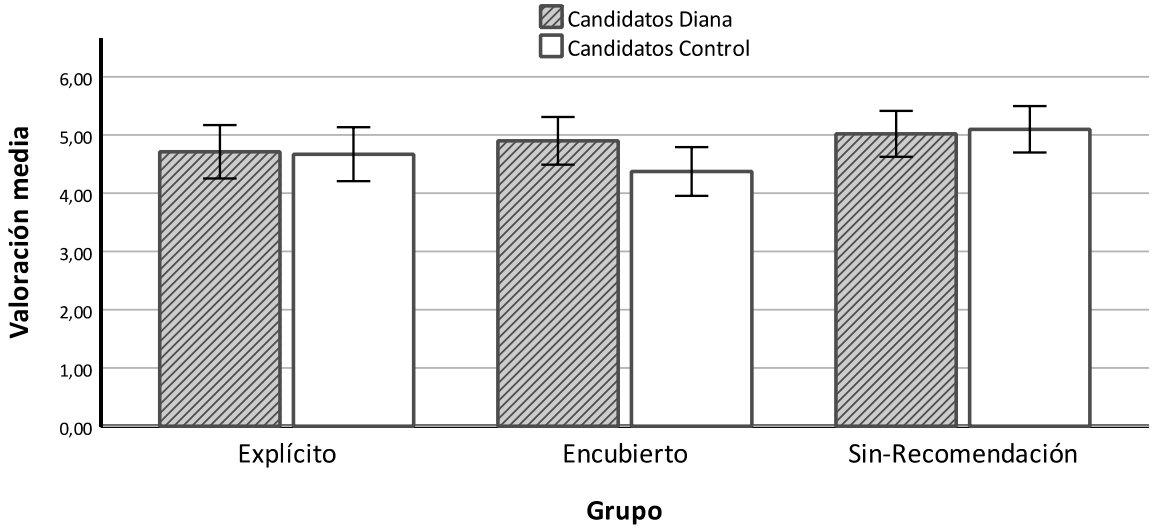
Resultados y Discusión

Los resultados se muestran en la Figura 9. Un ANOVA mixto[4] 2 (candidato: diana, control) x 3 (grupo: explícito, encubierto, sin-recomendación) no mostró efecto principal del candidato ($F(1, 177) = 2.64$, $p = .106$, $\eta^2_p = 0.015$), ni del grupo ($F(2, 177) = 1.43$, $p = .243$, $\eta^2_p = 0.016$), aunque sí se produjo interacción Candidato x Grupo ($F(2, 177) = 13.37$, $p = .037$, $\eta^2_p = 0.037$). Las comparaciones a posteriori mediante la prueba de Tukey revelaron que, al igual que en el experimento previo, las diferencias entre candidatos diana y candidatos control se producían en el grupo encubierto ($t(177) = 3.02$, $p = .034$, $d = 0.33$), pero no en el grupo explícito ($t(177) = 0.24$, $p = 1.00$, $d = 0.02$), ni tampoco en el grupo sin-recomendación ($t(177) = -0.45$, $p = .998$, $d = -0.05$).

[4] Al igual que en experimentos anteriores, dado que las puntuaciones a los candidatos fueron recogidas con una escala Likert, reportamos adicionalmente los resultados del test Wilcoxon para las medias de candidatos diana y control por grupo. Explícito: $Z = 0.221$, $p = .825$. Encubierto: $Z = 2.815$, $p = .005$. Sin-Recomendación: $Z = 0.386$, $p = .699$.

Figura 9

Valoración Media de los Candidatos Diana y Control en los Grupos Explícito, Encubierto y Sin-Recomendación en el Experimento 4



Nota. Las barras de error representan un 95% de CI.

Sin embargo, durante los análisis encontramos dos cuestiones problemáticas. Por un lado, se producía interacción entre el balanceo (utilizado para asegurar que los estímulos se mostraran como dianas y como controles) y el grupo ($F(2, 174) = 4.01$, $p = .020$, $\eta^2_p = 0.044$). Por otro lado, un 44% de los participantes del grupo explícito afirmaban no haber percibido el distintivo de compatibilidad, lo cual era un porcentaje considerable respecto al 22% reportado en los Experimentos 1 y 3. Esta baja visibilidad del distintivo con la nueva base de datos podría deberse a diferentes motivos, entre ellos la diferente forma de las imágenes (véase Figura 8), pero, independientemente de la causa, el aumento en el número de participantes que no percibieron el distintivo ponía en entredicho la validez de los resultados de este experimento.

Los resultados en esta réplica del Experimento 3 parecen corroborar los resultados del mismo en cuanto a la eficacia de la recomendación encubierta en el contexto de las citas románticas y también la falta de efectividad de la recomendación explícita en este campo. Asimismo, y al igual que en el Experimento 3, la diferencia de resultados respecto a los Experimentos 1 y 2 en contexto político plantean la pregunta de si el contexto de decisión pudiera tener un peso clave en la eficacia del estilo persuasivo. Es posible que la recomendación explícita sea más efectiva en política y la recomendación encubierta en citas. Sin embargo, los problemas encontrados por la interacción del balanceo con el grupo y la falta de percepción del distintivo de compatibilidad demandaban un nuevo experimento que permitiera afianzar estas conclusiones. Así, abordamos un nuevo experimento bajo la hipótesis de que el contexto de decisión (política frente a citas) pudiera ser sensible a los diferentes estilos de influencia de la recomendación algorítmica (explícita frente a encubierta).

Experimento 5. Recomendación explícita y encubierta en contexto político y de citas

El objetivo de este nuevo experimento fue replicar, por un lado, los resultados del Experimento 1 donde la recomendación explícita surtió efecto en el contexto político y, por otro, los resultados del Experimento 3 y del Experimento 4 donde la sugerencia encubierta fue eficaz en el contexto de citas. Nuestra hipótesis era que, dependiendo del tipo de recomendación utilizada, explícita o encubierta, el algoritmo influiría en las preferencias de los participantes en un contexto u otro. En concreto esperábamos que la recomendación encubierta del algoritmo resultara más eficaz en el contexto de citas y la explícita en el contexto político.

Dado que a lo largo de esta serie de experimentos habíamos aplicado diversas modificaciones en el procedimiento y alguna de ellas podía haber afectado a los resultados, mantuvimos los parámetros que lograron la manipulación exitosa mediante recomendación explícita en el Experimento 1 y mediante sugerencia encubierta en el Experimento 3. De esta forma, podíamos replicar ambos experimentos y medir qué papel jugaba el contexto en los resultados.

Método

Participantes y Materiales

Reclutamos 400 participantes a través de la plataforma online de Prolific Academic (57.5% mujeres, 0.8% no reportado). Como en el Experimento 3, solicitamos a la plataforma que invitase a participar en nuestra investigación a usuarios de etnia blanca/caucásica y de edades comprendidas entre los 30 y los 45 años (M = 35.80, SD =

4.30), de forma que de nuevo coincidieran con la etnia y la edad de los candidatos de la base de datos del experimento, la de Karras y colaboradores (2018), ya utilizada en el Experimento 4 y calibrada en un experimento previo (véase Apéndice B).

La invitación a participar fue redactada en inglés y el experimento se llevó a cabo también en inglés. De forma automatizada, se asignó al azar a los participantes a uno de los cuatro grupos experimentales: grupo contexto político-recomendación explícita (n = 100), grupo contexto político-recomendación encubierta (n = 102), grupo contexto citas-recomendación explícita (n = 98) y grupo contexto citas-recomendación encubierta (n = 100). El análisis de sensibilidad indicó que, con este tamaño de muestra, contábamos con una potencia del 90% para detectar efectos pequeños (η^2_p = 0.019).

Diseño y Procedimiento

El diseño y el procedimiento se asemejaron a los de experimentos anteriores. Diseñamos un experimento de 2 (candidato: diana, control) x 2 (grupo: explícito, encubierto) x 2 (contexto: político, citas). La Tabla 1 muestra un resumen del diseño experimental, comparado con el resto de experimentos. La variable candidato se manipuló intra-sujetos al igual que en los experimentos anteriores, mientras que las variables grupo y contexto, inter-sujetos.

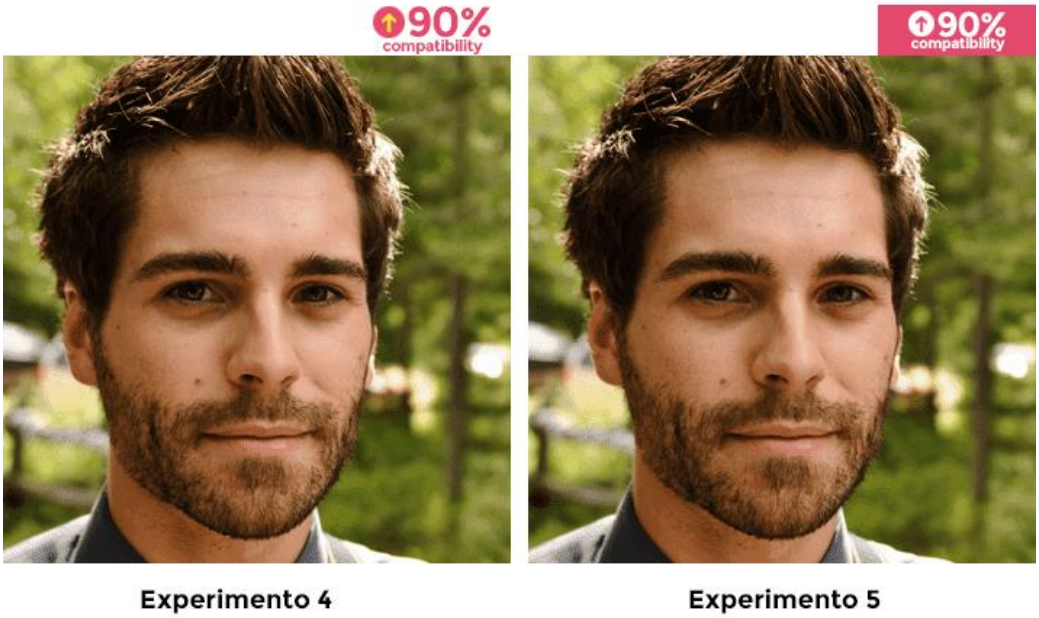
Dependiendo del contexto y tras aceptar el consentimiento informado online y completar las preguntas habituales sobre género y edad, algunos participantes indicaron su orientación política, mientras que otros informaron si tenían pareja o no y su preferencia por hombres o mujeres en un entorno como las plataformas de citas.

Al igual que en los experimentos previos, durante la Fase 0 todos los participantes rellenaron el test de personalidad ficticio y valoraron su conformidad con él. Más tarde fueron expuestos a 40 fotografías de candidatos ficticios durante la Fase 1. En el contexto de citas, todas las fotografías eran de hombres o de mujeres según la preferencia indicada por los participantes, mientras que en el contexto político la mitad de los candidatos eran mujeres y la otra mitad hombres. Al igual que en los experimentos previos, en esta fase las 40 fotografías eran extras, de relleno, en el grupo explícito. Sin embargo, solo 20 de las fotografías eran de relleno en el grupo encubierto, dado que los otros 20 ensayos se componían de cuatro candidatos diana pre-expuestos cinco veces cada uno.

Como las fotografías empleadas en los Experimentos 1, 2 y 3 no eran demasiado actuales ni visualmente atractivas, esta fue la única modificación que introdujimos, utilizando de nuevo la base fotográfica más actual del Experimento 4 (Karras y cols., 2018), con un ajuste en el color del distintivo de compatibilidad para que ganara visibilidad (véase Figura 10).

Figura 10

Distintivo modificado en el Experimento 5 respecto al Experimento 4



Nota. Fotografía de la base de datos fotográfica de Karras y colaboradores (2018), con licencia CC BY 2.0, nombre DSC_3929.jpg y autor Jean-Simon Asselin. Accesible desde https://www.flickr.com/photos/acelain/4852007896

Como en los Experimentos 3 y 4, durante la Fase 2 los participantes valoraron a ocho candidatos en una escala del 1 al 9, con los símbolos de una x y un corazón en los extremos de esta. En el contexto de las citas, los participantes indicaban en la escala su disposición a enviar un mensaje por la web de citas a los candidatos, y en el contexto político su disposición a votarles. Cuatro de los candidatos mostrados, los candidatos diana, eran los que habían sido pre-expuestos durante la Fase 1 en el grupo encubierto y los otros cuatro eran nuevos y actuaban como candidatos control. En el grupo explícito, los candidatos diana mostraban el distintivo de compatibilidad.

En todos los casos, el tiempo de exposición de los candidatos se limitó a 2 segundos, como en el Experimento 1. Todas las imágenes se presentaron en orden aleatorio para cada participante, evitando que un mismo candidato se visualizara más de dos veces seguidas. Además, se aumentó el número de contrabalanceos de los estímulos en su papel como candidatos diana o candidatos control dado los problemas surgidos en el Experimento 4 (véase Tabla 3). Como final del experimento, se recogió el porcentaje de participantes que indicaba haber visto el distintivo de compatibilidad o notado la repetición de candidatos, hasta qué punto consideraban que el distintivo o la repetición habían afectado a sus puntuaciones y, por último, se facilitó a los participantes el acceso al pago además de una explicación sobre los verdaderos objetivos del experimento.

Tabla 3

Contrabalanceos de los Estímulos Diana y Control en el Experimento 5

	D1	D2	D3	D4	C1	C2	C3	C4
Balanceo 1	A	B	C	D	E	F	G	H
Balanceo 2	E	F	G	H	A	B	C	D
Balanceo 3	A	D	F	G	E	H	B	C
Balanceo 4	E	H	B	C	A	D	F	G

Nota. D = Estímulo Diana, recomendado; C = Estímulo Control, no recomendado; A-H = Fotografías de los candidatos.

Resultados y Discusión

Tras comprobar que en esta ocasión, a diferencia del Experimento 4, no se producía un problema de visibilidad del distintivo de compatibilidad en el grupo

explícito (el porcentaje de participantes que afirmaban no haberlo visto se redujo a los niveles de los anteriores experimentos: un 20.7%), realizamos un ANOVA mixto[5] 2 (candidato: diana *vs.* control) x 2 (grupo: explícito *vs.* encubierto) x 2 (contexto: político *vs.* citas) sobre las preferencias de los participantes hacia los candidatos. El ANOVA reveló una triple interacción (Candidato x Grupo x Contexto). Los resultados pueden observarse resumidos en la Tabla 4.

Tabla 4

Análisis de la Varianza de las Valoraciones por Grupo, Candidato y Contexto en el Experimento 5

	df	*F*	*p*	η^2_p
Candidato	1	55.56	< .001	0.123
Grupo	1	0.39	.531	0.001
Contexto	1	53.05	< .001	0.118
Grupo x Candidato	1	0.08	.078	0.000
Grupo x Contexto	1	0.20	.656	0.001
Candidato x Contexto	1	0.09	.765	0.000
Candidato x Grupo x Contexto	1	5.61	.018	0.014
Total	396			

Para ahondar en esta triple interacción, realizamos comparaciones dentro de cada contexto de decisión. Como esperábamos, la recomendación explícita fue efectiva en el contexto político, de modo que los participantes en el grupo explícito

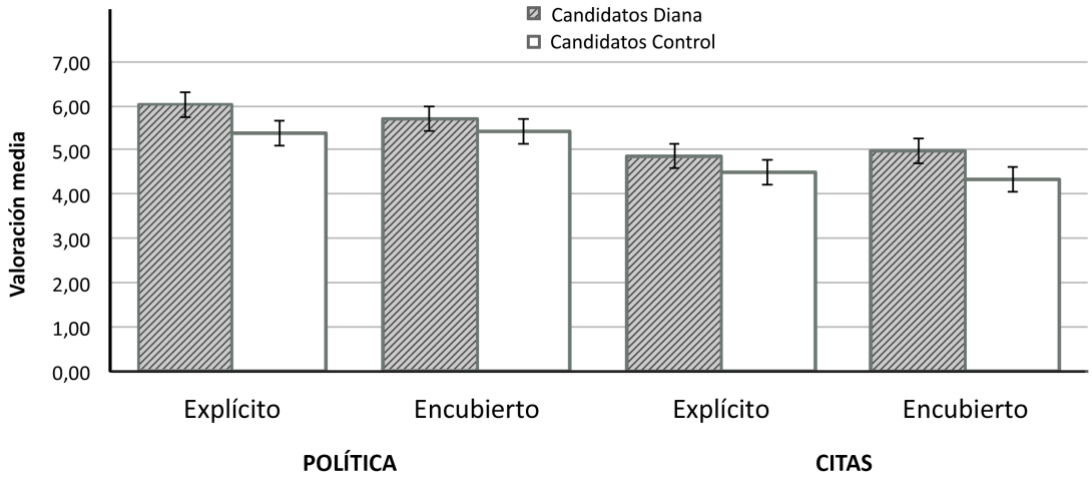
[5] Como en anteriores experimentos, dado que las puntuaciones a los candidatos fueron recogidas con una escala Likert, reportamos también los resultados del test Wilcoxon para las medias de candidatos diana y control por grupo y contexto. En Política: Explícito (Z = 4.342, p = .000); Encubierto: (Z = 2.742, p = .006). En Citas: Explícito (Z = 2.735, p = .006); Encubierto: (Z = 3.900, p = .000). En ambos casos, al realizar análisis más simples dentro de los grupos, en todos los casos se producen diferencias significativas.

mostraron una mayor disposición a votar a los candidatos diana que a los candidatos control, $t(396) = 4.90$, $p < .001$, $d = 0.49$. Esto replica los resultados del Experimento 1.

Además, y en línea con los resultados del Experimento 2, no se observaron diferencias significativas entre los candidatos diana y los candidatos control cuando la recomendación en el contexto político fue encubierta, $t(396) = 2.27$, $p = .310$, $d = 0.26$, a pesar de haber pre-expuesto a los candidatos en cinco ocasiones, como en el contexto de citas del Experimento 3 donde sí habíamos encontrado diferencias en la valoración de los candidatos tras la manipulación encubierta en el contexto de citas. Los resultados se muestran en la Figura 11.

Figura 11

Valoración Media de los Candidatos Diana y Candidatos Control por Grupo y Contexto en el Experimento 5



Nota. Las barras de error representan un 95% de CI.

Por otro lado, y tal y como esperábamos, los participantes del grupo encubierto indicaron una mayor preferencia por los candidatos diana que por los candidatos control en el contexto de las citas, $t(396) = 4.92$, $p < .001$, $d = 0.41$. Una diferencia que, como habíamos predicho, no fue estadísticamente significativa en el grupo explícito, $t(396) = 2.80$, $p = .097$, $d = 0.21$.

La réplica de los Experimentos 1 y 3 nos permite afianzar los resultados encontrados, además de descartar el posible efecto de las diferencias menores entre procedimientos, como por ejemplo la restricción de tiempo en la valoración de candidatos que no afectó a los resultados a pesar de que cambió entre el Experimento 3 y este Experimento 5.

Además, parece que el contexto de decisión (política o citas) podría desempeñar un papel importante en la eficacia de la persuasión de un algoritmo, resultando la decisión política más sensible a la recomendación explícita (Experimentos 1, 2 y 5) y la decisión de citas a la sugerencia encubierta (Experimentos 3, 4 y 5). Sin embargo, esta es una conclusión que es necesario matizar, y así lo abordamos en la discusión de esta serie experimental. Pero antes, incluimos un nuevo capítulo para ahondar en un elemento presente en todos nuestros experimentos sobre la que hasta ahora no hemos hecho hincapié: el test ficticio de personalidad y el falso perfilado de los participantes.

Experimento 6. Confiabilidad de la recomendación algorítmica en contexto político

Tanto los medios de comunicación como las investigaciones en el campo de la IA aseguran que los algoritmos son capaces de inferir todo tipo de detalles sobre nuestra persona, como por ejemplo nuestro estado de ánimo a partir del ritmo de pulsaciones en nuestro teclado del ordenador (Epp y cols., 2011), o nuestra personalidad mediante los "Me gusta" que damos en Facebook o las páginas por las que navegamos (Kosinski y cols., 2012, 2013), o que incluso pueden perfilar nuestra personalidad con mayor precisión que nuestros amigos y familiares (Youyou y cols., 2015). Para ello, los algoritmos no solo se alimentan de los registros de interacción dentro de sus propios sistemas de recomendación, sino que adquieren información extra a grandes empresas de datos. Información sobre nuestro nivel educativo, el número de niños a nuestro cargo, nuestra religión, etnia u opiniones políticas, nuestras búsquedas en Internet, datos de compra, uso de tarjetas de crédito, ingresos o préstamos solicitados, nuestra estabilidad económica, canciones que nos gustan, ubicación, o nuestros planes previstos como tener un bebé o cambiar de trabajo (Acxiom, 2015; Christl, 2017; Solon y Farivar, 2019; WPP's Data Alliance, 2016).

Probablemente debido a esta información que los algoritmos tienen sobre nosotros, las personas atribuimos a estos sistemas una gran capacidad para conocer nuestra personalidad. En los experimentos de Warshaw y colaboradores (2015) y de Springer y colaboradores (2017), por ejemplo, cuando el algoritmo generaba un perfil de la personalidad y del estado de ánimo de los participantes, estos consideraban que

el perfilado era tan acertado que preferían no retocarlo y publicarlo online directamente. A partir de la información que tienen los algoritmos sobre nosotros, además, podrían ajustar sus sugerencias con base en nuestra personalidad y preferencias. Matz y colaboradores (2017) afirman que es posible mejorar la eficacia de los recomendadores de anuncios en Facebook si el texto de estos anuncios hace referencia al perfil psicológico de los individuos; un perfil que en sus experimentos con más de 3.5 millones de participantes era inferido a partir de los “Me gusta” de los participantes en Facebook y sus puntuaciones en el cuestionario Pool Internacional de Ítems de Personalidad (IPIP) (Kosinski y cols., 2015). Bajo esta premisa de que los algoritmos pueden perfilar nuestra personalidad con gran detalle, el historiador Yuval H. Harari afirma que en un futuro cercano “los algoritmos no se rebelarán ni nos esclavizarán; más bien, serán tan buenos a la hora de tomar decisiones por nosotros que sería una locura no seguir sus consejos" (Harari, 2016, p. 253). Argumentos como este, ciertos o no, han sido utilizados para encumbrar a los algoritmos como eficientes decisores en multitud de contextos. Es el caso de las plataformas de citas. A través de autoinformes, pero también mediante el registro de las interacciones de los usuarios en sus plataformas, compañías como eHarmony, PerfectMatch o Chemistry aseguran que pueden recomendar a los candidatos más compatibles gracias al perfilado algorítmico de sus usuarios (Finkel y cols., 2012). Logg y colaboradores (2019), por el contrario, afirman que las personas no necesitamos conocer demasiada información sobre los algoritmos con los que interactuamos para mostrar apreciación ante sus recomendaciones. Así, en su trabajo, los autores muestran que es posible lograr la aceptación del consejo algorítmico sin que los participantes sean previamente

perfilados y sin necesidad de ofrecer detalles sobre el funcionamiento interno del sistema.

