Publicado en: J. D. Gutiérrez & R. Manrique (Eds.), IA en acción: Desarrollo, impacto e implicaciones éticas en la sociedad. Bogotá: Ediciones Uniandes (2025).

Transparencia, explicabilidad y confianza en los sistemas de aprendizaje automático

Andrés Páez[1]

Resumen

Uno de los principios éticos mencionados más frecuentemente en los lineamientos para el desarrollo de la inteligencia artificial (IA) es la transparencia algorítmica. Sin embargo, no existe una definición estándar de qué es un algoritmo transparente ni tampoco es evidente por qué la opacidad algorítmica representa un reto para el desarrollo ético de la IA. También se afirma a menudo que la transparencia algorítmica fomenta la confianza en la IA, pero esta aseveración es más una suposición a priori que una tesis basada en evidencia empírica. Tampoco se discute mucho hasta qué punto es técnicamente posible volver transparente la caja negra de la IA a través de los métodos de explicabilidad. En este capítulo haré un análisis de la interrelación entre los conceptos de transparencia, explicabilidad y confianza. Inicialmente, analizaré los diferentes tipos de opacidad algorítmica para entender mejor cuál es el problema al que nos enfrentamos. En las secciones subsiguientes trazaré la relación entre explicabilidad y transparencia, y presentaré los límites de los métodos actuales de explicabilidad. En la sección final, examinaré la evidencia empírica acerca de la relación entre transparencia y confianza en las decisiones automatizadas basadas en sistemas de IA.

1. Introducción

La falta de transparencia de muchos algoritmos de inteligencia artificial (IA) es considerada uno de los principales obstáculos para su desarrollo ético. Una reciente revisión de los lineamientos que han sido propuestos en diferentes lugares del mundo para el desarrollo de la IA arrojó que la transparencia era el principio ético más comúnmente citado; aparece en 73 de los 84 lineamientos examinados (Jobin et al., 2019). La falta de transparencia mencionada en estos lineamientos en realidad se refiere a dos problemas diferentes. Por una parte, existe una preocupación por la *implementación* transparente de aquellos sistemas de IA que afectan directamente la vida y los

[1] Profesor Titular del Departamento de Filosofía e Investigador del Centro de Investigación y Formación en Inteligencia Artificial (CinfonIA) de la Universidad de los Andes, Bogotá, Colombia.

derechos de las personas (Coddou & Smart, 2021). La transparencia algorítmica en este sentido incluye elementos tan diversos como: informar a las personas que una decisión que las afecta está basada en una herramienta algorítmica, implementar mecanismos de rendición de cuentas y exigir garantías de que los algoritmos no tienen efectos discriminatorios, entre otros (Llamas et al., 2022; GPAI, 2024). Por otra parte, la transparencia se puede referir a la posibilidad de *comprender* el funcionamiento interno del algoritmo, la forma en que procesa los datos de entrada y llega a una predicción o una clasificación. El Reglamento General de Protección de Datos de la Unión Europea (GDPR), por ejemplo, requiere que se les proporcione a los usuarios "información significativa acerca de la lógica involucrada" en los sistemas de decisión automatizada (Regulación EU 2016/679, Articulo 13). En este capítulo sólo me ocuparé de este segundo sentido de transparencia[2].

La transparencia algorítmica en el segundo sentido puede verse opacada por dos razones diferentes. Por una parte, muchos algoritmos complejos, como las redes neuronales profundas, procesan en paralelo un volumen enorme de datos subsimbólicos a través de múltiples capas ocultas de nodos interconectados. Esta arquitectura hace que el funcionamiento interno del algoritmo sea epistémicamente inaccesible para cualquier ser humano, incluyendo sus diseñadores y desarrolladores. Sin embargo, la opacidad no siempre es el resultado de la complejidad técnica. En muchos casos el funcionamiento del algoritmo es escondido intencionalmente y protegido como un secreto industrial, especialmente en los sistemas de decisión automatizados de uso comercial. De ese modo se genera un tipo de opacidad totalmente diferente. En la primera parte del capítulo examinaremos estas dos formas diferentes de entender la opacidad algorítmica. Cada una de ellas generará retos éticos diferentes que tendremos que analizar.