En nuestros experimentos anteriores habíamos asumido que la confianza de los participantes hacia nuestro algoritmo ficticio era un requisito para lograr influir de forma explícita en sus preferencias y que esta confianza se había logrado gracias a las supuestas preguntas de personalidad y el informe basado en el efecto Forer. Sin embargo, este supuesto no lo habíamos puesto a prueba hasta el momento, por lo que decidimos incluir un experimento con el objetivo de comprender qué impacto tenía la Fase 0 de confiabilidad sobre nuestros resultados con recomendación explícita. Nuestra hipótesis era que tanto el test de personalidad como el supuesto informe personalizado, facilitados a los participantes en los anteriores experimentos, era un elemento necesario para generar confianza hacia el algoritmo, así como para que los participantes aceptaran la recomendación algorítmica explícita en el contexto político. Nuestro informe de personalidad basado en el efecto Forer (1949) había recibido unas valoraciones de precisión entre moderadas y altas en todos los experimentos (*M* = 6.71, *SD* = 1.65, en una escala de 1 a 9, en el Experimento 1; *M* = 6.89, *SD* = 1.65 en el Experimento 2; *M* = 6.78, *SD* = 1.56 en el Experimento 3; *M* = 6.83, *SD* = 1.41 en el Experimento 4; y *M* = 6.85, *SD* = 1.46 en el Experimento 5).

Método

Participantes y Materiales

Para este experimento contamos con 300 participantes[6] (45% mujeres, 1.7% no reportado) con edades comprendidas entre los 18 y los 70 años (*M* = 27.5, *SD* = 9.05) y etnia blanca/caucásica, reclutados mediante la plataforma Prolific Academic. En esta ocasión consideramos que la franja de edad podía ser más amplia, como en los Experimentos 1 y 2 pero a diferencia de experimentos en el contexto de citas, puesto que en el contexto político no parecía necesario que los participantes coincidieran en edad con los candidatos. La muestra se dividió aleatoriamente en dos grupos: grupo experimental forer; *n* = 150; y grupo control noforer, *n* = 150.

La base fotográfica utilizada fue de nuevo la de Karras y colaboradores (2018), también usada en los Experimentos 4 y 5 (véase Apéndice B). Dado que habíamos consolidado el procedimiento en los experimentos anteriores, decidimos pre-registrarlo en AsPredicted. Puede consultarse online: https://aspredicted.org/vi32d.pdf.

Diseño y Procedimiento

Planteamos un diseño mixto entre grupos (con efecto Forer o sin él) e intra-sujetos (valoración de candidatos diana y control) y replicamos parte del procedimiento del Experimento 5, en concreto el del grupo explícito en el contexto político, para uno de los grupos (grupo forer), siendo el otro grupo (noforer) idéntico,

[6] Dado que en los meses anteriores se había alertado de la presencia de bots automatizados suplantando a participantes humanos en la plataforma de reclutamiento Amazon Mechanical Turk, quisimos asegurarnos de que no ocurría lo mismo en Prolific Academic. Para ello, incluimos al inicio del experimento una operación matemática sencilla (una suma) en formato de imagen. Dado que no hubo ninguna respuesta incorrecta, no se descartó a ningún participante de la muestra y corroboramos la fiabilidad de la plataforma y no fue necesario repetir esta prueba en los siguientes experimentos.

pero sin la fase de perfilado e informe de personalidad. Tras el consentimiento informado online, las preguntas demográficas habituales (género y edad) y registrar si la profesión o estudios de los participantes tenía relación con la tecnología, inteligencia artificial, robots o algoritmos, por si esto pudiera afectar a los resultados ("¿Su trabajo o estudios están relacionados con la tecnología, la inteligencia artificial, los robots o los algoritmos?" / "Are your work or studies related to technology, artificial intelligence, robots or algorithms?"), repetimos la Fase 0 de los experimentos anteriores, aunque solo para los participantes del grupo forer. Así, después de indicar su orientación política, los participantes completaron el cuestionario de personalidad y recibieron el supuesto informe personal elaborado por el algoritmo (véanse Figuras 2 y 3). Por su parte, el grupo control noforer solo respondió a la pregunta de la orientación política, saltándose el resto de preguntas y el informe. Como novedad, todos los participantes señalaron el nivel de confiabilidad que asociaban a la recomendación algorítmica en el contexto de decisiones políticas ("¿Hasta qué punto considera confiables los algoritmos de inteligencia artificial para recomendar candidatos políticos compatibles?" / "How trustworthy do you consider artificial intelligence algorithms for recommending compatible political candidates?"). Además, en esta ocasión decidimos prescindir de la Fase 1 con imágenes de relleno, puesto que no parecía un requisito para la influencia explícita perseguida en este experimento (véase Tabla 5).

Tabla 5

Resumen del Diseño del Experimento 6

Grupo	Fase 0	Fase Valoraciones
Forer	Test ficticio de personalidad e informe	D1-D4 (*) C1-C4
NoForer	Sin test ni informe	D1-D4 (*) C1-C4

Nota. Las imágenes pueden ser D = Diana, recomendado; o C = Control, no recomendado; (*) = distintivo "+90% compatibilidad"; el papel de los estímulos como candidatos diana (DI-D4) y control (C1-C4) fue contrabalanceado.

La siguiente tarea para los participantes consistió en la valoración de ocho candidatos: cuatro candidatos diana, que portaban el distintivo de compatibilidad, y cuatro candidatos control. Al igual que en los experimentos de recomendación explícita, los ocho candidatos a puntuar se presentaron con un límite de dos segundos de visualización. De nuevo se aumentó el número de contrabalanceos en las imágenes que hacían las veces de estímulos diana y control (véase Tabla 6).

Tabla 6

Contrabalanceos de los Estímulos Diana y Control en el Experimento 6

	D1	D2	D3	D4	C1	C2	C3	C4
Balanceo 1	A	C	F	H	B	D	E	G
Balanceo 2	B	C	F	G	A	D	E	H
Balanceo 3	A	D	E	F	B	C	E	G
Balanceo 4	B	E	G	H	A	C	D	F
Balanceo 5	C	D	E	H	A	B	F	G
Balanceo 6	A	B	D	G	C	E	F	H

Nota. D = Estímulo Diana, recomendado; C = Estímulo Control, no recomendado; A-H = Fotografías de los candidatos.

Tras esta tarea, pedimos a los participantes que volvieran a responder a la pregunta sobre la confianza en los algoritmos en decisiones políticas que habían rellenado al inicio del experimento. Además, debían valorar cómo de efectivo consideraban que había sido el algoritmo en su labor de recomendarles candidatos ("¿Cómo ha sido de eficaz nuestro algoritmo a la hora de recomendar candidatos?" / "How effective has our algorithm been in recommending candidates?"), cómo habían cambiado sus expectativas sobre el desempeño del algoritmo respecto al inicio del experimento ("Teniendo en cuenta sus expectativas al principio del experimento, ¿cómo realizó nuestro algoritmo la tarea?" / "Considering your expectations at the beginning of the experiment, how well did our algorithm perform the task?") y hasta qué punto consideraban confiable el consejo algorítmico en todo tipo decisiones ("¿Hasta qué punto considera confiable el consejo que pueden ofrecer los algoritmos de inteligencia artificial, en general, en cualquier otro ámbito de decisión?" / "To what point do you consider trustworthy the advice that artificial intelligence algorithms can offer, in general, in any other decision-making matter?"). Como cierre del experimento, se recogió el porcentaje de participantes que afirmaba haber visto el distintivo de compatibilidad y su valoración sobre si este podía haber afectado a sus puntuaciones. Además, se facilitó a los participantes el acceso al pago y se les explicó el verdadero propósito del experimento.

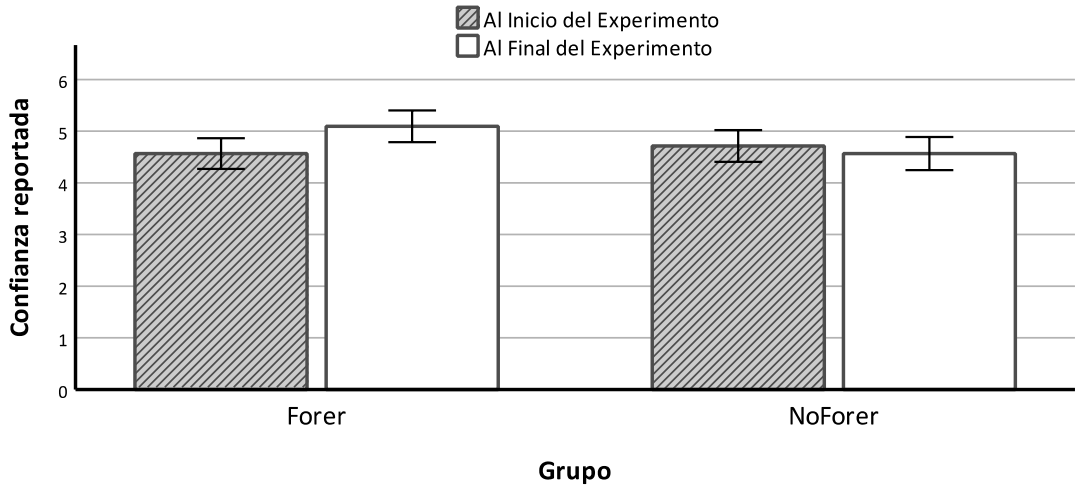
Resultados y Discusión

Contradiciendo nuestra hipótesis, un ANOVA mixto[7] 2 (candidato: diana *vs.* control) x 2 (grupo: forer *vs*. noforer) solo reveló un efecto principal del candidato (F(1, 298) = 25.22, $p < .001$, $\eta^2_p = 0.078$). No se produjo ni efecto principal de grupo (F(1, 298) = 0.84, $p = .361$, $\eta^2_p = 0.003$), ni interacción Candidato x Grupo (F(1, 298) = 0.50, $p = .480$, $\eta^2_p = 0.002$). Los candidatos recomendados de forma explícita por el algoritmo se valoraron mejor que los candidatos control tanto en el grupo forer ($M_{Diana} = 5.63$, $SD_{Diana} = 1.31$, $M_{Control} = 5.23$, $SD_{Control} = 1.29$, t(298) = 4.05, $p <.001$, $d = 0.31$) como en el noforer ($M_{Diana} = 5.45$, $SD_{Diana} = 1.37$, $M_{Control} = 5.16$, $SD_{Control} = 1.32$, t(298) = 3.05, $p =.013$, $d = 0.22$), por lo que esta fase de confiabilidad del algoritmo a base de preguntas de personalidad no parece ser un requisito para ejercer la influencia. Sin embargo, la manipulación del efecto Forer sí tuvo un efecto en las percepciones autoinformadas de los participantes hacia los algoritmos. Así, encontramos que, en el grupo forer, la confiabilidad de los algoritmos políticos aumentó desde el inicio hasta el final del experimento (t(298) = -3.83, $p = .002$, $d = -0.33$), mientras que no se produjo tal efecto en el grupo noforer (t(249) = 0.93, $p = .792$, $d = 0.08$). En ambos grupos, la confianza reportada al inicio del experimento era prácticamente igual (t(378) = 0.40, $p = .978$, $d = 0.05$; véase Figura 12). La profesión o estudios de los participantes no tuvo ningún impacto en los resultados.

[7] Reportamos también los análisis no-paramétricos realizados. Los resultados del test Wilcoxon para las medias de candidatos diana y control por grupo fueron: En grupo forer ($Z = 3.848$, $p = .000$); En noforer ($Z = 3.182$, $p = .001$).

Figura 12

Confianza en los Algoritmos Políticos por Grupo al Inicio y al Final del Experimento



Nota. Las barras de error representan un 95% de CI.

Además, los participantes del grupo forer consideraron al algoritmo del experimento más eficaz (M = 5.14, SD = 1.64, $t(298) = 1.77$, $p = .039$, $d = 0.20$) que los participantes del grupo noforer (M = 4.79, SD = 1.81), e indicaron que el desempeño del algoritmo cumplió mejor con sus expectativas (M = 6.13, SD = 1.69, $t(298) = 4.75$, $p < .001$, $d = 0.5$) en comparación con el grupo control noforer (M = 5.16, SD = 1.85). Solo la confiabilidad de los algoritmos en general, es decir, en cualquier ámbito de decisión, se mantuvo en niveles similares entre grupos (M_{Forer} = 5.32, SD_{Forer} = 1.68; $M_{NoForer}$ = 5.31, $SD_{NoForer}$ = 1.80; $t(298) = 0.03$, $p = .487$, $d = 0.01$).

Los resultados del experimento parecen indicar que la Fase 0 de confiabilidad no resulta determinante para que el algoritmo ejerza su influencia mediante la recomendación explícita a los participantes, puesto que, tanto en el grupo noforer como en el grupo forer, los participantes se encontraban más inclinados a votar a los

candidatos diana que a los candidatos control en el contexto político. Sin embargo, esta fase sí incrementó la eficacia percibida del algoritmo y el cumplimiento de las expectativas, además de aumentar la confiabilidad de los algoritmos en el contexto de las decisiones políticas, aunque no en los algoritmos en general. Si bien puede que no hubiera sido necesario en los experimentos previos la incorporación de esta fase como requisito previo a la influencia, parece que esta sí generó credibilidad y confiabilidad hacia el algoritmo, por lo que no podemos descartar que haya contribuido a los resultados observados.

La Unión Europea afirma que para garantizar una IA confiable es necesario que la IA favorezca "la agencia y la supervisión humanas", posea "solidez y seguridad técnicas", garantice "la privacidad y la gobernanza de los datos", proporcione "transparencia", respete "la diversidad, la no discriminación y la equidad", promueva "el bienestar social y ambiental" y permita "la rendición de cuentas" (European Commission, 2019). Sin embargo, tal y como muestran nuestros resultados, parece que los algoritmos necesitan mucho menos que todo esto para generar confianza en su interacción con las personas, lo cual, en última instancia, podría resultar peligroso según cuáles sean sus fines.

Discusión de la Serie Experimental 1

A lo largo de seis experimentos encontramos que los algoritmos pueden influir en decisiones humanas importantes, tanto de manera explícita como encubierta. Además, parece que el contexto puede afectar al estilo de persuasión algorítmica más eficaz para cada situación.

En el Experimento 1 sobre recomendación explícita en política, observamos que el consejo algorítmico inclinaba las preferencias de voto de los participantes hacia los candidatos diana. En el Experimento 2 sobre recomendación encubierta en política, no logramos influir en los participantes, pero sí conseguimos influirles de manera encubierta en un contexto diferente, el de las citas en el Experimento 3. El aumento de la familiaridad de los candidatos mediante su mayor pre-exposición en este experimento permitió que los candidatos diana fueran mejor valorados que los candidatos control. No obstante, en este experimento la recomendación explícita no consiguió afectar a las preferencias de citas de los participantes, al contrario de lo sucedido en el contexto político del Experimento 1. Tras un intento de réplica en el Experimento 4, en el que sufrimos algunos problemas de balanceo y visibilidad del distintivo de compatibilidad, los resultados de los experimentos previos se replicaron y ampliaron con el Experimento 5, donde pusimos a prueba ambos estilos de influencia y ambos contextos. Por un lado, nuestro algoritmo (ficticio) fue capaz de influir de nuevo en la disposición a votar a los candidatos diana, aceptándose su recomendación explícita en el contexto político (como en el Experimento 1), pero no su sugerencia encubierta en este mismo contexto político (como en el Experimento 2). Por otro lado,

los participantes no aceptaron la recomendación explícita de candidatos en el contexto de las citas, pero sus preferencias sí fueron influenciadas de manera encubierta, replicando así también los resultados del Experimento 3 y el Experimento 4. Esta réplica de los experimentos anteriores se logró además a pesar de los muchos cambios introducidos en los procedimientos durante toda la serie (incluido el conjunto fotográfico), lo que sugiere la solidez de los resultados.

Aunque nuestros experimentos podrían parecer relacionados con las investigaciones de apreciación y aversión a los algoritmos, son muchas las diferencias entre nuestros experimentos y estos estudios. Como ya indicamos, nuestro propósito era simplemente comprobar si la recomendación algorítmica puede influir en las preferencias de las personas, sin que ello implique necesariamente una apreciación o aversión de la recomendación. Por ejemplo, en nuestros experimentos de recomendación encubierta, puede que los participantes no fueran siquiera conscientes de la recomendación ofrecida por el algoritmo. Pero, aunque lo hubieran sido, que siguieran la recomendación tampoco implica apreciación.

A diferencia de la literatura sobre apreciación y aversión al algoritmo, consideramos que nuestros experimentos fueron más ecológicos, ya que ofrecieron la recomendación algorítmica explícita a los participantes de forma semejante a cómo se presenta en los recomendadores de uso habitual como Tinder o Netflix, es decir: 1) mostrando el grado de compatibilidad de los candidatos con los participantes; 2) registrando el comportamiento en lugar de usar autoinformes; 3) especificando que la recomendación proviene de un “algoritmo de inteligencia artificial”, por ser este concepto más representativo del estado de la tecnología hoy en día; y 4) sin

contraponer la aceptación de la recomendación algorítmica al consejo de un humano, dado que esta es una situación más natural en los contextos estudiados.

Nuestros resultados también podrían relacionarse con la literatura sobre tecnología persuasiva. En esta literatura proliferan los trabajos sobre el efecto de las recomendaciones de los algoritmos en el comportamiento o las actitudes cuando los algoritmos muestran a los usuarios una explicación personalizada de los motivos por los que se les recomienda un objeto o contenido específico (Gkika y Lekakos, 2014; Noorbehbahani y Zarein, 2018). Lo que hemos visto con los experimentos de esta serie es que el hecho de que los algoritmos no aporten una explicación sobre sus recomendaciones no invalida el poder persuasivo de estos, afectando igualmente al comportamiento de las personas, lo que sugiere que este poder podría ser incluso mayor del que se les suele atribuir a estas recomendaciones algorítmicas.

La literatura sobre tecnología persuasiva suele apoyarse en los modelos de proceso dual que discutimos en la Introducción, y lo hace bien abordándolos desde la perspectiva de la Psicología Social en referencia a la persuasión y el cambio de actitudes, o bien desde la Psicología Básica y la investigación sobre juicios y toma de decisiones. Como también explicamos en la Introducción, desde la Psicología Social, modelos como el Modelo Sistemático Heurístico de Chaiken y colaboradores (1989) o el Modelo de Probabilidad de Elaboración de Petty y Cacioppo (1986) teorizan que la persuasión puede producirse por dos vías, similares a los dos tipos de procesamiento planteados por las teorías de proceso dual en Psicología de los juicios y la toma de decisiones (Evans y Stanovich, 2013; Samson y Voyer, 2012; Stanovich y West, 2000). Por un lado, cuando las personas se encuentran motivadas para procesar activamente

una comunicación persuasiva, centrándose en la información y los argumentos ofrecidos, la persuasión se lograría por la *ruta central* de la persuasión (Sistema 2, Tipo 2 o sistema reflexivo, en Psicología de los juicios y la toma de decisiones). Por otro lado, ante la falta de motivación o la escasez de recursos (de tiempo, cognitivos…) de la persona, la persuasión se podría conseguir a través de la llamada *ruta periférica* (Sistema 1, Tipo 1 o sistema automático), utilizando para ello claves heurísticas que desencadenarían la aceptación sin que ello conllevase una gran reflexión. Aunque en nuestros experimentos no ofrecíamos a los participantes ninguna explicación razonada sobre sus recomendaciones, nuestro distintivo de compatibilidad en los experimentos de recomendación explícita y la repetición de los candidatos en los experimentos de recomendación encubierta podrían haber actuado como señales periféricas para desencadenar la persuasión del algoritmo.

Respecto a las limitaciones de nuestro trabajo, hay que señalar que quizá el contexto experimental pudo parecer poco realista a los participantes dado que sus decisiones no acarreaban consecuencias en la vida real. Sin embargo, al ser una limitación que hubiese afectado por igual a todas las condiciones, entendemos que los resultados encontrados no podrían explicarse con base en ello.

Otra limitación que podría achacarse a nuestro trabajo es que los participantes solo podían evaluar a los candidatos mediante sus fotografías, ya que no se les ofrecía ninguna referencia sobre sus programas políticos ni sobre sus perfiles en las plataformas de citas. No obstante, como señalamos anteriormente, el aspecto físico resulta ser una de las características que más afecta a los votantes en el contexto de decisión política (Palmer y Peterson, 2015; White y cols., 2013) y a la interacción entre

candidatos en las webs de citas (Hern, 2014). Utilizar otros parámetros, como mostrar la ideología de cada candidato en el contexto político, por ejemplo, hubiera implicado probablemente un gran impacto en la preferencia de los participantes, al ser este el atributo con mayor peso en una votación como indican Bonneau y Cann (2015) y Sances (2018). Sin embargo, en esa situación los resultados se hubieran podido interpretar como producto de dicho atributo y no de la influencia del algoritmo, por lo que preferimos simplificar y reducir el número de variables e interacciones potenciales dado el propósito de nuestro experimento.

Cabe señalar que, aunque somos conscientes de que los resultados encontrados no implican que la influencia resulte tan poderosa como para que las personas con ideología política de extrema izquierda voten a la extrema derecha o viceversa, consideramos que nuestros experimentos muestran la capacidad de los algoritmos para inclinar las preferencias hacia unos candidatos cuando no existe una fuerte preferencia previa.

Por otro lado, no descartamos que quizá la influencia conseguida por la recomendación algorítmica también pueda alcanzarse con otro tipo de estrategias de influencia, independientes de los algoritmos. Por ejemplo, mostrar información sobre el comportamiento social en un sistema de recomendación (conocer las preferencias de otros usuarios) puede afectar significativamente a las decisiones (Zhu y Huberman, 2014). Gunaratne y colaboradores (2018), de hecho, compararon ambos tipos de influencia (social y algorítmica) en un estudio sobre recomendaciones de ahorro para la jubilación. Los autores encontraron que la recomendación del algoritmo ejercía más influencia que la lograda al mostrar las preferencias sociales. Sin embargo, la variedad

de parámetros que pueden estar incidiendo en ese resultado es inmensa (desde la descripción del algoritmo, hasta el volumen y autoridad de las personas que componían la recomendación social), por lo que no resulta generalizable y quizá fuera incluso posible invertir ese resultado utilizando una recomendación social más relevante y creíble que la del algoritmo. Aunque esta comparación y la comparación con la recomendación humana podrían tener interés en aplicaciones concretas, no eran objetivos de nuestros experimentos.

Es importante remarcar que nuestro falso y sencillo algoritmo consiguió dirigir las preferencias de los participantes sin necesidad de establecer perfiles individualizados de estos. Nuestro ficticio test de personalidad, muy lejos de realizar un perfilado sofisticado como el que supuestamente elaboran los sistemas de recomendación algorítmica, solo se utilizó para aumentar la confianza de los participantes, pero no fue determinante en el cambio de sus preferencias tal y como se detalla en el Experimento 6. Esto no descarta que un algoritmo más sofisticado, como aquellos con los que las personas interactúan en su vida cotidiana a través de Internet, pueda ejercer una influencia mucho mayor. A modo de ejemplo, imaginemos la influencia que el algoritmo de Facebook, con el volumen ingente de datos que posee de sus usuarios, puede alcanzar con su nuevo servicio de citas en línea (Sharp, 2019).

Otra lectura interesante que podemos extraer de esta primera serie experimental es que existe un número infinito de variaciones posibles para ajustar los algoritmos, sus formatos y parámetros. Dado que algunas variaciones pueden producir importantes diferencias en su eficacia, cabe preguntarse si el algoritmo de

recomendación encubierta hubiera podido influir también en las decisiones políticas de haber utilizado diferentes parámetros o un heurístico diferente al de familiaridad.

Como muestran nuestros experimentos, encontrar los mejores parámetros para que la recomendación algorítmica sea más persuasiva es una cuestión de experimentación. Y aquí es importante señalar que la velocidad con la que los científicos académicos humanos podemos realizar nuevos experimentos y recoger nuevos datos es muy lenta, en comparación con la facilidad con la que muchas empresas de IA realizan experimentos con millones de seres humanos diariamente a través de Internet. Estas compañías pueden probar tantas hipótesis como quieran con muestras tan grandes como necesiten hasta encontrar las recomendaciones algorítmicas más eficaces, por lo que su capacidad para influir en las decisiones, el comportamiento y las actitudes de las personas, tanto de forma explícita como encubierta, es claramente mucho mayor que la mostrada en la presente investigación. Su capacidad para conocer lo que impulsa el comportamiento humano y cómo dirigirlo en un entorno tecnológico está, en orden de magnitud, por delante de la psicología académica y otras ciencias sociales (Lazer y cols., 2009), por lo que es necesario aumentar la cantidad de estudios científicos disponibles públicamente sobre la influencia de los algoritmos de IA en las decisiones y el comportamiento humano.

En resumen, nuestros experimentos muestran que la recomendación algorítmica, al mostrarse de forma explícita o encubierta, puede afectar a las decisiones a pesar de que los sistemas de recomendación pueden parecer entornos donde los usuarios conservan su autonomía y libertad de elección completa (Araujo y cols., 2020).

Dado que el objetivo de esta serie experimental era investigar empíricamente esta capacidad de influencia de los algoritmos y no tanto identificar las causas de las diferencias observadas entre los contextos de política y citas, abordaremos este tema a continuación. Para ello, evaluaremos hasta qué punto la falta de aceptación de la recomendación explícita en citas podría deberse a dos factores relacionados con las capacidades que las personas asociamos a los algoritmos: la posible subjetividad de la tarea y el contexto de decisión; y la atribución de ciertas habilidades, consideradas exclusivas de los humanos, a los algoritmos.

Parte III. La influencia de las capacidades algorítmicas en las decisiones y juicios

Capítulo 4. **Atribución de capacidades a los algoritmos**

Al final del capítulo anterior, abríamos la puerta a investigar cuáles podrían ser las razones por las que el algoritmo resultó influyente con recomendación explícita en política, pero no en el contexto de las citas. Aunque siempre cabe la posibilidad de que nuestros resultados pudieran haber sido diferentes de usar otros parámetros, lo cierto es que desde la literatura de la aversión al algoritmo se ha tratado de explicar la falta de eficacia del consejo explícito en algunos dominios con base en varios factores que podrían resultar interesantes para su investigación y que están relacionados con las capacidades que las personas asociamos a los algoritmos. Entre estos factores estaría la objetividad de las tareas de decisión y las habilidades atribuidas a los algoritmos (Castelo, 2019; Castelo y cols., 2019; Inbar y cols., 2010; Logg y cols., 2019; Longoni y Cian, 2020).

Empecemos por el primero de estos factores. Comúnmente se perciben como tareas objetivas aquellas que requieren un análisis lógico, basado en reglas. Las tareas subjetivas, por el contrario, serían aquellas que requieren de emoción o intuición (Inbar y cols., 2010). Según los escasos estudios que abordan esta relación entre los algoritmos y la objetividad de las tareas, las personas rechazarían el consejo algorítmico en tareas hedónicas, con alta emocionalidad implicada (como podría ser el contexto de citas), prefiriendo en estas la recomendación humana (Castelo, 2019). Por su parte, en tareas utilitarias, más racionales y lógicas (como podría ser, al menos en

principio, el contexto de voto político), las personas preferirían la recomendación algorítmica por encima del consejo humano (Logg y cols., 2019).

Esta cuestión ha sido abordada experimentalmente por Castelo y colaboradores (2019). En el primer experimento de su trabajo, por ejemplo, los autores presentaban a los participantes, reclutados a través de Amazon Mechanical Turk, una lista de tareas para indicar el grado de objetividad, trascendencia y familiaridad de estas. Actividades artísticas como componer canciones fueron calificadas como tareas poco objetivas (30 puntos sobre 100), al igual que tareas como la recomendar una pareja romántica (con 26 puntos), similar a la de nuestros experimentos en el contexto de citas. En el extremo contrario, tareas como conducir un coche (69 puntos), diagnosticar una enfermedad (77) o predecir el tiempo (68) recibieron puntuaciones altas en objetividad. La tarea de predecir los resultados de unas elecciones democráticas, la más similar a la utilizada en nuestros experimentos de persuasión explícita, se evaluó con 57 puntos de objetividad sobre 100.

A continuación, una muestra diferente de participantes declaraba cuánto confiarían que en que cada una de las tareas fuera desempeñada por un algoritmo o por un humano. Los autores encontraron que, aunque en promedio los participantes confiaban más en el desempeño humano, su confianza en los algoritmos era mayor en las tareas valoradas como objetivas. Según los autores, el grado de objetividad atribuido a una tarea de decisión podría ser un rasgo que determinara cuándo se produce apreciación hacia la recomendación explícita de un algoritmo y cuándo causa aversión, una teoría compartida por otros investigadores como Logg y colaboradores (2019).

Por su parte, Longoni y Cian (2020) señalan que las personas consideran los sistemas de recomendación algorítmica más competentes que los consejeros humanos cuando la decisión implica sopesar atributos utilitarios, pero no cuando se trata de evaluar atributos hedónicos, a partir de un razonamiento similar al de Castelo y colaboradores (2019). Así, las personas confiarían más en los algoritmos a la hora de recomendar elementos utilitarios porque la toma de decisión implica basarse en hechos, lógica y criterios racionales, al contrario que en las decisiones con atributos hedónicos implicados, donde las personas suelen juzgar la recomendación a partir de emociones, criterios sensoriales e intuición.

Este impacto en la confianza hacia la recomendación algorítmica dependiendo de la objetividad/utilidad vs. subjetividad/hedonismo de la tarea podría estar a su vez relacionado con el otro factor mencionado al comienzo de este capítulo: el de las habilidades atribuidas a los algoritmos de IA.

Como ya mencionamos en la Introducción, según Sundar (2008), las personas asocian a los algoritmos de IA rasgos o habilidades positivas, como la objetividad, la falta de sesgo y la neutralidad, al tiempo que también les atribuyen rasgos negativos como inflexibilidad, carencia de emociones y frialdad (Sundar, 2020). Por ejemplo, Jago (2019) mostró, en su trabajo sobre la autenticidad de un algoritmo, que los participantes valoraban diferente la obra artística de un algoritmo que la de un humano, aunque en realidad la autoría de la obra pertenecía al algoritmo en todo momento. El autor utilizó como medidas dos dimensiones de autenticidad: la autenticidad tipo, es decir, si la obra se consideraba auténtica para poder clasificarse como arte; y la autenticidad moral, por la que una obra se consideraba genuina si

reflejaba los valores o motivaciones de su creador. Según sus resultados, cuando los participantes creían que la pieza artística era obra de un humano, la consideraban más auténtica que cuando sabían que era obra del algoritmo, pero solo en términos de autenticidad moral, no de autenticidad tipo. Es decir, los participantes aceptaban que se calificara la obra del algoritmo como arte, pero no consideraban aceptable que la pieza pudiera reflejar la motivación, la esencia o los valores del artista.

Por todo ello, decidimos explorar la respuesta de las personas a los algoritmos en contextos de decisión subjetiva con una nueva serie experimental que desarrollamos en el siguiente capítulo.

Capítulo 5. **Serie Experimental 2**

Esta serie cuenta con tres nuevos experimentos enfocados a comprender si la atribución de ciertas capacidades a los algoritmos puede influir en la falta de aceptación de la recomendación algorítmica o en los juicios sobre su desempeño en tareas consideradas poco objetivas, como la búsqueda de pareja (con 26 puntos sobre 100 en el trabajo de Castelo y colaboradores, 2019) o la composición musical (con 30 puntos).

El primer experimento se orienta a comprender si la subjetividad de la tarea de citas pudiera estar provocando la falta de aceptación de la recomendación explícita en este contexto. Los otros dos experimentos, por su parte, evalúan si la atribución de habilidades humanas a los algoritmos afecta a la percepción y juicio sobre su desempeño en un contexto subjetivo diferente al usado hasta ahora: el contexto del arte.

Experimento 7. Objetividad y subjetividad de la tarea en contexto de citas

Como adelantábamos, Castelo y colaboradores (2019) han abordado el efecto de la subjetividad de la tarea en la aversión al algoritmo en uno de los pocos trabajos existentes sobre el tema. En él, los autores describen a los participantes una tarea aparentemente objetiva (estimar un valor bursátil) de forma diferente según su grupo experimental. En un grupo, la tarea se presenta con un enfoque subjetivo, sugiriendo a los participantes que la mejor forma de realizarla es confiando en la intuición humana.

En el otro grupo, la tarea se describe como objetiva, explicando que la mejor estrategia para llevarla a cabo es considerando datos como la oferta y la demanda del sector o el precio del valor bursátil. Además, los autores manipulan la información facilitada sobre cuáles son las capacidades actuales de los algoritmos. Mientras que a unos participantes se les transmite que los algoritmos actualmente guardan una alta semejanza con los humanos al poseer habilidades subjetivas, a otros se les indica que la semejanza de los algoritmos con los humanos es baja porque los sistemas actuales solo destacan en habilidades objetivas. Según sus resultados, cuando el algoritmo se presenta con baja semejanza humana, las personas confían en los algoritmos si la tarea se describe como objetiva, pero no si se presenta como subjetiva. Sin embargo, si el algoritmo se presenta con alta semejanza humana, las personas confían en el algoritmo en tareas descritas como objetivas y en tareas descritas como subjetivas.

Para contribuir a este campo y arrojar algo de luz sobre la falta de influencia de la recomendación explícita en el contexto de citas de nuestra primera serie experimental, decidimos abordar un nuevo experimento con el objetivo de comprobar si presentar la tarea de citas como una tarea objetiva (o subjetiva) y recordar a los participantes las habilidades objetivas (o subjetivas) del algoritmo podía incrementar la influencia de su recomendación. Dado que nuestra tarea de partida era aparentemente subjetiva (preferencia por candidatos en el contexto de citas) y no objetiva (estimar un valor bursátil) como en Castelo y colaboradores (2019), nuestra hipótesis fue que la influencia explícita se lograría en la tarea de citas si la presentábamos como subjetiva, pero además señalábamos que los algoritmos cuentan con las habilidades necesarias para desempeñar este tipo de tareas subjetivas.

Método

Participantes y Materiales

Reclutamos a 240 participantes caucásicos (55.8% mujeres, 0.8% no reportado), con edades entre los 30 y los 45 años (M = 36.3; SD = 4.38) para que, como en anteriores experimentos, coincidieran en la etnia y edad de los candidatos a mostrar. De nuevo el experimento se realizó en inglés. Los participantes fueron repartidos aleatoriamente en dos grupos: objetivo (n = 121) y subjetivo (n = 119). El análisis de sensibilidad indicó que contábamos con una potencia del 90% para detectar efectos de tamaño pequeño (η^2_p = 0.017).

La base fotográfica utilizada fue la misma de los Experimentos 4, 5 y 6 (véase Apéndice B), de Karras y colaboradores (2018). El experimento fue pre-registrado en AsPredicted. Puede consultarse online: https://aspredicted.org/x2va7.pdf

Diseño y Procedimiento

Además de inspirarnos en el trabajo de Castelo y colaboradores (2019) para las instrucciones a los participantes, utilizamos gran parte del procedimiento utilizado en nuestro Experimento 5; en concreto el procedimiento del contexto de citas y el grupo explícito, con algunos cambios que indicamos a continuación. La Tabla 7 muestra un resumen del diseño experimental.

Tabla 7

Resumen del Diseño del Experimento 7

Grupo	Fase 0	Fase 1	Fase 2
Objetivo	Habilidades objetivas de los algoritmos y test de personalidad (sin informe)	Atributos objetivos de la tarea	D1-D4 (*) C1-C4
Subjetivo	Habilidades subjetivas de los algoritmos y test de personalidad (sin informe)	Atributos subjetivos de la tarea	D1-D4 (*) C1-C4

Nota. Las imágenes pueden ser D = Diana, recomendado; o C = Control, no recomendado; (*) = distintivo "+90% compatibilidad"; el papel de los estímulos como candidatos diana (DI-D4) y control (C1-C4) fue contrabalanceado.

Como en el Experimento 5, tras aceptar el consentimiento informado online los participantes respondieron a las preguntas habituales de género, edad, si tenían pareja o no y su preferencia por hombres o mujeres en el contexto de las citas. Además, en esta ocasión también se les preguntó si habían usado anteriormente plataformas de búsqueda de pareja (“¿Has utilizado alguna vez páginas web o aplicaciones de citas (como Tinder, Meeting...)?” / “Have you ever used dating websites or applications (such as Tinder, Meeting...)?”). A continuación, según el grupo experimental, en la Fase 0 los participantes leyeron unas instrucciones en las que se describían cuáles eran las habilidades de los algoritmos actualmente. En el grupo subjetivo, se explicó que los algoritmos de IA hoy son capaces de desempeñar tareas subjetivas, basadas en la intuición, tales como componer música, predecir coincidencias entre candidatos de citas o escribir poesía. Por contra, a los participantes del grupo objetivo se les recordaron las tareas objetivas que los algoritmos son capaces de realizar, como predecir o comprar valores bursátiles, ofrecer direcciones o diagnosticar

enfermedades. Al igual que en el Experimento 6, en ambos grupos introdujimos una pregunta sobre la confiabilidad en la recomendación algorítmica en el contexto de citas (“¿Hasta qué punto considera fiables los algoritmos de inteligencia artificial para recomendar candidatos compatibles en los sitios web de citas?” / “How trustworthy do you consider artificial intelligence algorithms for recommending compatible candidates on dating websites?”). Esta pregunta se presentó tanto en la Fase 0 como al final del experimento para conocer si se producía un cambio en el juicio de los participantes. Como final de la Fase 0, los participantes rellenaron el test de personalidad ficticio como en experimentos anteriores, aunque en esta ocasión, no se les pre-expuso a los candidatos de relleno, puesto que eliminarlos no había impactado en la influencia del algoritmo en el Experimento 6. Tampoco se mostró a los participantes de ambos grupos el informe supuestamente personalizado del algoritmo, ausente también en el grupo noforer del Experimento 6. El objetivo era evitar que este informe, al estar redactado de forma ambigua y subjetiva, pudiera entrar en contradicción con la manipulación de objetividad de los algoritmos.

A continuación, en la Fase 1, facilitamos a los participantes unas instrucciones sobre los atributos objetivos o subjetivos de la tarea de recomendar un candidato de citas. Nótese que estas instrucciones sobre la tarea eran diferentes (y complementarias) a las instrucciones sobre las habilidades de los algoritmos proporcionadas en la Fase 0. En el grupo subjetivo, se indicaba que la recomendación de una pareja romántica era una tarea subjetiva, basada en la intuición, mientras que en el grupo objetivo se puso foco en la necesidad de basarse en los datos y las reglas para recomendar candidatos con éxito.

No hubo cambios en la Fase 2 de puntuación. Como en anteriores experimentos, se puntuaron ocho candidatos en una escala del 1 al 9, cuatro de ellos candidatos diana, señalados con el distintivo de "+90% compatibilidad", y cuatro de ellos candidatos control. Cada fotografía fue visualizada durante 2 segundos y presentada en orden aleatorio para cada participante. Las imágenes fueron contrabalanceadas en su papel como candidatos recomendados o candidatos control, evitándose más de dos repeticiones seguidas del mismo candidato (véase Tabla 8).

Tabla 8

Contrabalanceos de los Estímulos Diana y Control en el Experimento 7

	D1	D2	D3	D4	C1	C2	C3	C4
Balanceo 1	A	B	C	D	E	F	G	H
Balanceo 2	E	F	G	H	A	B	C	D
Balanceo 3	A	D	F	G	E	H	B	C
Balanceo 4	E	H	B	C	A	D	F	G

Nota. D = Estímulo Diana, recomendado; C = Estímulo Control, no recomendado; A-H = Fotografías de los candidatos.

Antes de finalizar el experimento, los participantes contestaron de nuevo a la pregunta sobre la confiabilidad en los algoritmos en el contexto de citas y unas preguntas adicionales sobre su confianza en los algoritmos en la toma de decisiones en general ("¿Hasta qué punto considera confiable el consejo que pueden ofrecer los algoritmos de inteligencia artificial, en general, en cualquier otro ámbito de decisión?" / "To what point do you consider trustworthy the advice that artificial intelligence algorithms can offer, in general, in any other decision-making matter?"); el nivel de

cumplimiento de sus expectativas tras interactuar con nuestro algoritmo (“Teniendo en cuenta sus expectativas al principio del experimento, ¿cómo realizó nuestro algoritmo la tarea?” / “Considering your expectations at the beginning of the experiment, how well did our algorithm perform the task?”); y su juicio sobre la eficacia de nuestro algoritmo (“¿Cómo ha sido de eficaz nuestro algoritmo a la hora de recomendar candidatos?” / “How effective has our algorithm been in recommending compatible candidates?”). Para terminar, facilitamos a los participantes las preguntas utilizadas en los experimentos previos sobre la visualización del distintivo de recomendación, y sobre si consideraban que el distintivo les había influido en sus puntuaciones, además del acceso al pago y a la explicación del objetivo real del experimento.

Resultados y Discusión

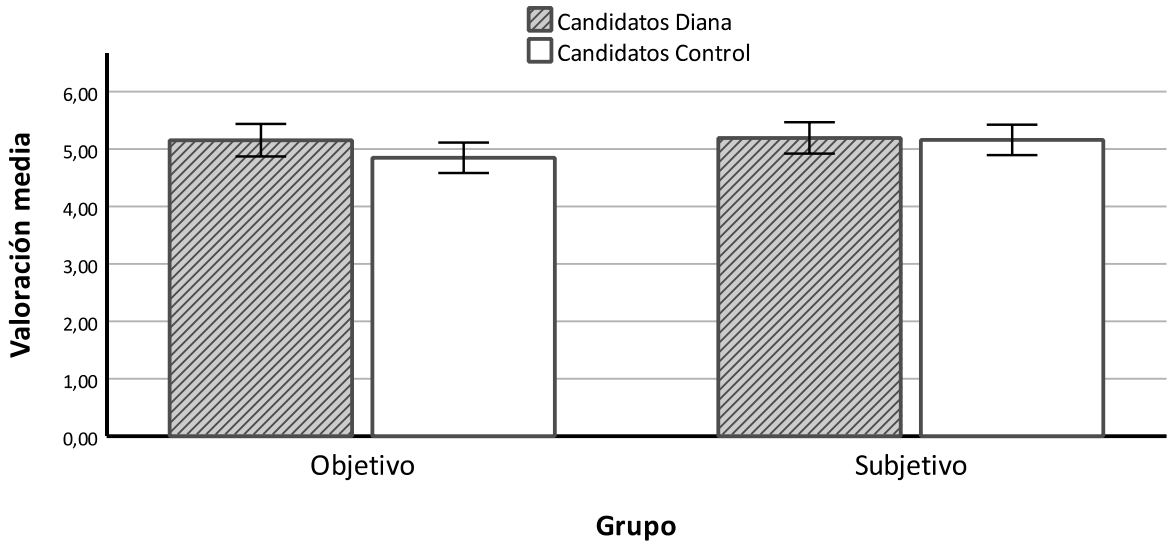
Realizamos un ANOVA mixto[8] con el candidato como factor intra-sujetos (diana *vs.* control) y el grupo como factor inter-sujetos (subjetivo *vs*. objetivo) sobre la valoración de los candidatos. Aunque no encontramos un efecto principal del grupo ($F(1, 238) = 0.98$, $p = .323$, $\eta^2_p = 0.004$) ni interacción Candidato x Grupo ($F(1, 238) = 2.81$, $p = .095$, $\eta^2_p = 0.012$), sí se producía un efecto principal del candidato ($F(1, 238) = 4.38$, $p = .037$, $\eta^2_p = 0.018$). Como en el Experimento 3, consideramos necesario realizar las comparaciones a posteriori a pesar de la falta de interacción por si esta había sido atenuada por la presencia de controles tanto en el candidato como en el grupo. Así, para comprender mejor los resultados obtenidos, realizamos las

[8] Realizamos, además, un análisis no-paramétrico como en los anteriores experimentos. Los resultados del test Wilcoxon para las medias de candidatos diana y control por grupo fueron: En grupo objetivo ($Z = 2.550$, $p = .011$); En grupo subjetivo ($Z = 0.196$, $p = .844$).

comparaciones a posteriori mediante la prueba de Tukey y observamos que la diferencia entre los candidatos recomendados y los controles no se producía en el grupo subjetivo como esperábamos (M_{Diana} = 5.19, $M_{Control}$ = 5.16, t(238) = 0.29, p = .099, d = 0.02), sino en el grupo objetivo (M_{Diana} = 5.15, $M_{Control}$ = 4.85, (t(238) = 2.68, p = .039, d = 0.20; véase Figura 13). Es decir, la influencia del algoritmo tenía lugar cuando describíamos la tarea objetiva pero no cuando la describíamos como subjetiva.

Figura 13

Valoración Media de los Candidatos Diana y Control por Grupo



Nota. Las barras de error representan un 95% de CI.

Además, encontramos que la confianza en los algoritmos, para la tarea de citas, que los participantes indicaron, al inicio y al final del experimento, aumentó significativamente (M_{Inicio} = 4.75, M_{Final} = 5.09, t(239) = -2.86, p = .002, d = -0.19). Esta diferencia fue significativa en el grupo subjetivo ($M_{Subjetivo\ Inicio}$ = 4.78, $M_{Subjetivo\ Final}$ = 5.24, t(118) = -2.62, p = .005, d = -0.24), pero no en el objetivo ($M_{Objetivo\ Inicio}$ = 4.72,

$M_{Objetivo\ Final}$ = 4.93, t(120) = -1.36, p = .088, d = -0.12), lo cual resulta curioso puesto que el grupo subjetivo era en el que no se había logrado la influencia. La confianza al inicio para ambos grupos fue prácticamente la misma (p = .994).

No hubo tampoco diferencias entre grupos en el resto de medidas de autoinforme. Ambos grupos declararon niveles similares de efectividad por parte del algoritmo ($M_{Objetivo}$ = 5.46, $M_{Subjetivo}$ = 5.61, t(238) = -0.66, p = .513), de expectativa cumplida ($M_{Objetivo}$ = 5.74, $M_{Subjetivo}$ = 6.04, t(238) = -1.32, p = .187) y de confianza en los algoritmos en general ($M_{Objetivo}$ = 5.16, $M_{Subjetivo}$ = 5.51, t(238) = -1.64, p = .102) .

A pesar de que en este experimento, al igual que en el Experimento 6, eliminamos la pre-exposición a los candidatos de relleno y no se mostró el falso informe de personalidad para evitar que este pudiera entrar en conflicto con el enmarcado objetivo de la tarea en uno de los grupos, no podemos descartar que estos cambios hayan podido afectar a los resultados, dado que su eliminación en el Experimento 6 había sido en el contexto político y no en el de citas.

Aun así, parece que estos resultados apuntan a que la percepción sobre la objetividad de una tarea puede moldearse, como ya señalaban Castelo y colaboradores (2019, p. 10). Reformular una tarea subjetiva como la de las citas para describirla desde un prisma de atributos objetivos parece haber favorecido la aceptación de la recomendación explícita del algoritmo en un contexto donde hasta ese momento solo había funcionado la influencia encubierta. Consideramos que este hallazgo es relevante dado el considerable impacto a nivel social que la objetividad percibida de una tarea puede causar fuera del laboratorio. Como respuesta a la expansión de la IA, o quizá como factor de su expansión precisamente, contextos de

decisión hasta ahora considerados complejos y subjetivos comienzan a presentarse ante los ciudadanos como tareas objetivas, perfectos para la toma de decisión algorítmica. Sería el caso de la IA utilizada para inferir rasgos de personalidad o predecir futuros comportamientos en base al análisis visual de posturas o rasgos faciales de las personas. Un uso que además no cuenta con una base científica que respalde la validez de tales correlaciones (Barrett y cols., 2019) y que recuerda a investigaciones pseudocientíficas ya superadas como la eugenesia o la frenología, las cuales afirmaban que ciertos comportamientos criminales podían detectarse a partir de mediciones "objetivas" de los rasgos faciales. La preocupación por este posible uso problemático de los algoritmos de reconocimiento facial ha sido plasmada recientemente en una carta al Gobierno de España por un grupo de más de 70 académicos españoles (Agudo y cols., 2021). Nuestros resultados indican que presentar estas tareas como objetivas puede hacer a los ciudadanos más vulnerables a la recomendación algorítmica.

Esta objetivación de las tareas complejas para ser analizadas y gestionadas por modelos de IA implica en ocasiones la utilización de variables *proxys* excesivamente simplistas, que carecen de una visión completa del problema de decisión, o que incluso pueden llegar a estar sesgadas. Los proxys son variables que se correlacionan con el valor inferido y se utilizan como variable principal para realizar predicciones. Por ejemplo, los costes previos en atención médica pueden ser utilizados por un algoritmo de IA como proxy para predecir qué personas necesitan de atención médica adicional. Es el caso sobre el que Obermeyer y colaboradores alertaron en su trabajo de 2019. Los autores encontraron que el uso de esta variable como proxy, por parte de un

algoritmo de un gran hospital universitario de EE.UU., provocaba que las personas de raza negra se vieran perjudicadas al perpetuarse la situación históricamente discriminatoria de emplear menores cantidades de dinero en el cuidado de estas personas.

La tendencia a considerar muchas tareas de decisión como objetivables (sin que necesariamente lo sean) podría estar incluso favoreciendo la delegación de la toma de decisiones en los algoritmos en cuestiones de índole personal, institucional y social. Un ejemplo de ello podría ser la decisión del Gobierno japonés de invertir 19 millones de dólares para impulsar la tasa de natalidad del país a través de servicios de búsqueda de parejas con inteligencia artificial (Quach, 2020).

Aunque según Castelo y colaboradores (2019), tareas como recomendar música, una película o una pareja romántica son consideradas por sus participantes como tareas poco objetivas (22, 23 y 26 puntos sobre 100 respectivamente), el uso de algoritmos de recomendación como Spotify, Netflix o Tinder va en aumento, lo que podría cambiar la percepción de la subjetividad de estas tareas a corto o medio plazo. Longoni y Cian (2020), por ejemplo, muestran en un experimento cómo la valoración de un producto (en su caso, un pastel) puede contemplarse desde diferentes criterios dependiendo de si es un humano o una IA quien selecciona los ingredientes de la receta para su elaboración. En su trabajo, cuando los ingredientes habían sido elegidos por una IA, el pastel recibía puntuaciones más altas en atributos utilitarios (por ejemplo, propiedades alimentarias beneficiosas) pero más bajas en atributos hedónicos (por ejemplo, deleite para los sentidos), al contrario de lo que sucedía cuando el que había seleccionado los ingredientes de la receta era un humano.

Quizá en un tiempo, con los algoritmos de IA presentes en multitud de contextos de decisión, la distinción entre tareas objetivas y subjetivas se reduzca al mínimo, derivando en una general aceptación de la recomendación algorítmica explícita. Sin embargo, y a pesar de que la objetivación de tareas puede ser un tema para abordar en futuras investigaciones, nuestro verdadero interés con esta nueva serie experimental era comprender cuándo y por qué no se aceptaba la recomendación explícita en contextos de decisión subjetivos. Los datos del presente experimento siguen mostrando la falta de aceptación a la recomendación explícita en citas cuando la tarea se presenta como subjetiva, a pesar de nuestro intento por enfatizar las habilidades subjetivas del algoritmo. Por ello, abordamos dos nuevos experimentos con un enfoque muy diferente a los anteriores y un nuevo contexto: el campo del arte.

Experimento 8. Emoción y sensibilidad algorítmica en contexto de arte

La expansión y desarrollo de la IA en los últimos años ha supuesto su desembarco en dominios que hasta ahora se consideraban terreno exclusivo para seres con habilidades propiamente humanas (Wegner y Gray, 2017), tales como escribir novelas (Jozuka, 2016), pintar cuadros (Christie's, 2018), idear trucos de magia (Williams y McOwan, 2014) o componer música (Adams, 2010; Deah, 2018).

En el caso de la música, aunque la contribución de la IA ha sido amplia en volumen, su trabajo no ha sido muy bien recibido por crítica y público. Muestra de ello es el caso de David Cope, un profesor de la Universidad de California que lleva más de dos décadas generando composiciones musicales con IA. Sus primeras muestras al público, una pieza musical similar a las de Bach en un concurso de la Universidad de Oregón y otra pieza con el estilo de Mozart en el Festival Barroco de Santa Cruz (G. Johnson, 1997) fueron recibidas con rechazo, desprecio e incluso ira (Friedel, 2018). Cope no pudo lograr que músicos reconocidos interpretaran sus composiciones públicamente ni siquiera años después (Saenz, 2009). La crítica no fue más benévola y calificó su obra de mera imitación, carente de sentido y alma (G. Johnson, 1997).

Desde entonces, los numerosos avances tecnológicos actuales no parecen haber cambiado la percepción de la capacidad artística de los algoritmos, al menos en el contexto de la música clásica. Las reacciones insatisfechas del público y las críticas negativas recibidas por su reciente conclusión de las sinfonías inacabadas de Mahler (Zappei, 2019) y Schubert (Mantilla, 2019) confirman este rechazo.

Estos hechos nos llevan a pensar que quizá la aversión al algoritmo también se produzca en el terreno de las artes. O al menos los pocos estudios existentes relacionados con el tema parecen apuntar en esta dirección. Por ejemplo, Ragot y colaboradores (2020) encontraron que los participantes en su estudio otorgaron mejores puntuaciones a los cuadros creados por humanos, en términos de gusto, belleza, novedad y significado, que a los creados por una IA.

Asimismo, en un estudio reciente, Hong y colaboradores (2020) pidieron a sus participantes que evaluaran piezas musicales compuestas por una IA o por un humano en términos de atractivo estético, creatividad y destreza, además de evaluar en qué medida se habían violado las expectativas previas de los participantes y cuál era su actitud hacia la IA creativa. A pesar del diseño de este experimento, los autores no abordaron si los participantes valoraban diferente los trabajos de artistas artificiales y los de artistas humanos. No obstante, sí concluyeron que la aceptación de las habilidades creativas de los algoritmos era un requisito necesario para una evaluación positiva de su rendimiento artístico.

Aunque, como hemos indicado, no hay muchos estudios que midan la valoración de la IA como artista, sí que algunos trabajos aportan hallazgos interesantes en este campo utilizando como método una forma modificada del Test de Turing, es decir, mostrando a los participantes obras artísticas de algoritmos y de humanos y pidiéndoles que adivinen su autoría. Por ejemplo, Moffat y Kelly (2006) presentaron algunas piezas musicales a una pequeña muestra de participantes, solicitándoles que las evaluaran e intentaran adivinar si estas habían sido compuestas por humanos o por ordenadores. Independientemente del género escuchado, los participantes preferían

las obras que creían que habían sido compuestas por humanos. De forma similar, Chamberlain y colaboradores (2018) mostraron a sus participantes varias piezas de arte visual creadas por humanos o por ordenadores para que adivinaran su autoría y las evaluaran. Cuando a los participantes les gustaban las obras de arte, asumían que el artista era humano.

Estos estudios sugieren que quizá la preferencia por las obras de arte creadas por humanos no se encuentre tanto en la calidad objetiva de la obra sino en los prejuicios que la gente tiene hacia la música creada por máquinas. Sin embargo, dado que en todos estos estudios las obras de arte de IA y humanas mostradas a los participantes eran diferentes, no es posible afirmar si la evaluación se debía a la obra de arte en sí o a la autoría de esta.

Como contrapunto, el trabajo de algunos compositores de IA ha recibido mejores críticas. Es el caso de la banda de death metal Dadabots, compuesta por una red neuronal artificial, que suma un total de 10 discos en el mercado (Merino, 2019).

Como señalan Chamberlain y colaboradores (2018), no existen suficientes investigaciones que aborden en términos psicológicos la interacción persona-algoritmo y su relación con el arte a pesar de que la IA se está convirtiendo en un actor habitual en este campo. Por todo ello, abordamos un nuevo experimento para comprobar si las personas atribuirían sensibilidad y capacidad de inducir emociones a las IAs, en comparación con los humanos, al contemplar una obra de arte musical y visual. A diferencia de estudios anteriores en los que los participantes compararon obras de arte humanas con obras de arte realizadas por IA, todos nuestros participantes fueron expuestos únicamente a obras de arte creadas por una IA y la manipulación

experimental consistía en confesar a algunos de ellos que el artista era una IA, mientras que a otros se les decía que la autoría era humana.

Nuestro objetivo con el presente experimento era determinar si la experiencia ante una misma obra de arte se juzgaría de forma diferente dependiendo de la autoría de esta. Nuestra hipótesis era que los participantes atribuirían una menor capacidad a la IA, respecto a los artistas humanos, para interpretar con sensibilidad una obra de arte y para evocar emociones en el público.

Método

Participantes y Materiales

Reclutamos una muestra de 249 participantes (55% mujeres, 8% no reportado), mediante el procedimiento de bola de nieve, utilizando un mensaje de WhatsApp enviado a varios grupos de contactos en España, los cuales también contribuyeron a su difusión. Este mensaje consistía en una invitación para participar en un experimento sobre "música y sentimientos" e incluía un enlace al experimento online.

Los participantes fueron asignados aleatoriamente a uno de los dos grupos: artista IA (*n* = 115) o artistas humanos (*n* = 134) y todos ellos vieron el mismo vídeo[9] en el que una IA improvisaba una melodía al piano mientras pintaba sobre un lienzo siguiendo el ritmo de la música en el que no se ve al autor de la obra (ni IA ni humanos).

Diseño y Procedimiento

El diseño de este experimento puede revisarse resumido en la Tabla 9.

[9] El vídeo muestra la instalación artística digital, creada con inteligencia artificial, "Water Color Melody Machine", de Karlos G. Liberal: https://player.vimeo.com/video/325421701.

Tabla 9

Resumen del Diseño del Experimento 8

Grupo	Instrucciones	Tarea	Preguntas
Artista IA	El artista es una IA	Ver el vídeo	Emoción y Sensibilidad
Artistas Humanos	Los artistas son humanos		

Tras aceptar el consentimiento informado online, los participantes leyeron unas instrucciones diferentes en cada grupo como introducción al vídeo. Estas instrucciones fueron nuestra manipulación experimental. Al grupo artista IA se le confesó que el artista era una IA ("Te presentamos a WCMM, una Inteligencia Artificial que improvisa al piano mientras pinta sobre el lienzo"). Utilizamos el término IA (y no el de algoritmo) por ser el término más similar a los utilizado en los trabajos de la literatura mencionados.

Por su parte, al otro grupo se les ocultó la verdadera autoría de la obra y se atribuyó la autoría a artistas humanos. Para controlar el género de los artistas humanos, a la mitad de los participantes se les dijo que el compositor y el pintor eran hombres ("Te presentamos a Javier Aldaz y Miguel Beltrán, dos artistas que improvisan al piano mientras pintan sobre el lienzo"), y a la otra mitad se le dijo que las artistas eran mujeres ("Te presentamos a Ana Aldaz y María Beltrán, dos artistas que improvisan al piano mientras pintan sobre lienzo").

Escogimos utilizar un formato audiovisual como obra a valorar, donde se combinaba composición musical y visual, para mostrar la abrumadora capacidad

creadora de la IA hoy en día. Consideramos que un vídeo donde una IA improvisa al piano mientras pinta sobre un lienzo mostraría adecuadamente su actual potencial.

Tras ver el vídeo se preguntó a todos los participantes sobre la actuación de los artistas. Los experimentos existentes hasta el momento se han centrado en la evaluación de la calidad de la obra generada por la IA (Hong y cols., 2020) o en la experiencia del público (es decir, si les gustó la actuación; por ejemplo, Moffat y Kelly, 2006). Por lo tanto, decidimos extender esos hallazgos averiguando si se valoraba de diferente manera la capacidad para provocar emoción y la sensibilidad del artista cuando un público generalista y no experto pensaba que la obra había sido creada por una IA o por humanos. Para ello, utilizamos dos sencillas preguntas. Por un lado, los participantes debían puntuar la emoción que la obra les había evocado ("Ahora que has visto el vídeo de esta Inteligencia Artificial / de estas artistas / de estos artistas, ¿hasta qué punto dirías que te ha emocionado?"). Y por otro, debían puntuar qué sensibilidad atribuían al artista ("¿Y cómo calificarías su sensibilidad?"). Como nos interesaba el juicio subjetivo de un público no experto, no especificamos cómo los participantes debían entender los términos de emoción y sensibilidad. Las respuestas se recogieron utilizando una escala Likert del 0 a 10 y las medias de estas dos preguntas fueron las variables dependientes de nuestro experimento.

Resultados y Discusión

Tras asegurarnos de que no existían diferencias entre los participantes que creían que los artistas humanos eran mujeres u hombres, ni en la emoción inducida ($M_{Hombres}$ = 4.19, $SD_{Hombres}$ = 2.66, $M_{Mujeres}$ = 3.99, $SD_{Mujeres}$ = 2.81, $t(132) = 0.44$, $p = .659$,

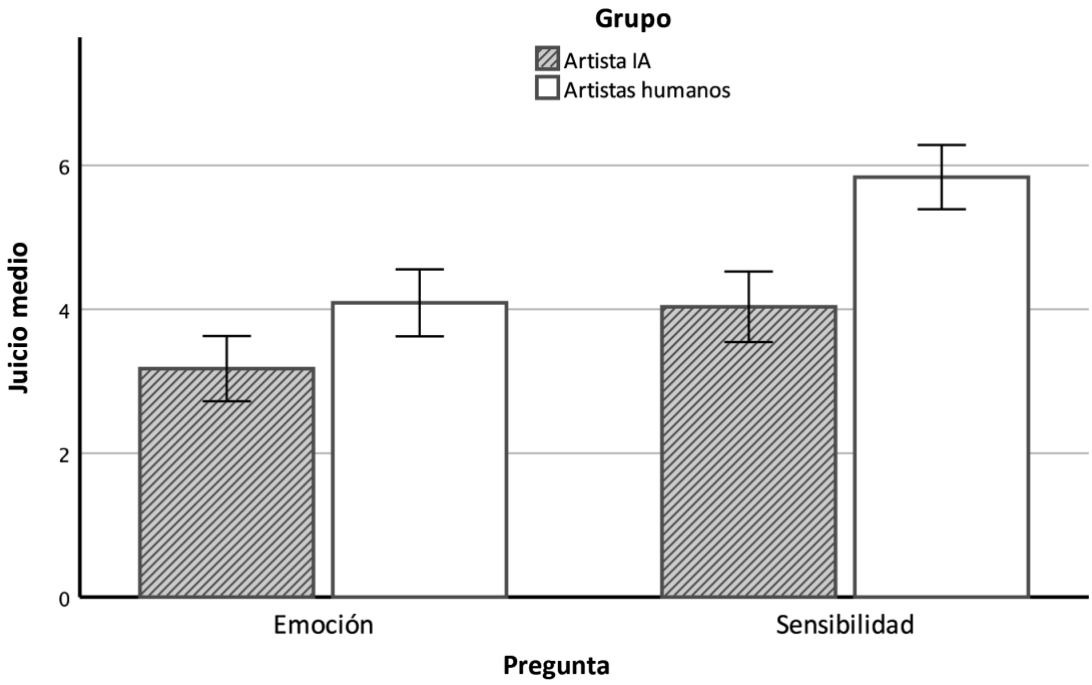
d = 0.08), ni en la sensibilidad atribuida ($M_{Hombres}$ = 5.84, $SD_{Hombres}$ = 2.44, $M_{Mujeres}$ = 5.84, $SD_{Mujeres}$ = 2.79, *t*(132) = 0.00, *p* = 1.00, *d* = 0.00), agrupamos los datos de hombres y mujeres artistas en el grupo de artistas humanos.

En consonancia con nuestra hipótesis, tal y como se muestra en la Figura 14, los participantes informaron de una mayor emoción cuando pensaban que el artista era humano que cuando pensaban que era una IA. Esto se confirmó mediante una prueba *T-Student*[10] para muestras independientes ($M_{Artistas\ Humanos}$ = 4.09, $SD_{Artistas\ Humanos}$ = 2.73; $M_{Artista\ IA}$ = 3.17, $SD_{Artista\ IA}$ = 2.45; *t*(247) = 2.76, *p* = .003, *d* = 0.35). Además, como también puede apreciarse en la misma figura, los participantes atribuyeron también una mayor sensibilidad al artista cuando pensaban que era humano ($M_{Artistas\ Humanos}$ = 5.84, $SD_{Artistas\ Humanos}$ = 2.61; $M_{Artista\ IA}$ = 4.03, $SD_{Artista\ IA}$ = 2.66; *t*(247) = 5.38, *p* < .001, *d* = 0.68), lo que también coincide con nuestras predicciones.

[10] Indicamos también los resultados con la prueba *T* de *Welch* dado que en este experimento las pruebas de normalidad indicaban violación de los supuestos. Sin embargo, gracias a que la muestra era lo suficientemente grande, las diferencias en los resultados son mínimas. Emoción (*t*(246) = 2.79, *p* = .003, *d* = 0.35). Sensibilidad (*t*(240) = 5.37, *p* < .001, *d* = 0.68).

Figura 14

Emoción y Sensibilidad de los Artistas Reportadas por Grupo



Nota. Las barras de error representan un 95% de CI.

Nuestros resultados muestran que saber que la IA había sido la autora de una obra de arte audiovisual parece reducir la valoración de la experiencia y del artista. Estos resultados replican y amplían los encontrados por la literatura anterior sobre la diferente apreciación del arte creado por humanos o por algoritmos.

Es importante destacar que, dado que en nuestro experimento la obra valorada era la misma en ambos grupos, los resultados muestran que los juicios de los participantes no se deben a obra de arte en sí, sino a sus prejuicios previos sobre las capacidades del artista.

Experimento 9. Creatividad del algoritmo en contexto de arte

Abordamos un nuevo experimento con el objetivo de replicar los resultados del experimento anterior y recoger además algunas variables utilizadas en investigaciones previas que facilitaran la comparación de resultados y ampliaran el alcance de nuestra investigación.

No son muchos los estudios sobre cómo las personas juzgan el arte realizado por la IA y, como hemos mencionado, la mayoría de los estudios existentes se centran en comprobar si las máquinas pueden exhibir un comportamiento como compositoras que sea indistinguible del de los humanos, con una prueba similar al test de Turing pero en el contexto del arte (Yang y cols., 2017). Aunque otros estudios se enfocan en evaluar si se producen diferencias en la calidad de las composiciones musicales creadas por diferentes modelos informáticos respecto a las de los humanos (Chu y cols., 2017; Pearce y Wiggins, 2007), son pocos los trabajos que evalúan la experiencia de las personas con el arte creado por la IA, como es el caso de nuestro experimento previo, y los que existen difieren considerablemente en cuanto a los objetivos y los métodos utilizados (véase la Tabla 10).

Tabla 10

Principales estudios que evalúan la obra artística de la IA

Artículo	Obra	Autoría revelada	Género	Variables dependientes
Hong y cols. (2020)	**Diferente** en cada grupo: obras de arte humanas o de IA	Antes del juicio	Música	Escala de 9 ítems para valorar la calidad musical (Hickey, 1999). Actitudes hacia la IA creativa. Escala de violación de expectativas (Burgoon, 2015).
Ragot y cols. (2020)	**Diferente** en cada grupo: obras de arte humanas o de IA	Antes del juicio	Pintura	Tres juicios relacionados con la calidad (belleza percibida, novedad y significado). Gusto por la obra de arte.
Hong y Curran (2019)	**Diferente** en cada grupo: obras de arte humanas o de IA	Antes del juicio	Pintura	Ocho juicios de calidad recopilados de diferentes estudios, como la originalidad, el grado de expresividad o el valor estético. Actitudes hacia la IA creativa (ítem binario).
Chamberlain y cols. (2018)	**Diferente** en cada grupo: obras de arte humanas o de IA	Tras el juicio. Test de Turing	Pintura	Atractivo de la obra de arte.
Moffat y Kelly (2006)	**Diferente** en cada grupo: obras de arte humanas o de IA	Tras el juicio. Test de Turing	Música	Gusto por la obra de arte. Disfrute de la obra de arte.
Experimento 8 de esta tesis	**Idéntica** para todos los grupos: obra de arte de IA	Antes del juicio	Música y Pintura	Emoción experimentada con la obra de arte. Sensibilidad del artista.

Experimento 9 de esta tesis	**Idéntica** para todos los grupos: obra de arte de IA	Antes del juicio	Música	Emoción experimentada con la obra de arte. Sensibilidad del artista. Subescala de emociones estéticas prototípicas (Schindler y cols., 2017), que incluye medida de gusto (Moffat y Kelly, 2006; Ragot, Martin y Cojean, 2020). Juicio sobre el nivel de calidad de la obra. Actitudes hacia la IA creativa (Hong, Peng y Williams, 2020).

Tal y como puede observarse en la Tabla 10, mientras algunos estudios revelan la autoría artística de la IA antes de que los participantes juzguen la obra de arte (Hong y cols., 2020; Hong y Curran, 2019; Ragot y cols., 2020), como ocurre en nuestros experimentos, otros informan de ella después de recoger los juicios de los participantes con un procedimiento similar al que utiliza el mencionado test de Turing (Chamberlain y cols., 2018; Moffat y Kelly, 2006). Además, no existe uniformidad en las variables recogidas en estos estudios. Algunos estudios utilizan juicios sobre la calidad de la obra como variable dependiente principal del experimento, a través de medidas definidas por los propios autores (por ejemplo, originalidad o valor estético de Hong y Curran, 2019; o belleza percibida y significado de Ragot y colaboradores, 2020). Sin embargo, en el caso de Hong y colaboradores (2020), los autores usaron una escala validada de 9 ítems, basada en la escala original de 18 ítems de Hickey (1999). Se trata de una escala para facilitar a los profesores de música la evaluación de las composiciones de sus alumnos, por lo que los ítems requieren un cierto conocimiento

profesional de la materia y son muy diferentes a los utilizados por otros investigadores para la misma variable. Mientras tanto, otros estudios se centran en evaluar la experiencia de los participantes más que la calidad de la obra en sí, con medidas más subjetivas, como el atractivo (Chamberlain y cols., 2018), el gusto (Moffat y Kelly, 2006; Ragot y cols., 2020) o el disfrute de la obra (Moffat y Kelly, 2006). Esto último sería un planteamiento similar al seguido en nuestros experimentos, en los que evaluamos la emoción experimentada y la sensibilidad atribuida al artista.

También hay diferencias entre estudios en el tipo de arte evaluado. Hong y colaboradores (2020) y Moffat y Kelly (2006) recogieron los juicios de los participantes sobre la música generada por la IA, mientras que Chamberlain y colaboradores (2018), Hong y Curran (2019) y Ragot y colaboradores (2020) registraron los juicios sobre el desempeño de la IA en pintura.

En lo que sí coinciden ampliamente los trabajos mencionados en la Tabla 10 es en que todos ellos muestran obras de arte creadas por IAs a un grupo de participantes, mientras que el otro grupo evalúa obras de arte creadas por artistas humanos. Sin embargo, creemos que este diseño no permite a los investigadores conocer si son los prejuicios sobre la autoría de la obra de arte los que provocan las diferencias en los juicios o es la obra de arte la que es cualitativamente diferente. Por este motivo, en nuestro experimento previo ambos grupos evaluaron la obra de arte de una IA. Este fue el procedimiento que mantuvimos en nuestro siguiente experimento, incorporando además algunas de las medidas mencionadas anteriormente con el fin de facilitar las comparaciones entre estudios y consolidar los resultados obtenidos.

En el nuevo experimento decidimos también simplificar el estímulo artístico. En lugar de utilizar la obra de arte de vídeo que combinaba música y pintura, presentamos a los participantes una obra de arte puramente musical. Además, añadimos una fase final al final del experimento para evaluar si los participantes mantenían sus juicios al comentarles que, a diferencia de lo que se les había dicho inicialmente, en realidad el autor de la obra era humano (o una IA, según el grupo). Al igual que en el experimento anterior, nuestra hipótesis era que las medidas de emoción y sensibilidad recibirían puntuaciones más bajas cuando los participantes supieran que el artista era una IA que cuando se les dijera que era un humano.

Método

Participantes y Materiales

Reclutamos 250 participantes (47.6% mujeres, 0.8% no binario) mayores de 18 años (*M* = 26.6, *SD* = 8.62) a través de la plataforma Prolific Academic. El análisis de sensibilidad para esta muestra, muy similar a la del experimento anterior, mostró que contábamos con una potencia del 80% para detectar efectos pequeños (*d* = 0.31). La muestra se dividió aleatoriamente en dos grupos: artista AI (*n* = 125) y artista humano (*n* = 125). Como en el experimento anterior, la obra de arte mostrada a todos los participantes fue la misma. En esta ocasión se trataba de una pieza puramente musical, compuesta e interpretada por una IA[11]. El experimento, esta vez en lengua

[11]La obra de arte escuchada puede consultarse en http://bit.ly/2Onl4DF. Se trata de una versión abreviada de la pieza original a1_98137.mid, generada por la IA Music Transformer (https://magenta.github.io/listen-to-transformer/#a1_98137.mid)

inglesa, se registró previamente en AsPredicted (https://aspredicted.org/blind.php?x=wi8r8r)[12].

Diseño y Procedimiento

El diseño fue muy similar al del Experimento 8, con una fase extra al final para reevaluar el juicio de los participantes ante el cambio de las instrucciones. El diseño de este experimento puede revisarse resumido en la Tabla 11.

Tabla 11

Resumen del Diseño del Experimento 9

Grupo	Instrucciones	Tarea	Preguntas	Cambio de instrucciones
Artista IA	La pieza es de una IA	Escuchar el audio	Juicios sobre la experiencia y el artista	La pieza era de un humano
Artista Humano	La pieza es de un artista			La pieza era de una IA

Antes de escucharla la pieza musical, los participantes aceptaron el consentimiento informado online y leyeron unas instrucciones diferentes según el grupo. Al grupo artista IA se le confesó que escucharía una pieza musical compuesta e interpretada por una IA, mientras que al grupo artista humano se le dijo que la pieza estaba compuesta e interpretada por un artista (sin especificar su género ni su condición humana). No especificamos en las instrucciones de este grupo que el artista fuera humano para no levantar sospechas sobre el propósito del experimento en los

[12] El enlace permite acceder a una versión anónima del pre-registro preparada para la revisión por pares, dado que este experimento forma parte de un artículo sin publicar aún.

participantes como advierten Hong y Curran (2019) y porque, si no se informa a las personas sobre la posible autoría de una IA, estas piensan por defecto que el artista es humano, como se muestra en Chamberlain y colaboradores (2018).

Para evitar que los participantes continuaran por error con el resto del experimento si el archivo de audio no se cargaba inmediatamente y además garantizar que los participantes escucharan al menos una parte de la pieza musical antes de avanzar, el botón para pasar a la siguiente página no aparecía hasta transcurridos 35 segundos.

Tras escuchar la obra de arte, se preguntó a todos los participantes sobre la emoción que la música les había provocado y la sensibilidad que atribuían al artista, utilizando para ello las mismas preguntas del experimento previo. Además, añadimos otra medida sobre la emoción utilizando la subescala de emociones estéticas prototípicas de Schindler y colaboradores (2017), que en su trabajo reporta una alta consistencia interna ($\alpha = .77$). Se trata de una escala de 5 puntos que evalúa la intensidad con la que se siente una emoción e incluye ítems como la fascinación, el asombro o la medida de gusto utilizada en Moffat y Kelly (2006) y Ragot, Martin y Cojean (2020). Esto nos permitía complementar la medida de emoción de un solo ítem usada previamente. Además, pedimos a los participantes que calificaran la calidad de la obra de arte (“¿Cuál era el nivel de calidad de la obra de arte?” / "What was the quality level of the artwork?") en una escala del 0 al 10.

Una vez valorada su experiencia, los participantes indicaron su grado de conformidad con los ítems de Hong y colaboradores (2020) sobre las actitudes hacia la IA creativa, su incomodidad ante la presencia de algoritmos en el arte (“La inteligencia

artificial que puede realizar obras de arte mejor que los humanos me incomoda” / “Artificial intelligence that can perform artworks better than humans makes me uncomfortable”; y “Me siento mal conmigo mismo si consumo arte realizado por la inteligencia artificial” / “I feel bad about myself if I consume art performed by artificial intelligence”) y su juicio sobre lo necesario que era poseer ciertas habilidades humanas para componer música (“Componer música es una tarea que requiere poseer emociones humanas” / “Composing music is a task that requires the possession of human emotions”; y “Componer música es una tarea relacionada con, y una parte muy importante, de lo que significa ser humano” / “Composing music is a task related to, and a very important part, of what it means to be human”).

A continuación, preguntamos a los participantes por su género y edad, su experiencia profesional o formativa con la IA y la tecnología con una pregunta binaria (“¿Su trabajo o estudios están relacionados con la tecnología, la inteligencia artificial, los robots o los algoritmos?” /"Are your work or studies related to technology, artificial intelligence, robots or algorithms?"), y su experiencia con la música con una pregunta similar (“¿Su trabajo o estudios están relacionados con la música?” / "Are your work or studies related to music?"). Además, los participantes indicaron hasta qué punto les gustaba la música clásica (“¿Hasta qué punto diría que le gusta la música clásica?” / “To what extent would you say you like classical music?”), ya que, según Hong y colaboradores (2020), la preferencia por el género musical de la obra escuchada puede condicionar su valoración. Por último, antes de informar a los participantes sobre el propósito real del experimento y facilitar el acceso al pago por su participación, añadimos una nueva fase respecto al experimento previo. En ella, cambiamos la

historia que se había presentado al inicio a los dos grupos. Al grupo artista IA se le dijo que, en realidad, la autoría de la pieza pertenecía a un artista humano, mientras que al grupo artista humano se le dijo que, en realidad, la obra de arte era de una IA. Con esta nueva información solicitamos a los participantes que volvieran a calificar la emoción experimentada con la pieza musical y la sensibilidad atribuida al artista, utilizando las mismas preguntas de un solo ítem utilizadas previamente. El objetivo de esta fase era evaluar si el cambio de las instrucciones podía implicar una diferencia en la valoración subjetiva de su experiencia.

Resultados y Discusión

Los resultados del Experimento 8 se replicaron con respecto a la variable de la sensibilidad del artista. Las pruebas *T-Student*[13] para muestras independientes confirmaron que los participantes atribuían una mayor sensibilidad al artista cuando creían que era un humano (M = 6.90, SD = 1.73) que cuando sabían que era una IA (M = 5.53, SD = 2.39; $t(248) = 5.19$, $p < .001$, $d = 0.66$). Sin embargo, a diferencia del experimento previo, en esta ocasión los participantes no indicaron una menor emoción cuando sabían que la obra de arte había sido compuesta por una IA (M = 5.18, SD = 2.43) que cuando creían que el artista era humano (M = 5.58, SD = 2.02; $t(248) = 1.39$, $p = .083$, $d = 0.18$). Cabe destacar que la media de la emoción reportada por los participantes de ambos grupos aumentó con respecto al experimento anterior ($M_{Artistas\ Humanos}$ = 4.09, $SD_{Artistas\ Humanos}$ = 2.73; $M_{AI\ artista}$ = 3.17, $SD_{Artista\ IA}$ = 2.45 en el

[13] Al igual que en el Experimento 8, indicamos además los resultados con la prueba *T-Welch* porque los supuestos de normalidad y heteroscedasticidad se violaron según las pruebas correspondientes. De nuevo, las diferencias son mínimas. Sensibilidad ($t(226) = 5.19$, $p < .001$, $d = 0.66$). Emoción ($t(240) = 1.39$, $p = .083$, $d = 0.18$).

Experimento 8). Es posible que esta mejor recepción de la obra de arte pudiera estar afectando a las diferencias entre grupos, aunque esto podría deberse a muchos factores diferentes. Las diferencias entre grupos en emoción sí resultaron significativas cuando analizamos las puntuaciones en la subescala de emociones estéticas prototípicas. La escala, que mostró una gran consistencia interna en nuestro experimento (α = 0.91), reveló que se producía una mayor emoción en el grupo artista humano (M = 3.09, SD = 0.85) que en el grupo artista IA (M = 2.90, SD = 0.90; $t(248)$ = 1.72, p = .043, d = 0.22). En esta línea, también encontramos diferencias entre los dos grupos en las valoraciones de calidad. De nuevo, los participantes del grupo artista humano (M = 7.33, SD = 1.66) valoraron con puntuaciones más altas la calidad de la obra de arte que los del grupo de artista IA (M = 6.33, SD = 1.99; $t(248)$ = 4.31, p < .001, d = 0.55).

Dado que, según Hong y colaboradores (2020), aceptar las capacidades creativas de la IA era un requisito necesario para la evaluación positiva de la obra de arte, analizamos si esta medida correlacionaba en nuestro experimento con las variables de emoción, sensibilidad y calidad de la obra de arte. Así, encontramos una correlación positiva entre la aceptación de la creatividad de la IA y la emoción (medida en la pregunta de un ítem, r = 0.27, p < .001; y en la subescala de emociones estéticas prototípicas, r = 0.29, p < .001); entre la aceptación de la creatividad y la sensibilidad atribuida al artista (r = 0.17, p = .004); y entre la aceptación de la creatividad y la valoración de la calidad de la obra de arte (r = 0.21, p < .001).

Además, los participantes no manifestaron una alta incomodidad por la inclusión de la IA en el arte (consistencia interna de los dos ítems de incomodidad, α =

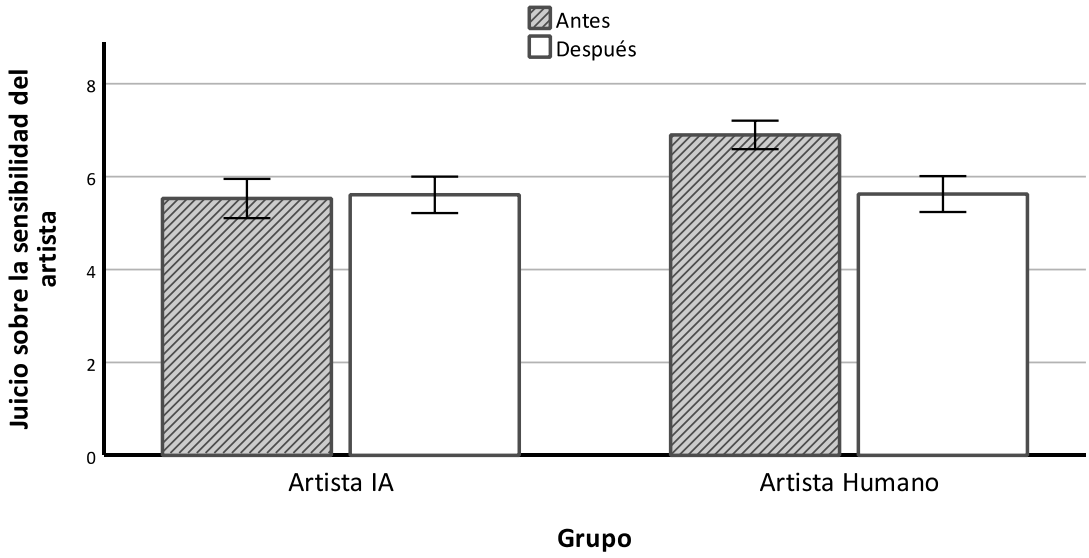
0.70; *M* = 3.45 sobre 10, *SD* = 2.49), aunque sí consideraban que componer música implica poseer habilidades de naturaleza humana como la emoción (consistencia de los dos ítems de habilidades α = 0.75; *M* = 6.52 sobre 10, *SD* = 2.43). Aunque estas dos variables no correlacionaron con la emoción, la sensibilidad o la calidad de la obra de arte (p_s > .05), encontramos una correlación negativa entre la aceptación de la creatividad en las IAs y sentir incomodidad con ellas (*r* = -0.34, *p* < .001) y entre la aceptación de la creatividad y entender el arte como esencialmente humano (*r* = -0.24, *p* < .001). En resumen, cuanto más se les considera creativas a las IAs, menor incomodidad supone que estas desempeñen tal labor, y menos se asocia el arte con habilidades de naturaleza exclusivamente humana.

Por último, analizamos si el cambio de información sobre la autoría de la obra de arte al final del experimento (diciéndole al grupo artista IA que el artista era en realidad humano y al grupo artista humano que el artista era una IA) afectaba a sus puntuaciones de emoción y sensibilidad. Para ello, realizamos 2 ANOVAs mixtos con los juicios de emoción y sensibilidad como variables dependientes, el momento de la medición (es decir, antes y después de que se cambiara la información sobre la autoría) como factores intra-sujetos y el grupo como factor inter-sujetos. Con respecto a la emoción, no encontramos efecto principal del momento de medición ($F(1, 248) = 0.78$, $p = .378$, $\eta^2_p = 0.003$), ni del grupo ($F(1, 248) = 1.75$, $p = .187$, $\eta^2_p = 0.007$), ni interacción Momento de medición x Grupo ($F(1, 248) = 0.09$, $p = .769$, $\eta^2_p = 0.000$). Es decir, la emoción reportada no cambió después de que a los participantes se les dijera que el autor era diferente a lo que creían al inicio del experimento (es decir, humano o IA). Sin embargo, sí encontramos diferencias en la sensibilidad (véase Figura 15), con

un efecto principal del momento de medición ($F(1, 248) = 29.3$, $p < .001$, $\eta^2_p = 0.106$), efecto de grupo ($F(1,248) = 7.82$, p = .006, $\eta^2_p = 0.031$), así como interacción Momento de medición x Grupo ($F(1,248) = 37.7$, $p < .001$, $\eta^2_p = 0.132$). Analizando las comparaciones a posteriori mediante la prueba de Tukey descubrimos que la diferencia se producía en el grupo artista humano. En él, los participantes atribuían más sensibilidad al artista al principio, cuando creían que era humano ($M = 6.90$, $SD = 1.73$) que al final cuando se les decía que el artista en realidad era una IA ($M = 5.62$, $SD = 2.18$; $t(248) = 8.17$, $p = .001$, $d = 0.65$).

Figura 15

Sensibilidad Atribuida a los Artistas Antes y Después de Recibir Información Contradictoria Sobre la Autoría



Nota. Las barras de error representan un 95% de CI. La figura muestra la sensibilidad atribuida al artista en cada grupo al inicio, así como la sensibilidad atribuida después de recibir la información contradictoria (autoría de la IA en el grupo humano y autoría humana en el grupo IA).

Como no contábamos con suficientes expertos en música en la muestra (n = 23), no pudimos analizar si la experiencia influía en sus juicios de emoción o sensibilidad. Lo que sí encontramos fue que, como señalaron Hong y colaboradores (2020), el gusto por el género musical de la obra, música clásica, correlacionaba positivamente con la emoción (recogida en la pregunta de un ítem, r = 0.41, p < .001; y en la subescala de emociones estéticas prototípicas, r = 0.31, p < .001). También esa medida correlacionó con la sensibilidad atribuida al artista (r = 0.25, p < .001) y con la puntuación de calidad reportada sobre la obra (r = 0.15, p = .008).

En resumen, nuestros resultados replican y amplían los del experimento previo. Los participantes atribuyen menor sensibilidad al artista cuando saben que la obra de arte es de una IA que cuando creen que es de un humano. Aunque no replicamos el efecto encontrado sobre la emoción recogida con la pregunta de un solo ítem, sí observamos diferencias en la emoción entre grupos cuando usamos una medida más sensible, la subescala de 8 ítems de emociones estéticas prototípicas. La ausencia de emoción reportada entre grupos en la medida de una pregunta puede estar relacionada con el hecho de haber utilizado una pieza artística diferente (solo musical en este experimento frente a musical y pictórica en el anterior). De hecho, la obra en este experimento obtuvo valoraciones más altas que la obra del experimento anterior, algo que no es de extrañar porque las técnicas de IA en este campo de la composición musical han mejorado sustancialmente en el tiempo transcurrido entre los experimentos (1 año y 10 meses).

El cambio en la sensibilidad atribuida al artista al recibir información contraria sobre la autoría apoya la idea de que los prejuicios sobre las capacidades de la IA en el

arte influyen en la evaluación de su rendimiento, dado que los participantes modificaron la atribución de sensibilidad (aunque no el juicio de emoción) al decirles que la obra de arte pertenecía a un humano y viceversa. Además, según nuestros datos, los participantes que atribuyeron capacidades creativas a las IAs mostraron también una menor incomodidad con su presencia en el arte y consideraron menos necesarias en esta área las habilidades de naturaleza exclusivamente humana.

Por último, nuestros resultados replican los resultados encontrados por Hong y colaboradores (2020) sobre cómo la atribución de capacidades creativas a la IA y el gusto por el género musical evaluado afectan al juicio de la calidad de la obra de arte. En nuestro experimento, ampliamos estos resultados al mostrar que esa atribución de creatividad también afecta a la emoción experimentada y la sensibilidad asociada con el artista. Y lo hacemos además con un diseño experimental donde queda evidente que el juicio de los participantes se debe a la información facilitada sobre la autoría de la obra y no a la obra en sí misma.

Discusión de la Serie Experimental 2

A lo largo de los tres experimentos de este apartado, hemos abordado dos de los factores que se han propuesto desde la literatura de aversión al algoritmo como explicaciones de la falta de aceptación de la recomendación algorítmica, y que están relacionados con la objetividad atribuida a las tareas de decisión y las habilidades atribuidas a los algoritmos.

En el Experimento 7 sobre la objetividad o subjetividad de las tareas de decisión, encontramos que el algoritmo influyó sobre las valoraciones de los participantes en el contexto de citas cuando la tarea y las habilidades del algoritmo se describieron como objetivas, pero no cuando se describieron como subjetivas. Como ya apuntaban Castelo y colaboradores (2019), la percepción de objetividad de las tareas puede moldearse y esta objetividad percibida puede a su vez favorecer la influencia y la eficacia percibida de los algoritmos en dichas tareas. Nuestros resultados parecen apoyar esto. En nuestro experimento, presentar la tarea de citas como objetiva favoreció la influencia de la recomendación explícita del algoritmo en un contexto donde, durante toda la serie experimental 1, solo habíamos logrado influencia encubierta.

Respecto al trabajo de Castelo y colaboradores (2019), que utilizaban una tarea aparentemente objetiva de partida (estimar un valor bursátil), nuestro Experimento 7 añade nueva evidencia al mostrar que también es posible moldear la objetividad percibida de una tarea aparentemente subjetiva de partida (la recomendación de un candidato en el contexto de citas) y que esta presentación de la tarea favorezca la

aceptación de la recomendación algorítmica explícita. Como señalábamos anteriormente, sería necesario seguir investigando hasta qué punto la descripción de objetividad de las tareas puede favorecer la influencia del algoritmo en áreas de decisión complejas que hasta el momento se percibían como subjetivas.

Dado que el Experimento 7 seguía mostrando que la recomendación algorítmica explícita no es aceptada cuando la tarea se interpreta como subjetiva, nuestros Experimentos 8 y 9 se centraron en comprender mejor qué habilidades se atribuyen a la los algoritmos de IA (y, de forma más global, a la IA) en contextos aparentemente subjetivos; en concreto, en el contexto del arte, dado que, como vimos anteriormente, tareas como la composición musical se perciben como tan poco objetivas como la tarea de citas. En dos experimentos en los que recogíamos el grado de emoción vivida y sensibilidad atribuida al artista al experimentar una obra de arte creada por una IA, encontramos que los participantes informaban de una mayor emoción y atribuían una mayor sensibilidad al artista cuando pensaban que era humano que cuando sabían que era una IA. Además, también encontramos que la atribución de capacidades creativas a la IA y el gusto por el género musical también afectaron a la evaluación de la calidad de la obra de arte, así como a la emoción experimentada, a la sensibilidad asociada con el artista, a sentir una menor incomodidad con el arte de la IA y a una menor necesidad de percibir habilidades humanas en el artista.

Nuestros resultados ofrecen pruebas empíricas de que, como ya señaló David Cope (G. Johnson, 1997), el valor de una obra de arte no descansa en la pieza en sí (ya que en nuestro experimento fue la misma para ambos grupos), sino en las

percepciones y atribuciones subjetivas del público y en sus creencias previas sobre las capacidades artísticas de los artistas. Saber que la IA ha sido la generadora de una pieza musical parece reducir la emoción experimentada con ella, así como la valoración de la sensibilidad del artista y de la calidad de la obra. Quizá el conocimiento de la autoría de la IA con respecto a la obra de arte podría haber desencadenado ciertos estereotipos negativos sobre la IA en los participantes, considerando estos que la IA no cuenta con las capacidades necesarias para emocionar y transmitir sensibilidad con su arte. Hong y colaboradores (2020) señalaban en su trabajo que las actitudes previas hacia la IA creativa podrían estar afectando a las valoraciones de la obra de arte y del artista. Nuestros experimentos replican y extienden sus hallazgos al utilizar un diseño experimental donde se demuestra claramente que el efecto encontrado se debe a los prejuicios sobre el artista y no a la obra, dado que la pieza artística mostrada fue la misma para todos los participantes.

En cualquier caso, sería interesante en futuras investigaciones explorar otras posibles razones para esta infravaloración. Castelo y colaboradores (2019) sugirieron, por ejemplo, que la minusvaloración de la IA podría estar provocada también por su falta de presencia cotidiana en determinados entornos, por las creencias previas que la gente pueda tener sobre su rendimiento, o por la inquietud que provoca la idea de que la IA pueda realizar tareas que hasta ahora solo se asociaban a los humanos. Con respecto a este último punto, se han propuesto dos categorías de habilidades humanas que pueden proyectarse (o no) en los algoritmos de IA (Castelo, 2019; Lee, 2018). La investigación sobre deshumanización, por ejemplo, se refiere a ellas como habilidades de singularidad humana y habilidades de naturaleza humana (Loughnan y

Haslam, 2007). Las habilidades de singularidad humana serían atributos que distinguen a los humanos de otros animales pero que se acepta que estén compartidos con las máquinas (normalmente de naturaleza cognitiva, como la racionalidad y la lógica). Las habilidades de naturaleza humana serían atributos que los humanos compartirían con otros animales, pero no con las máquinas (atributos de naturaleza emocional, la intuición o la imaginación). Según Castelo (2019), las personas confiarían en las decisiones algorítmicas para tareas que requieren de habilidades de singularidad humana (como la cognición) en tareas mecánicas y objetivas, mientras que no lo harían para tareas más hedónicas y subjetivas donde se requieren habilidades de naturaleza humana (como la emoción). Por ejemplo, Lee (2018) evaluó el nivel de confianza depositado en gerentes humanos o algoritmos para tomar decisiones que requerían habilidades de singularidad, como estimar los componentes de una máquina que podían funcionar mal en una fábrica, o habilidades de naturaleza humana, como analizar currículos y seleccionar a los mejores candidatos a entrevistar para un puesto de trabajo. El autor encontró que, cuando se necesitaban habilidades de singularidad, los participantes declaraban confiar por igual en las decisiones algorítmicas y en las humanas, valorándolas igual de justas. Sin embargo, los participantes juzgaban las decisiones algorítmicas como más injustas y confiaban menos en ellas cuando se requerían habilidades de naturaleza humana.

En el campo de la música, compositores, críticos e incluso espectadores consideran que ciertas habilidades atribuidas a los humanos (y no a las máquinas), como la creatividad, la sensibilidad, la emoción e incluso el "alma" (Adams, 2010), resultan requisitos esenciales para una ejecución artística de calidad. Por lo tanto,

saber que una obra artística ha sido elaborada por una IA, a la que no se le presuponen estas habilidades exclusivamente humanas, puede conducir a una peor valoración de las composiciones musicales generadas por ella, tal y como se ha observado en nuestros experimentos.

Creemos que esta segunda serie experimental contribuye a entender qué capacidades creativas, de sensibilidad y de emoción atribuyen las personas a la IA cuando valoran su trabajo artístico. Nuestro hallazgo de una preferencia sesgada por la música supuestamente creada por humanos se debe a prejuicios previos sobre las habilidades de la IA y no a la pieza artística en sí misma, dado que en nuestros experimentos era siempre la misma, y, probablemente a que el arte no es un campo en el que la objetividad y la neutralidad esperadas de la IA puedan aportar algún valor positivo o extra a la obra artística respecto a lo que ya aportan los humanos.

Así pues, parece que aún queda camino por recorrer hasta que la presencia activa de la IA en terrenos tan exclusivamente humanos como el arte y la música se convierta en algo familiar y valorado. Quizá el punto de inflexión llegue si algún día la "auténtica humanidad" (Adams, 2010) deja de considerarse un requisito indispensable para la creación artística.

Parte IV. **Discusión General**

Capítulo 6. **Discusión General**

"Ahora mismo los algoritmos te están observando (. …) Y cuando estos algoritmos te conozcan mejor de lo que te conoces tú, lograrán controlarte y manipularte, y tú poco podrás hacer al respecto" (Harari, 2018, p. 302). Es imposible saber cómo de proféticas son estas palabras del historiador Yuval H. Harari, pero, basándonos en los resultados de los experimentos de esta tesis, lo que sí podemos afirmar es que los algoritmos ya son capaces influir en nuestras decisiones. Para ello, podrían estar utilizando diferentes estrategias persuasivas. Por un lado, tal y como muestran nuestros cinco primeros experimentos, la recomendación algorítmica podría estar influyendo en nuestras preferencias de forma explícita o encubierta en diferentes contextos. Aunque la recomendación explícita parece que resultaría más efectiva en el contexto de la política (Experimentos 1 y 5) y la encubierta en citas (Experimentos 3, 4 y 5), en todos los experimentos (salvo en el Experimento 2) el algoritmo influyó de una forma u otra en los participantes, por lo que su capacidad persuasiva se mostró robusta.

En nuestros experimentos las diferentes estrategias de persuasión, explícitas y encubiertas, se han presentado como dos alternativas por separado, aunque no tienen por qué ser excluyentes, pudiendo combinarse para sumar fuerzas. Por ejemplo, en el mencionado trabajo de Banker y Khetani (2019), cuando se solicitaba a los participantes que eligieran entre una selección de productos de carácter utilitario (por ejemplo, cargador de móvil portátil en el primer experimento de su estudio), el algoritmo les recomendaba explícitamente un producto. Pero, además, en alguna de

las condiciones del experimento, a esta recomendación explícita se le añadía una encubierta, utilizando para ello el conocido como efecto señuelo (*decoy effect* o *asymmetric dominance effect*). Debido a este efecto, el producto recomendado dominaba claramente en alguna característica (por ejemplo, la capacidad de carga) a otro de los productos disponibles para su selección (aunque no a todos los demás), por lo que resultaba más atractivo en la comparación. Esta recomendación encubierta, unida a la recomendación explícita del algoritmo, provocaba que el producto fuera escogido en el 77% de las ocasiones, frente al 48% de la condición sin recomendación. Aunque los autores no comprobaron la influencia de la recomendación explícita o encubierta por separado, sus resultados muestran que la combinación de ambos tipos de recomendación puede manipular las preferencias de los participantes que reciben la recomendación respecto a los que no la reciben.

Como señalábamos en capítulos anteriores, existe un número amplísimo de variaciones posibles para ajustar los parámetros del algoritmo y sus recomendaciones de forma que sean efectivas, por lo que nuestros experimentos muestran en realidad la potencialidad de una influencia que se encuentra actualmente en manos de intereses de terceros y en continuo avance y perfeccionamiento. Por ello, entendemos que son necesarios más trabajos que profundicen en la respuesta a la recomendación algorítmica y a los factores que contribuyen a su aceptación o rechazo.

Otra estrategia persuasiva se podría estar utilizando para aumentar la confianza que depositamos en los algoritmos, y así influir en nuestras decisiones, es simular que son capaces de conocer nuestra personalidad (Experimento 6). Nuestra primera serie experimental mostró que ni siquiera era necesaria la existencia de un

algoritmo real en el sistema de recomendación para influir en las preferencias de los participantes, al igual que no era preciso ofrecer un informe real sobre la personalidad de los usuarios para que estos confiaran en el algoritmo. En el Experimento 6 sobre la Fase 0 de confiabilidad, independientemente de si los participantes habían completado o no el test de personalidad y habían recibido o no el supuesto informe por parte del algoritmo, sus valoraciones para los candidatos recomendados, respecto a los candidatos control, fueron más altas. Por lo tanto, no parece que proporcionar excesiva información sobre un algoritmo sea un requisito para que este influya en las preferencias de las personas, al menos en el contexto político que fue el evaluado. Del igual modo que tampoco parece necesario que se ofrezca recomendaciones realmente personalizadas para que los usuarios confíen en el desempeño del algoritmo.

Y la última estrategia persuasiva que se podría estar siguiendo para que las recomendaciones algorítmicas fueran aceptadas es manipulando la percepción de objetividad de las tareas de decisión. Según los resultados encontrados en el Experimento 7, parece que la objetividad percibida de una tarea podría moldearse. Esto permitiría que una tarea a priori subjetiva se presentara como objetiva y, por ello, más adecuada para ser llevada a cabo por los algoritmos, ya que a estos se les presuponen capacidades adecuadas para este tipo de tareas, como la de procesar grandes volúmenes de datos, o atributos como la neutralidad, la credibilidad o la fiabilidad (Sundar, 2020).

Además, es muy probable que la imparable penetración de los algoritmos de IA en todo tipo de mercados y contextos de decisión modifique la forma en la que abordamos ciertas tareas. Un ejemplo de ello es el estudio de Longoni y Cian (2020) ya

mencionado. En él, los participantes variaban los criterios por los que consideraban adecuado un producto (en este caso, un pastel) dependiendo de si había sido una IA o un humano quien había seleccionado los ingredientes de la receta. Mientras que en el caso del humano se valoraban positivamente los atributos hedónicos del pastel, cuando el seleccionador de los ingredientes había sido una IA los criterios que se ponían en valor eran los atributos utilitarios (como sus propiedades alimentarias). En esta línea, M. P. Burden (2012) advierte que los algoritmos y el diseño de las plataformas que los albergan están cambiando la forma en que consideramos aspectos tan relevantes como la amistad o la búsqueda de pareja. Ejemplo de ello sería Snapchat, una aplicación de mensajería que registra el número de días seguidos en los que dos contactos interactúan entre ellos. Los usuarios deben utilizar la aplicación a diario para poder mantener su "racha"[14] de conversaciones ininterrumpidas y que su contador de interacciones no se ponga a cero, dado que este conteo se ha convertido dentro de la plataforma en un indicador del grado de amistad entre sus usuarios (en su mayoría adolescentes).

Es posible que el paso del tiempo también llegue a afectar a los prejuicios que las personas tienen sobre las habilidades de los algoritmos de IA. Las recientes denuncias de sesgo en los algoritmos, por ejemplo, podrían llegar a empañar la imagen de objetividad que hasta ahora se le ha presupuesto a la IA. O, por el contrario, la presencia cada vez más habitual de los algoritmos de IA en nuestras decisiones diarias podría favorecer la atribución de un mayor número de capacidades positivas a los

[14] Explicación de cómo funcionan las rachas según Snapchat: https://support.snapchat.com/es/a/snapstreaks#

algoritmos, hasta ahora no vinculadas a ellos (como la capacidad creativa, por ejemplo). En nuestros Experimentos 8 y 9 sobre IA en el contexto de arte, cuando los participantes eran conscientes de que la obra artística pertenecía al algoritmo, sus prejuicios previos sobre las habilidades de la IA provocaban la minusvaloración de la calidad de la obra, de su emoción reportada y de la sensibilidad que atribuían al artista. Sin embargo, a su vez, las creencias previas de los participantes sobre las capacidades creativas de la IA modulaban estos juicios. Aunque estos resultados podrían apuntar a la falta de aceptación de los algoritmos en terrenos subjetivos como el arte (en línea con los de los Experimentos 2, 4 y 5 de recomendación encubierta en citas), consideramos que puede ser cuestión de tiempo que se transforme en aceptación e, incluso, que los prejuicios actuales sobre las habilidades y atributos de los algoritmos de IA cambien sustancialmente. Pongamos como ejemplo la recomendación de películas. En el mencionado experimento de Castelo y colaboradores (2019) sobre la objetividad de las tareas, sus participantes confiaban más en las recomendaciones de un humano (76 puntos sobre 100) que en las de un algoritmo (59 puntos) en la tarea de recomendar películas. Sin embargo, esta preferencia por la recomendación humana frente a la algorítmica no parece cristalizarse en comportamiento real de las personas fuera del laboratorio. Plataformas de vídeo como Netflix, HBO y similares son ampliamente utilizadas (el 40% de los hogares españoles durante 2019, por ejemplo, contaban con alguna de estas plataformas; CNMC, 2020). En ellas, son algoritmos los que recomiendan el contenido a consumir, por lo que las personas estarían aceptando su recomendación de forma continuada en una tarea que se considera subjetiva y para la que, supuestamente, se

confía más en el desempeño humano que en el algorítmico. Por ello, aunque en el estudio de Castelo y colaboradores (2019) sus participantes también reportaban menor confianza en la recomendación algorítmica que en la humana en la tarea de componer canciones (81 puntos de confianza para el humano y 43 para el algoritmo) y en la tarea de recomendar una pareja romántica (59 puntos para el humano y 37 para el algoritmo), un mayor incremento de la presencia de los algoritmos en el contexto del arte o de la búsqueda de pareja, al igual que ocurre con la recomendación de películas dado el éxito de las plataformas de vídeo, podría ser el ingrediente que falta para que las personas acepten recomendaciones algorítmicas en tareas subjetivas.

A lo largo de esta tesis hemos ahondado en cómo la presencia y la recomendación del algoritmo en diversos contextos puede influir en las decisiones y juicios de las personas. Consideramos que, en vista de los resultados de nuestros experimentos, las personas corremos el riesgo de subestimar la capacidad de influencia de los algoritmos en los sistemas de recomendación y sobreestimar la libertad de elección que las personas tenemos ellos (Araujo y cols., 2020). En muchos de los sistemas de recomendación que hoy utilizamos, como Tinder, Netflix o Google Search, el contenido se encuentra desde el inicio previamente filtrado y personalizado (por perfil, por consumo o por geolocalización, por ejemplo) aunque el usuario no sea consciente de ello. Esto implica que en realidad la elección del usuario no se produce sobre la totalidad de la oferta de contenidos o servicios, sino sobre la preselección que el algoritmo ha hecho en base a los intereses del usuario o a los de su empresa propietaria (Duhigg, 2018). Aun así, en estos sistemas de recomendación la decisión

final aparenta encontrarse en nuestras manos, aunque los experimentos de esta tesis precisamente hayan puesto esto en cuestión.

Es muy probable que, conforme los sistemas de recomendación se implementen en multitud de contextos de decisión, poco a poco se vaya dando paso a la delegación completa de la decisión en los algoritmos. Esto supondrá que las personas solo intervengan en el proceso de decisión para corregir los errores del algoritmo, para detener la decisión automatizada, o para tratar de comprenderla (Naughton, 2019). De hecho, podemos encontrar ya algunos casos de decisiones delegadas en contextos cotidianos de poca relevancia, como el consumo de contenido de ocio. El algoritmo de Netflix, por ejemplo, invita al usuario a reproducir contenido aleatorio incluso antes de que este llegue a entrar en su perfil (Fernández, 2020). A nivel institucional son cada vez más las decisiones relevantes que se están delegando en los algoritmos. Por ejemplo, el Gobierno de España delegó en un algoritmo la adjudicación del bono social de electricidad, situación que fue denunciada por la Fundación Civio al identificar un malfuncionamiento del algoritmo (Rouco, 2019), lo que confirma también que este tipo de problemas (de falta de objetividad o eficiencia en los algoritmos) solo suelen ser detectados a posteriori.

Dada la necesidad de investigación sobre la interrupción de los algoritmos en las decisiones humanas, son varias las iniciativas que han comenzado a desarrollarse. Una importante línea de investigación en este sentido, por ejemplo, se ha centrado en la explicabilidad e interpretabilidad de los sistemas algorítmicos (Doshi-Velez y cols., 2019) con el objetivo de fomentar la transparencia en sus decisiones (Stoyanovich y Howe, 2018). Esto permitiría cumplir con las directrices de la Unión Europea para

garantizar una IA confiable que ya detallábamos anteriormente (European Commission, 2019). Una propuesta que requiere de importantes cambios técnicos en las propias IAs por la naturaleza opaca que muchas de ellas poseen. Aquí se ubica, por ejemplo, el programa de IA explicable (XAI) de la Agencia de Proyectos de Investigación Avanzados de Defensa de EE.UU. (Gunning, 2019). También iniciativas como la de la Unión Europea, que ha planteado recientemente nuevas normas y medidas para favorecer la excelencia y la confiabilidad en la IA, señalando de alto riesgo a "los sistemas o las aplicaciones de IA que manipulan el comportamiento humano para eludir la voluntad de los usuarios" en aquellos casos en los que se vea afectada "la seguridad, los medios de subsistencia y los derechos de las personas" (Comisión Europea, 2021). Una normativa muy necesaria que, eso sí, no contempla dentro del ámbito de riesgo las potenciales situaciones de influencia mostradas en nuestros experimentos.

Otra línea de investigación en marcha sobre la relación entre personas y algoritmos es la propuesta por Rahwan y colaboradores (2019), que sugieren el estudio del comportamiento de los algoritmos mediante intervenciones experimentales, tanto en el laboratorio como en contextos reales para así conocer cómo reaccionan los algoritmos en su interacción con las personas. En esta línea se encontrarían los trabajos de ingeniería inversa para comprender el funcionamiento de un algoritmo de propiedad privada o de caja negra, como el mencionado de ProPublica para descubrir los criterios de decisión del algoritmo COMPAS, que es empleado en algunos juzgados estadounidenses para determinar las condiciones de la libertad condicional de los acusados (Larson y cols., 2016). O experimentos comportamentales

como el realizado por Bikolabs, el laboratorio de investigación de la consultora tecnológica Biko, donde se muestra que ciertos sistemas de reconocimiento de objetos con IA, como el de Amazon Rekognition, albergan sexismo en el etiquetado de sus imágenes. Para descubrirlo, los autores analizaron el etiquetado asignado por la IA a fotografías de hombres y mujeres portando objetos históricamente estereotipados, y encontraron un claro sesgo en la eficacia del sistema y las etiquetas utilizadas (Agudo y Liberal, 2020). Esta tesis contribuye también a esta línea de investigación, aunque desde una perspectiva diferente, dado que mucha de la investigación existente se centra en el estudio del comportamiento del algoritmo, en ocasiones desde un punto de vista puramente técnico. Nuestros experimentos, por su parte, suponen una aportación relevante y con cierto valor ecológico, pero con el foco puesto, desde un punto de vista psicológico, en el comportamiento de las personas, en lugar de en el comportamiento del algoritmo.

Así, sumamos evidencia sobre cómo ciertas capacidades atribuidas al algoritmo, como la objetividad, pueden influir en las decisiones y afectar a los juicios sobre su desempeño. Además, esta tesis contribuye a aumentar la evidencia sobre las estrategias persuasivas que pueden estar utilizándose a la hora de mostrar las recomendaciones algorítmicas y su potencial de influencia, no tanto frente a la recomendación humana en general, sino frente al criterio propio.

Por todo ello, y siendo conscientes de que queda mucho camino por recorrer, esperamos que este trabajo contribuya a ampliar la poca evidencia empírica existente en el campo de la Psicología sobre este tema, al abordar la respuesta humana ante la recomendación del algoritmo de IA y la interacción con él en ámbitos tan relevantes

como la política, la búsqueda de pareja o el arte, y que además permita comprender mejor cómo la interacción con los algoritmos puede determinar nuestras decisiones, juicios y experiencias.

Referencias bibliográficas

Abakoumkin, G. (2011). Forming choice preferences the easy way: Order and familiarity effects in elections. *Journal of Applied Social Psychology*, *41*(11), 2689–2707. https://doi.org/10.1111/j.1559-1816.2011.00845.x

Acxiom. (2015). *Data services API: Data bundles*. https://developer.myacxiom.com/code/api/data-bundles/bundle/basicDemographics

Adams, T. (2010, July 11). David Cope: 'You pushed the button and out came hundreds and thousands of sonatas'. *The Guardian*. https://www.theguardian.com/technology/2010/jul/11/david-cope-computer-composer

Agudo, U., Casacuberta, D., Guersenzvaig, A., & Liberal, K. G. (2021, March 25). Sobre los riesgos del reconocimiento y análisis facial. *Ctxt*. https://ctxt.es/es/20210301/Firmas/35455/carta-gobierno-moratoria-reconocimiento-facial-racismo-sexismo-algoritmos.htm

Agudo, U., & Liberal, K. G. (2020, September 9). El automágico traje del emperador. *Bikolabs*. https://medium.com/bikolabs/el-automagico-traje-del-emperador-c2a0bbf6187b

Al-slaity, A. N. (2021). *Beyond recommendation accuracy: A human-like recommender system* [Doctoral dissertation, University of Ottawa]. http://dx.doi.org/10.20381/ruor-26103

Al Jazeera. (2018, March 28). Cambridge Analytica and Facebook: The scandal so far. *Al Jazeera.* https://www.aljazeera.com/news/2018/3/28/cambridge-analytica-and-facebook-the-scandal-so-far

Alexander, V., Blinder, C., & Zak, P. J. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology. *Computers in Human Behavior*, *89*, 279–288.

https://doi.org/10.1016/j.chb.2018.07.026

Alslaity, A., & Tran, T. (2019). Towards persuasive recommender systems. *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)*, 143–148. https://doi.org/10.1109/INFOCT.2019.8711416

Alvarado, O., Abeele, V. Vanden, Geerts, D., & Verbert, K. (2019). "I really don't know what 'thumbs up' means": Algorithmic experience in movie recommender algorithms. *Human-Computer Interaction – INTERACT 2019. Lecture Notes in Computer Science*, *vol 11748*, 521–541. https://doi.org/10.1007/978-3-030-29387-1_30

Angwin, J. (2017, November 29). Facebook to temporarily block advertisers from excluding audiences by race. *ProPublica*. https://www.propublica.org/article/facebook-to-temporarily-block-advertisers-from-excluding-audiences-by-race

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Angwin, J., Scheiber, N., & Tobin, A. (2017, December 20). Dozens of companies are using Facebook to exclude older workers from job ads. *ProPublica*. https://www.propublica.org/article/facebook-ads-age-discrimination-targeting

Angwin, J., Tobin, A., & Varner, M. (2017, November 21). Facebook (still) letting housing advertisers exclude users by race. *ProPublica*. https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin

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

00931-w

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, *37*(1), 93–110. https://doi.org/10.1016/0749-5978(86)90046-4

Arkes, H. R., Shaffer, V. A., & Medow, M. A. (2007). Patients derogate physicians who use a computer-assisted diagnostic aid. *Medical Decision Making*, *27*(2), 189–202. https://doi.org/10.1177/0272989X06297391

Bailenson, J. N., Iyengar, S., Yee, N., & Collins, N. A. (2008). Facial similarity between voters and candidates causes influence. P*ublic Opinion Quarterly, 72*(5), 935–961. https://doi.org/10.1093/poq/nfn064

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. https://doi.org/https://doi.org/10.1037/a0033872

Banerjee, S., & John, P. (2019). Nudge plus: incorporating reflection into behavioural public policy. *SSRN*. https://doi.org/10.2139/ssrn.3479690

Banker, S., & Khetani, S. (2019). Algorithm overdependence: How the use of algorithmic recommendation systems can increase risks to consumer well-being. *Journal of Public Policy & Marketing*, *38*(4), 500–515. https://doi.org/10.1177/0743915619858057

Barberia, I., Tubau, E., Matute, H., & Rodríguez-Ferreiro, J. (2018). A short educational intervention diminishes causal illusions and specific paranormal beliefs in undergraduates. *PLOS ONE*, *13*(1), e0191907. https://doi.org/10.1371/journal.pone.0191907

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732. http://dx.doi.org/10.2139/ssrn.2477899

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019).

Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Beer, D. (2017). The social power of algorithms. *Information, Communication & Society*, *20*(1), 1–13. https://doi.org/10.1080/1369118X.2016.1216147

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2020). Watch me improve—Algorithm aversion and demonstrating the ability to learn. *Business and Information Systems Engineering*, 1–14. https://doi.org/10.1007/s12599-020-00678-5

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

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421

Bonneau, C. W., & Cann, D. M. (2015). Party identification and vote choice in partisan and nonpartisan elections. *Political Behavior*, *37*(1), 43–66. https://doi.org/10.1007/s11109-013-9260-2

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968-1987. *Psychological Bulletin*, *106*(2), 265–289. https://doi.org/10.1037/0033-2909.106.2.265

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

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

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of

Facebook algorithms. *Information Communication and Society*, *20*(1), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

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

Burden, B. C. (2002). When bad press is good news. *Harvard International Journal of Press/Politics*, *7*(3), 76–89. https://doi.org/10.1177/1081180x0200700305

Burden, M. P. (2012). *The hidden persuasions of algorithms* [Doctoral dissertation, University of Alberta]. https://doi.org/10.7939/R3W600

Burgess, M. (2020, August 6). Police built an AI to predict violent crime. It was seriously flawed. *WIRED*. https://www.wired.co.uk/article/police-violence-prediction-ndas

Burgoon, J. K. (2015). Expectancy violations theory. In *The International Encyclopedia of Interpersonal Communication* (pp. 1–9). Wiley. https://doi.org/10.1002/9781118540190.wbeic102

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 205395171562251. https://doi.org/10.1177/2053951715622512

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2019). 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

Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). How Cambridge Analytica turned Facebook 'likes' into a lucrative political tool. *The Guardian*. https://www.theguardian.com/technology/2018/mar/17/facebook-cambridge-analytica-kogan-data-algorithm

Cameron, K. A. (2009). A practitioner's guide to persuasion: An overview of 15 selected persuasion theories, models and frameworks. *Patient Education and Counseling,*

74(3), 309–317. https://doi.org/10.1016/j.pec.2008.12.003

Cappelli, P., Tambe, P., & Yakubovich, V. (2018). Artificial intelligence in human resources management: Challenges and a path forward. *SSRN*. https://doi.org/10.2139/ssrn.3263878

Caraban, A., Karapanos, E., Gonçalves, D., & Campos, P. (2019, May 2). 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 503*, 1–15. https://doi.org/10.1145/3290605.3300733

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

Castelo, N. (2019). *Blurring the line between human and machine: Marketing artificial intelligence* [Doctoral dissertation, Columbia University]. https://doi.org/10.7916/d8-k7vk-0s40

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

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20–23. https://doi.org/10.1038/538020a

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic information processing within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 212–252). The Guilford Press.

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

Chamberlain, R., Mullin, C., Scheerlinck, B., & Wagemans, J. (2018). Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of*

Aesthetics, Creativity, and the Arts, *12*(2), 177–192. https://doi.org/10.1037/aca0000136

Chen, A. (2019, February 7). Why the future of life insurance may depend on your online presence. *The Verge*. https://www.theverge.com/2019/2/7/18211890/social-media-life-insurance-new-york-algorithms-big-data-discrimination-online-records

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

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

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

Christie's. (2018). *The first piece of AI-generated art to come to auction*. https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx

Christl, W. (2017). *Corporate surveillance in everyday life*. Cracked Labs. http://crackedlabs.org/en/corporate-surveillance

Chu, H., Urtasun, R., & Fidler, S. (2017). *Song from PI: A musically plausible network for pop music generation*. arXiv:1611.03477

Cialdini, R. B. (1993). *Influence. The psychology of persuasion*. HarperCollins.

Cialdini, R. B., & Sagarin, B. (2005). Principles of interpersonal influence. In T. C. Brock & M. C. Green (Eds.), *Persuasion: Psychological insights and perspectives* (pp.

146–169). Sage Publications, Inc.

CNMC (2020, May 29). *Cuatro de cada diez hogares con Internet consumen contenidos audiovisuales en plataformas online de pago*. https://www.cnmc.es/prensa/cnmc-panel-hogares-ott-audiovisual

Coffey, H. (2018, November 19). Airlines face crack down on use of 'exploitative' algorithm that splits up families on flights. *The Independent*. https://www.independent.co.uk/travel/news-and-advice/airline-flights-pay-extra-to-sit-together-split-up-family-algorithm-minister-a8640771.html

Collins, J. (2018, December 12). Simple heuristics that make algorithms smart. *Behavioral Scientist*. http://behavioralscientist.org/simple-heuristics-that-make-algorithms-smart/

Comisión Europea. (2021, April 21). Una Europa adaptada a la era digital: Inteligencia artificial. *Web Oficial de la Unión Europea*. https://ec.europa.eu/commission/presscorner/detail/es/ip_21_1682

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl, J. (2003). Is seeing believing?: How recommender system interfaces affect users' opinions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'03)*, 585–592. https://doi.org/10.1145/642611.642713

Cummings, M. (2004). Automation bias in intelligent time critical decision support systems. *AIAA, 2*, 557–562. https://doi.org/doi:10.2514/6.2004-6313

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems : theory and results* [Doctoral dissertation, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/15192

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

De Neys, W. (2021). On dual- and single-process models of thinking. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620964172

Deah, D. (2018). How AI-generated music is changing the way hits are made. *The Verge*. https://www.theverge.com/2018/8/31/17777008/artificial-intelligence-taryn-southern-amper-music

Del Castillo, C. (2018, November 2). ¿Y si la gran noticia falsa de 2018 son los datos sobre la influencia de Facebook? *Eldiario.es*. https://www.eldiario.es/tecnologia/potencial-publicitario-Facebook-fake-Zuckerberg_0_827667467.html

Dewey, C. (2021). Reframing single- and dual-process theories as cognitive models: Commentary on De Neys (2021). *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691621997115

Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56–62. https://doi.org/10.1145/2844110

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

Dietvorst, B. J. (2016). People reject (superior) algorithms because they compare them to counter-normative reference points. *Social Science Research Network*. https://dx.doi.org/10.2139/ssrn.2881503

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

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour and Information Technology*, *18*(6), 399–411. https://doi.org/10.1080/014492999118832

Dijkstra, J. J., Liebrand, W. B. G., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour and Information Technology*, *17*(3), 155–163. https://doi.org/10.1080/014492998119526

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., Weller, A., Wood, A. (2019). *Accountability of AI under the law: The role of explanation*. arXiv:1711.01134

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

Duhigg, C. (2018, February 20). The case against Google. *The New York Times*. https://www.nytimes.com/2018/02/20/magazine/the-case-against-google.html

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

Duportail, J. (2017, September 26). I asked Tinder for my data. It sent me 800 pages of my deepest, darkest secrets. *The Guardian*. https://www.theguardian.com/technology/2017/sep/26/tinder-personal-data-dating-app-messages-hacked-sold

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

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Eastwood, J., Snook, B., & Luther, K. (2012). What people want from their professionals: Attitudes toward decision-making strategies. *Journal of Behavioral Decision Making*, *25*(5), 458–468. https://doi.org/10.1002/bdm.741

Eiband, M., Völkel, S. T., Buschek, D., Cook, S., & Hussmann, H. (2019). When people and algorithms meet. *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*, 96–106. https://doi.org/10.1145/3301275.3302262

Einhorn, H. J. (1986). Accepting error to make less error. *Journal of Personality Assessment*, *50*(3), 387–395. https://doi.org/10.1207/s15327752jpa5003_8

Elish, M. C., & Boyd, D. (2018, November 13). Don't believe every AI you see. *The Ethical Machine*. https://ai.shorensteincenter.org/ideas/2018/11/12/dont-believe-every-ai-you-see-1

Epp, C., Lippold, M., & Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, 715–724. https://doi.org/10.1145/1978942.1979046

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*, *112*(33), E4512–E4521. https://doi.org/10.1073/pnas.1419828112

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). 'I always assumed that I wasn't really that close to [her]'. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing*

Systems (CHI '15), 153–162. https://doi.org/10.1145/2702123.2702556

Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (2017). "Be careful; Things can be worse than they appear" - Understanding biased algorithms and users' behavior around them in rating platforms. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, 62–71.

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

Evans, J. S. B. T. (2010). Intuition and reasoning: A dual-process perspective. *Psychological Inquiry, 21*(4), 313–326. https://doi.org/10.1080/1047840X.2010.521057

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

Fernández, Y. (2020, August 19). Reproducción aleatoria de Netflix: qué es y cómo usar esta función de contenido aleatorio. *Xataka*. https://www.xataka.com/basics/reproduccion-aleatoria-netflix-que-como-usar-esta-funcion-contenido-aleatorio

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics*. SAGE Publications.

Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T., & Sprecher, S. (2012). Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public Interest*, *13*(1), 3–66. https://doi.org/10.1177/1529100612436522

Fisher, H. (2018). *Helen Fisher's personality test*. The anatomy of love. https://theanatomyoflove.com/relationship-quizzes/helen-fishers-personality-test/

Fogg, B. J. (2002). *Persuasive technology: using computers to change what we think and do*. Ubiquity. https://doi.org/10.1145/764008.763957

Forer, B. R. (1949). The fallacy of personal validation: A classroom demonstration of gullibility. *The Journal of Abnormal and Social Psychology*, *44*(1), 118–123. https://doi.org/10.1037/h0059240

Friedel, F. (2018). *Machines who play (and compose) music*. The Friedel Chronicles. https://medium.com/@frederic_38110/machines-that-play-and-compose-music-70abfe9a8549

Fry, H., & Krzywinski, M. (2019). *Hola mundo: Cómo seguir siendo humanos en la era de los algoritmos*. Blackie Books

Geslevich, N. P. (2019). Consumer finance and AI: The death of second opinions? *New York University Journal of Legislation and Public Policy*, *22*(2), 319–374.

Gigerenzer, G. (2015). *Risk savvy: How to make good decisions.* Penguin Books.

Gigerenzer, G. (2015b). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology*, *6*(3), 361–383. https://doi.org/10.1007/s13164-015-0248-1

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, *5*(3–4), 303–336. https://doi.org/10.1561/105.00000092

Gillespie, T. (2014). The relevance of algorithms. In Tarleton Gillespie, Pablo J. Boczkowski, Kirsten A. Foot (Eds.), *Media Technologies: Essays on Communication, Materiality, and Society* (pp. 167–194). MIT Press. https://doi.org/10.7551/mitpress/9780262525374.003.0009

Gkika, S., & Lekakos, G. (2014). The persuasive role of explanations in recommender systems. *Second International Workshop on Behavior Change Support Systems (BCSS 2014), 1153,* 59-68.

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

Greenawalt, L. (2018, December 6). The history of U.S. housing segregation points to the devastating consequences of algorithmic bias. *The Ethical Machine*. https://ai.shorensteincenter.org/ideas/2018/12/1/federal-housing-history-shows-that-algorithmic-bias-carries-long-lasting-disastrous-consequences-29jz2

Greig, J. (2018, May 8). Welsh police facial recognition software has 92% fail rate, showing dangers of early AI. *TechRepublic*. https://www.techrepublic.com/article/welsh-police-facial-recognition-has-92-fail-rate-showing-dangers-of-early-ai/

Gretzel, U., & Fesenmaier, D. R. (2006). Persuasion in recommender systems. *International Journal of Electronic Commerce*, *11*(2), 81–100. https://doi.org/10.2753/JEC1086-4415110204

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Gunaratne, J., Zalmanson, L., & Nov, O. (2018). The persuasive power of algorithmic and crowdsourced advice. *Journal of Management Information Systems*, *35*(4), 1092–1120. https://doi.org/10.1080/07421222.2018.1523534

Gunning, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19*), ii–ii. https://doi.org/10.1145/3301275.3308446

Hallinan, B., & Striphas, T. (2016). Recommended for you: The Netflix Prize and the production of algorithmic culture. *New Media & Society*, *18*(1), 117–137. https://doi.org/10.1177/1461444814538646

Hansen, P. G., & Jespersen, A. M. (2013). Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, *4*(1), 3–28. https://doi.org/10.1017/s1867299x00002762

Harari, Y. N. (2016). *Homo Deus: Breve historia del mañana.* Debate.

Harari, Y. N. (2018). *21 lecciones para el siglo XXI*. Debate.

Harlan, E., & Schnuck, O. (2021, February 16). *Objective or biased*. BR24. https://web.br.de/interaktiv/ki-bewerbung/en/

Harris, A., Islam, S., Qadir, J., & Khan, U. A. (2017). *Persuasive technology for human development: Review and case study*. arXiv:1708.08758

Harwell, D. (2018, November 23). Wanted: The 'perfect babysitter.' Must pass AI scan for respect and attitude. *The Washington Post*. https://www.washingtonpost.com/technology/2018/11/16/wanted-perfect-babysitter-must-pass-ai-scan-respect-attitude/

Harwell, D. (2019, November 6). HireVue's AI face-scanning algorithm increasingly decides whether you deserve the job. *The Washington Post*. https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

Hern, A. (2014, July 29). OKCupid: We experiment on users. Everyone does. *The Guardian*. https://www.theguardian.com/technology/2014/jul/29/okcupid-experiment-human-beings-dating

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*(6), 973–986. https://doi.org/10.1177/1745691617702496

Hickey, M. (1999). Assessment rubrics for music composition. *Music Educators Journal*,

85(4), 26–52. https://doi.org/10.2307/3399530

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, *1*(3), 333–342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

Hill, K. (2020, June 24). Wrongfully accused by an algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Hong, J. W., & Curran, N. M. (2019). Artificial intelligence, artists, and art: Attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Transactions on Multimedia Computing, Communications and Applications*, *15*(2s), 1–16. https://doi.org/10.1145/3326337

Hong, J. W., Peng, Q., & Williams, D. (2020). Are you ready for artificial Mozart and Skrillex? An experiment testing expectancy violation theory and AI music. *New Media & Society*, 146144482092579. https://doi.org/10.1177/1461444820925798

Hortal, A. (2019). Nudging and educating: Bounded axiological rationality in behavioral insights. *Behavioural Public Policy*, 1–24. https://doi.org/10.1017/bpp.2019.2

Hosanagar, K. (2019, March 5). Free will in an algorithmic world. *OneZero*. https://onezero.medium.com/free-will-in-an-algorithmic-world-8d5acb550cb7

Hunch (website). (2021, June 9). In *Wikipedia.* https://en.wikipedia.org/wiki/Hunch_(website)

Inbar, Y., Cone, J., & Gilovich, T. (2010). People's intuitions about intuitive insight and intuitive choice. *Journal of Personality and Social Psychology*, *99*(2), 232–247. https://doi.org/10.1037/a0020215

Iyengar, S. (2011). *El arte de elegir*. Gestión 2000.

Jago, A. S. (2019). Algorithms and authenticity. *Academy of Management Discoveries*, *5*(1), 38–56. https://doi.org/10.5465/amd.2017.0002

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, *302*(5649), 1338–

1339. https://doi.org/10.1126/science.1091721

Johnson, G. (1997, November 11). Undiscovered Bach? No, a computer wrote it. *The New York Times*. https://www.nytimes.com/1997/11/11/science/undiscovered-bach-no-a-computer-wrote-it.html

Johnson, K., Ren, L., Kuchar, J., & Oman, C. (2002). Interaction of automation and time pressure in a route replanning task. *International Conference on Human–ComPuter Interaction in Aeronautics (HCI-Aero2002)*, 132–137. https://www.aaai.org/Library/HCI/2002/hci02-021.php

Jones, J. J., Bond, R. M., Bakshy, E., Eckles, D., & Fowler, J. H. (2017). Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election. *PLOS ONE*, *12*(4), e0173851. https://doi.org/10.1371/journal.pone.0173851

Jozuka, E. (2016, March 24). A japanese AI almost won a literary prize. *VICE*. https://www.vice.com/en_us/article/wnxnjn/a-japanese-ai-almost-won-a-literary-prize

Kahneman, D. (2012). *Pensar rápido, pensar despacio*. Debate.

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

Karlsen, R., & Andersen, A. (2019). Recommendations with a nudge. *Technologies*, *7*(2), 45. https://doi.org/10.3390/technologies7020045

Karras, T., Laine, S., & Aila, T. (2018). *A style-based generator architecture for generative adversarial networks*. arXiv:1812.04948

Kenyon, H. (2018, February 1). AI, please explain yourself. *Signal*. https://www.afcea.org/content/ai-please-explain-yourself

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information*

Communication and Society, *20*(1), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

Knight, W. (2017, April 11). The dark secret at the heart of AI. *MIT Technology Review*. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4–5), 441–504. https://doi.org/10.1007/s11257-011-9118-4

Koebler, J. (2020, June 29). Detroit police chief: Facial recognition software misidentifies 96% of the time. *Motherboard*. https://www.vice.com/en/article/dyzykz/detroit-police-chief-facial-recognition-software-misidentifies-96-of-the-time

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543–556. https://doi.org/10.1037/a0039210

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Kosinski, M., Stillwell, D., Kohli, P., and Bachrach, Y., & Graepel, T. (2012). *Personality and website choice.* ACM Web Sciences 2012. https://www.microsoft.com/en-us/research/publication/personality-and-website-choice/

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790.

https://doi.org/10.1073/pnas.1320040111

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

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

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

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

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

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *0*, 1–22. https://doi.org/10.1093/jcr/ucz013

Longoni, C., & Cian, L. (2020). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing*. https://doi.org/10.1177/0022242920957347

Loughnan, S., & Haslam, N. (2007). Animals and androids. *Psychological Science*, *18*(2), 116–121. https://doi.org/10.1111/j.1467-9280.2007.01858.x

Mantilla, J. R. (2019, February 5). Un algoritmo completa la misteriosa 'Sinfonía

inacabada' de Schubert. *El País*. https://elpais.com/cultura/2019/02/04/actualidad/1549284459_079024.html

Marsh, S. (2019, October 15). One in three councils using algorithms to make welfare decisions. *The Guardian*. https://www.theguardian.com/society/2019/oct/15/councils-using-algorithms-make-welfare-decisions-benefits

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(48), 12714–12719. https://doi.org/10.1073/pnas.1710966114

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

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

Merino, M. (2019, April 22). Un canal de Youtube emite de forma constante música 'death metal' generada sobre la marcha por una inteligencia artificial. *Xataka*. https://www.xataka.com/inteligencia-artificial/canal-youtube-emite-forma-constante-musica-death-metal-generada-marcha-inteligencia-artificial

Meske, C., & Potthoff, T. (2017). The DINU-Model - A process model for the design of nudges. *Proceedings of the 25th European Conference on Information Systems (ECIS)*, 2587–2597. http://aisel.aisnet.org/ecis2017_rip/11

Miller, A. P. (2018, July 26). Want less-biased decisions? Use algorithms. *Harvard Business Review*. https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

Miyzaki, S. (2012, September 28). Algorhythmics: Understanding micro-temporality in computational cultures. *Computational Culture*. http://computationalculture.net/algorhythmics-understanding-micro-temporality-in-computational-cultures/

Moffat, D. C., & Kelly, M. (2006). *An investigation into people's bias against computational creativity in music composition*. Proceedings of the 3rd international joint workshop on computational creativity (ECAI06 Workshop), 1–8.

Montoya, R. M., Horton, R. S., Vevea, J. L., Citkowicz, M., & Lauber, E. A. (2017). A re-examination of the mere exposure effect: The influence of repeated exposure on recognition, familiarity, and liking. *Psychological Bulletin*, *143*(5), 459–498. https://doi.org/10.1037/bul0000085

Morales, F. J., Moya, M., Gaviria, E., & Cuadrado, I. (2007). *Psicología Social* (3ª ed.). Mc Graw-Hill.

Nasir, M., Baucom, B. R., Georgiou, P., & Narayanan, S. (2017). Predicting couple therapy outcomes based on speech acoustic features. *PLOS ONE*, *12*(9), e0185123. https://doi.org/10.1371/journal.pone.0185123

Naughton, J. (2019, March 31). When new technology goes badly wrong, humans carry the can. *The Guardian*. https://www.theguardian.com/commentisfree/2019/mar/31/when-new-technology-goes-wrong-humans-carry-the-can

Neyland, D., & Möllers, N. (2016). Algorithmic IF … THEN rules and the conditions and consequences of power. *Information, Communication & Society ISSN:*, *20*(1), 45–62. https://doi.org/10.1080/1369118X.2016.1156141

Niiler, E. (2019, March 25). Can AI be a fair judge in court? Estonia thinks so. *WIRED*. https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/

Noorbehbahani, F., & Zarein, Z. (2018). The impact of demographic factors on persuasion strategies in personalized recommender system. *8th International*

Conference on Computer and Knowledge Engineering, ICCKE 2018, 104–109. https://doi.org/10.1109/ICCKE.2018.8566550

O'Neil, C. (2018). *Armas de destrucción matemática: Cómo el Big Data aumenta la desigualdad y amenaza la democracia*. Capitán Swing.

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

Oduor, M., Alahäivälä, T., & Oinas-Kukkonen, H. (2014). Persuasive software design patterns for social influence. *Personal and Ubiquitous Computing*, *18*(7), 1689–1704. https://doi.org/10.1007/s00779-014-0778-z

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

Palmer, C. L., & Peterson, R. D. (2015). Halo effects and the attractiveness premium in perceptions of political expertise. *American Politics Research*, *44*(2), 353–382. https://doi.org/10.1177/1532673X15600517

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

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.

Pearce, M. T., & Wiggins, G. A. (2007). Evaluating cognitive models of musical composition. In A. Cardoso & G. A. Wiggins (Eds.)*, Proceedings of the 4th international joint workshop on computational creativity* (pp. 73– 80). Goldsmiths, University of London.

Peskin, M., & Newell, F. N. (2004). Familiarity breeds attraction: Effects of exposure on the attractiveness of typical and distinctive faces. *Perception*, *33*(2), 147–157. https://doi.org/10.1068/p5028

Peters, A. (2018, January 11). Having a heart attack? This AI helps emergency dispatchers find out. *Fast Company*. https://www.fastcompany.com/40515740/having-a-heart-attack-this-ai-helps-emergency-dispatchers-find-out

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*(C), 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

Pheterson, M., & Horai, J. (1976). The effects of sensation seeking, physical attractiveness of stimuli, and exposure frequency on liking. *Social Behavior and Personality: An International Journal*, *4*(2), 241–247. https://doi.org/10.2224/sbp.1976.4.2.241

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

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468. https://doi.org/10.1002/bdm.542

Quach, K. (2020, December 9). Japan pours millions into AI-powered dating to get its people making babies again. *The Register*. https://www.theregister.com/2020/12/09/japan_ai_dating/

Raft, W., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, *207*(4430), 557–558. https://doi.org/10.1126/science.7352271

Ragot, M., Martin, N., & Cojean, S. (2020). AI-generated vs. human artworks. A perception bias towards artificial intelligence? *Extended Abstracts of the 2020 CHI*

Conference on Human Factors in Computing Systems (CHI EA '20), 1–10. https://doi.org/10.1145/3334480.3382892

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'Sandy', … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Rhodes, G., Halberstadt, J., & Brajkovich, G. (2001). Generalization of mere exposure effects to averaged composite faces. *Social Cognition*, *19*(1), 57–70. https://doi.org/10.1521/soco.19.1.57.18961

Rouco, F. (2019, September 22). La corriente que pide el código fuente de los programas que deciden las ayudas al ciudadano: que el algoritmo no discrimine. *Xataka*. https://www.xataka.com/legislacion-y-derechos/corriente-que-pide-codigo-fuente-programas-que-deciden-ayudas-al-ciudadano

Sabatini, J. (2017, October 3). The one reason nobody is talking realistically about driverless cars. *Car and Driver*. https://www.caranddriver.com/features/a15080116/the-one-reason-nobody-is-talking-realistically-about-driverless-cars-feature/

Saenz, A. (2009, October 9). Music created by learning computer getting better. *Singularity Hub*. https://singularityhub.com/2009/10/09/music-created-by-learning-computer-getting-better/

Sætra, H. S. (2019). When nudge comes to shove: Liberty and nudging in the era of big data. *Technology in Society*, *59*, 1–10. https://doi.org/10.1016/j.techsoc.2019.04.006

Sample, I. (2020, January 1). AI system outperforms experts in spotting breast cancer. *The Guardian*. https://www.theguardian.com/society/2020/jan/01/ai-system-outperforms-experts-in-spotting-breast-cancer

Samson, A., & Voyer, B. G. (2012). Two minds, three ways: dual system and dual process models in consumer psychology. *AMS Review, 2*(2), 48–71. https://doi.org/10.1007/S13162-012-0030-9

Sances, M. W. (2018). Ideology and vote choice in U.S. mayoral elections: Evidence from Facebook surveys. *Political Behavior*, *40*(3), 737–762. https://doi.org/10.1007/s11109-017-9420-x

Sandle, P. (2018, June 27). Feeling poorly? The app will see you now. *Reuters*. https://www.reuters.com/article/us-britain-healthcare-ai/feeling-poorly-the-app-will-see-you-now-idUSKBN1JN35S

Scheiber, N. (2017, April 2). How Uber uses psychological tricks to push its drivers' buttons. *The New York Times*. https://www.nytimes.com/interactive/2017/04/02/technology/uber-drivers-psychological-tricks.html

Schindler, I., Hosoya, G., Menninghaus, W., Beermann, U., Wagner, V., Eid, M., & Scherer, K. R. (2017). Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PLOS ONE*, *12*(6). https://doi.org/10.1371/journal.pone.0178899

Schwartz, B. (2005). *The paradox of choice: why more is less*. HarperCollins.

Schwienbacher, J. (2020). *Reactions on algorithms: A systematic literature review of algorithm aversion and algorithm appreciation* [Doctoral dissertation, University of Innsbruck]. https://diglib.uibk.ac.at/ulbtirolhs/content/titleinfo/5099158/

Seaver, N. (2019). Captivating algorithms: Recommender systems as traps. *Journal of Material Culture*, *24*(4), 421–436. https://doi.org/10.1177/1359183518820366

Sharp, N. (2019, September 5). It's Facebook official, dating is here. *About Facebook*. https://about.fb.com/news/2019/09/facebook-dating/

Sieck, W. R., & Arkes, H. R. (2005). The recalcitrance of overconfidence and its

contribution to decision aid neglect. *Journal of Behavioral Decision Making*, *18*(1), 29–53. https://doi.org/10.1002/bdm.486

Sinha, R., & Swearingen, K. (2001). *Comparing recommendations made by online systems and friends*. Proceedings of the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries.

Solon, O., & Farivar, C. (2019, April 16). Mark Zuckerberg leveraged Facebook user data to fight rivals and help friends, leaked documents show. *NBC News*. https://www.nbcnews.com/tech/social-media/mark-zuckerberg-leveraged-facebook-user-data-fight-rivals-help-friends-n994706

Soulié, J. (2015). *Testpolitico.com*. Testpolitico.com. http://www.testpolitico.com/test/

Spiesel, C. (2020). Technology's black mirror: Seeing, machines, and culture. *International Journal for the Semiotics of Law*, 1–17. https://doi.org/10.1007/s11196-019-09679-4

Springer, A., Hollis, V., & Whittaker, S. (2017). *Dice in the black box: User experiences with an inscrutable algorithm*. AAAI Spring Symposium Series. https://aaai.org/ocs/index.php/SSS/SSS17/paper/view/15372

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

Stoyanovich, J., & Howe, B. (2018, November 27). Follow the data! Algorithmic transparency starts with data transparency. *The Ethical Machine*. https://ai.shorensteincenter.org/ideas/2018/11/26/follow-the-data-algorithmic-transparency-starts-with-data-transparency

Sugden, R. (2017). Do people really want to be nudged towards healthy lifestyles? *International Review of Economics*, *64*(2), 113–123. https://doi.org/10.1007/s12232-016-0264-1

Suh, B. (2019, May 2). Can AI nudge us to make better choices? *Harvard Business Review*. https://hbr.org/2019/05/can-ai-nudge-us-to-make-better-choices

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital Media, Youth, and Credibility*, 73–100. https://doi.org/10.1162/dmal.9780262562324.073

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *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*, 1–9. https://doi.org/10.1145/3290605.3300768

Sunstein, C. R. (2017). Nudges that fail. *Behavioural Public Policy*, *1*(1), 4–25. https://doi.org/10.1017/bpp.2016.3

Susser, D. (2019). Invisible influence: Artificial intelligence and the ethics of adaptive choice architectures. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society,* 403–408. https://doi.org/10.1145/3306618.3314286

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, *8*(2). https://doi.org/10.14763/2019.2.1410

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science Magazine*, *361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

Thaler, R. H., & Sunstein, C. R. (2009). *Un pequeño empujón (Nudge): El impulso que necesitas para tomar mejores decisiones sobre salud, dinero y felicidad*. Taurus.

Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2019). My friends, editors, algorithms, and I: Examining audience attitudes to news selection. *Digital Journalism*, *7*(4), 447–469. https://doi.org/10.1080/21670811.2018.1493936

Tiffany, K. (2019, March 18). The Tinder algorithm, explained. *Vox*. https://www.vox.com/2019/2/7/18210998/tinder-algorithm-swiping-tips-dating-app-science

Tucker, P. (2019, March 26). The US military is creating the future of employee monitoring. *Defense One*. https://www.defenseone.com/technology/2019/03/us-military-creating-future-employee-monitoring/155824/

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

Varghese, S. (2019, October 21). The junk science of emotion-recognition technology. *The Outline*. https://theoutline.com/post/8118/junk-emotion-recognition-technology

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly: Management Information Systems*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Victor, D. (2016, March 25). Microsoft created a Twitter bot to learn from users. It quickly became a racist jerk. *The New York Times*. https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html?_r=0

Villareal, A. (2019, February 5). Google o Amazon hacen experimentos de psicología con nosotros sin avisarnos de ello. *El Confidencial*. https://www.elconfidencial.com/tecnologia/ciencia/2019-02-05/helena-matute-psicologia-experimental-cognitiva_1803274/

Waddell, T. F. (2019). Can an algorithm reduce the perceived bias of news? Testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. *Journalism & Mass Communication Quarterly*, *96*(1), 82–100. https://doi.org/10.1177/1077699018815891

Wærn, Y., & Ramberg, R. (1996). People's perception of human and computer advice. *Computers in Human Behavior*, *12*(1), 17–27. https://doi.org/10.1016/0747-5632(95)00016-X

Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). *Deep learning for identifying metastatic breast cancer*. arXiv:1606.05718

Warshaw, J., Matthews, T., Whittaker, S., Kau, C., Bengualid, M., & Smith, B. A. (2015). Can an algorithm know the "real you"? *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 797–806. https://doi.org/10.1145/2702123.2702274

Watson, C. (2018, April 11). The key moments from Mark Zuckerberg's testimony to Congress. *The Guardian*. https://www.theguardian.com/technology/2018/apr/11/mark-zuckerbergs-testimony-to-congress-the-key-moments

Wegner, D. M., & Gray, K. J. (2017). *The mind club: who thinks, what feels, and why it matter*. Penguin Books.

Weinmann, M., Schneider, C., & Brocke, J. v. (2016). Digital nudging. *Business and Information Systems Engineering*, *58*(6), 433–436. https://doi.org/10.1007/s12599-016-0453-1

White, A. E., Kenrick, D. T., & Neuberg, S. L. (2013). Beauty at the ballot box: Disease threats predict preferences for physically attractive leaders. *Psychological Science*, *24*(12), 2429–2436. https://doi.org/10.1177/0956797613493642

Wiggers, K. (2019, July 1). AI classifies people's emotions from the way they walk. *VentureBeat*. https://venturebeat.com/2019/07/01/ai-classifies-peoples-emotions-from-the-way-they-walk/

Williams, H., & McOwan, P. W. (2014). Magic in the machine: A computational magician's assistant. *Frontiers in Psychology*, *5*, 1283. https://doi.org/10.3389/fpsyg.2014.01283

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

WPP's Data Alliance (2016). *WPP's Data Alliance and Spotify announce global data partnership*. http://www.thedataalliance.com/blog/wpps-data-alliance-and-spotify-announce-global-data-partnership/

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

Yang, L.-C., Chou, S.-Y., & Yang, Y.-H. (2017). *MidiNet: A convolutional generative adversarial network for symbolic-domain music generation*. arXiv:1703.10847

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

Yoo, K.-H., Gretzel, U., & Zanker, M. (2013). *Persuasive recommender systems*. Springer New York. https://doi.org/10.1007/978-1-4614-4702-3

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2, Pt.2), 1–27. https://doi.org/10.1037/h0025848

Zappei, J. (2019, September 7). AI as good as Mahler? Austrian orchestra performs symphony with twist. *AFP*. https://news.yahoo.com/ai-good-mahler-austrian-orchestra-performs-symphony-twist-102216114.html

Zhu, H., & Huberman, B. A. (2014). To switch or not to switch. *American Behavioral Scientist*, *58*(10), 1329–1344. https://doi.org/10.1177/0002764214527089

Zittrain, J. (2019, July 23). The hidden costs of automated thinking. *The New Yorker*. https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking

Apéndice A

Primera calibración de estímulos

Para llevar a cabo los Experimentos 1, 2 y 3 de esta tesis, fue necesario calibrar los estímulos que se usarían como candidatos políticos o de citas. Con este objetivo, realizamos un experimento previo que nos permitiera seleccionar fotografías de la base de datos pública de Bainbridge y colaboradores (2013).

Método

Reclutamos 38 participantes, con edades entre 31 y 45 años (M = 39.3, SD = 4.25), mediante mensajes de Whatsapp y correos electrónicos en español a grupos de personas conocidas. La invitación a participar contenía un enlace al experimento de calibración online creado en la plataforma Qualtrics.

De las 10.168 fotografías de la base de datos fotográfica de Bainbridge y colaboradores (2013), 2.222 imágenes contaban con etiquetas (como la edad o el atractivo). De ellas, 2.063 estaban marcadas como no famosas y, de este grupo, 1.695 estaban etiquetadas como personas de raza blanca, mientras que el resto se repartían con otras seis etiquetas de raza. Por este motivo, decidimos utilizar las imágenes clasificadas como de raza blanca por ser las más numerosas.

Aunque las imágenes ya habían sido evaluadas en atractivo, este etiquetado había sido realizado por una muestra demasiado pequeña (solo 12 personas), por lo que decidimos calibrar de nuevo 100 de estas fotografías (50 hombres y 50 mujeres). Las fotografías finales seleccionadas fueron aquellas clasificadas con edades entre los

26 y los 45 años, con 4 puntos o más de media y mediana en atractivo (sobre 5) y etiquetadas como de raza blanca.

Se explicó a los participantes que iban a evaluar su gusto por 50 candidatos anónimos de una supuesta página de citas, por lo que debían indicar si preferían valorar a hombres o mujeres. Aunque los estímulos iban a usarse tanto para simular el contexto de citas como el contexto político, consideramos que recogeríamos una medida de atractivo más real si el escenario que planteábamos a los participantes en este experimento de calibración era el de valorar a candidatos de citas.

Las fotografías solo se mostraron a los participantes durante 2 segundos para que su valoración de gusto o atractivo respondiera a una primera impresión, recogida en una escala del 1 al 10. A partir de las valoraciones medias de las imágenes, ordenamos los rostros de más atractivos a menos atractivos. El objetivo era utilizar los rostros mejor valorados como estímulos en los siguientes experimentos.

Resultados y Discusión

Este experimento permitió identificar los 100 rostros de la base fotográfica pública de Bainbridge, Isola y Oliva (2013) que cumplieran con los requisitos necesarios para los Experimentos 1, 2 y 3 de esta tesis. Las 16 imágenes (8 hombres y 8 mujeres) calibradas como las más atractivas fueron posteriormente utilizadas, de forma contrabalanceada, como los candidatos recomendados y candidatos control. El resto de fotografías fueron usadas como estímulos de relleno, escogiéndolas siempre en orden de más a menos atractivas.

Apéndice B

Segunda calibración de estímulos

En este experimento de calibración evaluamos 80 fotografías extraídas de la base de datos fotográfica de Karras y colaboradores (2018), y compuesta por 70.000 fotografías.

Método

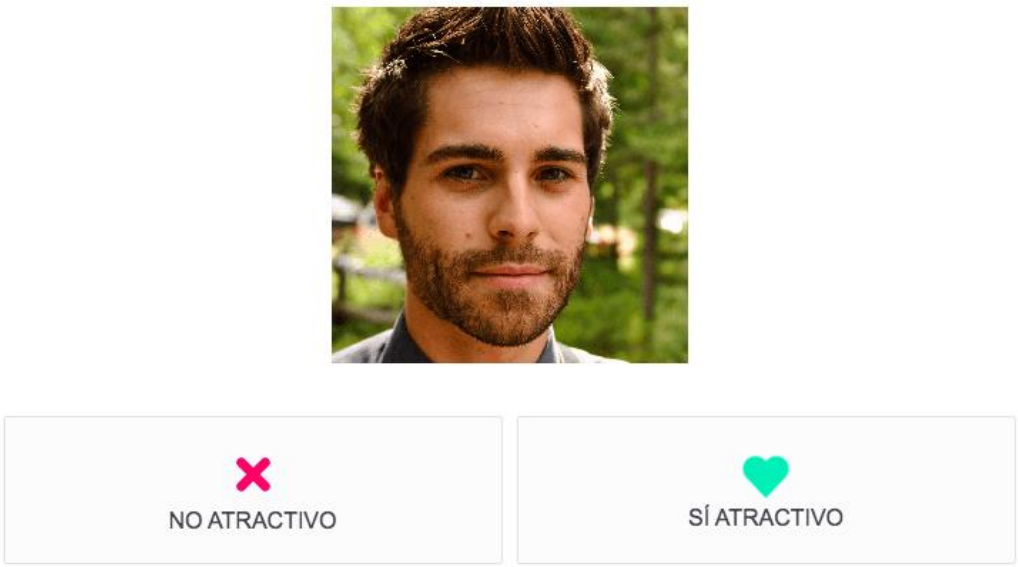
Reclutamos a 28 participantes (57.1% mujeres), con edades entre 26 y 46 años (*M* = 38.7, *SD* = 4.4), mediante mensajes de Whatsapp en español a círculos cercanos. En esta ocasión, determinamos calibrar 80 imágenes (40 hombres y 40 mujeres), las cuales fueron preseleccionadas manualmente por los experimentadores con base en varios criterios, dado que no se trataba de una base de datos etiquetada. Los criterios fueron: que los rostros mostrados aparentasen entre 25 y 45 años, que fueran de etnia caucásica y no famosos, siguiendo los criterios de los anteriores experimentos; que en las fotografías no aparecieran objetos o complementos que ocultaran parte del rostro (sombreros, gafas, micrófonos…); que el fondo de la imagen fuera neutro o estuviera desenfocado; y que los rostros se orientaran hacia la cámara.

El procedimiento seguido fue el siguiente. Tras la bienvenida y su aceptación a participar en el experimento, los participantes indicaron su edad y género, y visualizaron las 80 fotografías preseleccionadas, de una en una, en páginas diferentes. En esta ocasión todos los participantes valoraron todas las fotografías independientemente de su género. Por cada imagen, los participantes indicaron si consideraban el rostro mostrado como atractivo en una escala dicotómica (“No

atractivo", símbolo x / "Sí atractivo", símbolo corazón). Las fotografías fueron presentadas en orden aleatorio (véase Figura 16).

Figura 16

Ejemplo de Fotografía[15] *a Calibrar con Escala Dicotómica*



Al igual que en el experimento de calibración previo, una vez obtenidas las puntuaciones de todas las fotografías, ordenamos los rostros de más atractivo a menos atractivo para utilizar los mejor valorados como los candidatos diana y control en los Experimentos 4, 5, 6 y 7.

[15] Fotografía de la base de datos fotográfica de Karras y colaboradores (2018), con licencia CC BY 2.0, nombre DSC_3929.jpg y autor Jean-Simon Asselin. Accesible desde https://www.flickr.com/photos/acelain/4852007896

Resultados y Discusión

Esta calibración nos permitió utilizar imágenes con una apariencia más actual en los Experimentos 4, 5, 6 y 7. De nuevo, las 16 imágenes (8 hombres y 8 mujeres) calibradas como las más atractivas fueron posteriormente utilizadas, de forma contrabalanceada, como los candidatos diana y control. El resto de fotografías fueron usadas como estímulos extras. En esta ocasión, al no contar la base fotográfica con un etiquetado y valoración previa de los ítems, este experimento de calibración fue clave para determinar cuáles de las 80 fotografías seleccionadas eran las más atractivas y, por tanto, apropiadas para ser recomendadas por el algoritmo.