Para intentar disminuir el efecto de la opacidad producto de la complejidad técnica de los algoritmos, en años recientes se han desarrollado métodos de explicabilidad e interpretabilidad que intentan ofrecer atisbos acerca de su funcionamiento. El desarrollo de estos métodos, del que se ocupa una subdisciplina de las ciencias de la computación conocida como la IA explicable (XAI, por sus siglas en inglés), se enfrenta a numerosos retos y no es evidente que vaya a haber grandes avances en la comprensibilidad de los algoritmos en el futuro cercano. Estas limitaciones de la

[2] Adicionalmente, vale la pena aclarar que solo me ocupare de sistemas de IA *discriminatorios*, es decir, aquellos que llevan a cabo tareas de clasificación y predicción. La transparencia en los sistemas de IA *generativa* requiere de una aproximación diferente a la ofrecida aquí.

XAI, que examinaremos en la segunda parte del capítulo, plantean un reto importante al ideal de transparencia que se persigue en los marcos éticos de la IA.

Decir que un algoritmo es transparente epistémicamente es equivalente a decir que es comprensible. Naturalmente, la comprensibilidad es una cuestión de grado, determinada en gran medida por el conocimiento de fondo que tengan los usuarios del algoritmo. Un sistema de IA puede ser medianamente transparente para su desarrollador y totalmente opaco para el usuario final. Para poder caracterizar adecuadamente la transparencia algorítmica es necesario entonces hacer un análisis pragmático —contextual y situado— de qué significa *comprender* un sistema de IA. Hay varios sentidos en los que se puede decir que se comprende un algoritmo. Por una parte, podemos decir que, a través de un método de XAI como LIME (Ribeiro et al., 2016), entendemos cuáles elementos del *input* fueron mayormente responsables de un *output* dado. Por otra parte, la comprensión también se puede referir a entender el funcionamiento global del algoritmo a través de un algoritmo sustituto más simple que lo represente. En la tercera parte del capítulo analizaremos estos sentidos del concepto de comprensión para poder aclarar la meta de los métodos de explicabilidad

La última tarea de este capítulo será analizar la relación entre transparencia y confianza, especialmente en el contexto de las decisiones automatizadas basadas en sistemas de IA. En las relaciones humanas, la confianza es un concepto ético en la medida en que les atribuimos honestidad, justicia y competencia a los demás y depositamos en ellos la responsabilidad por nuestro bienestar y el cumplimiento de nuestros objetivos teóricos y prácticos. En el caso de los sistemas de IA, se pretende que la explicabilidad nos proporcione una forma de juzgar la confiabilidad e imparcialidad de las decisiones del sistema. Sin embargo, la evidencia empírica muestra que la explicabilidad no siempre promueve la confianza, y en algunos casos incluso puede disminuirla. Aunque todavía falta estudiar más a fondo el problema, no debemos tomar como un dogma la idea de que la transparencia y la confianza en la IA van de la mano.

2. Dos tipos de opacidad algorítmica

La opacidad algorítmica se puede entender de dos maneras muy diferentes. Por una parte, se puede referir a modelos cuya estructura, *dataset*, características (*features*), pesos y sesgos son propiedad de una compañía privada o pública, y son tratados como secretos industriales protegidos por leyes mercantiles y de derechos de autor. Estos modelos no son necesariamente complejos, pero su

opacidad se deriva del hecho de que las personas afectadas por sus decisiones están legalmente impedidas para acceder a sus datos y funcionamiento. Por otra parte, la opacidad algorítmica se puede referir a algoritmos que escapan la comprensión humana debido a su complejidad extrema, lo cual los hace epistémicamente inaccesibles. Algunos autores, como Rudin (2019), han sugerido que cuando las decisiones de un algoritmo afectan significativamente la vida de las personas, sólo se deberían utilizar algoritmos transparentes. Sin embargo, es bien sabido que hay una relación inversa entre la transparencia y la precisión de un algoritmo. Los algoritmos más precisos, como las redes neuronales profundas, son también los más opacos. Sacrificar la precisión de las decisiones en aras de la transparencia podría tener efectos muy negativos sobre los usuarios. En esta sección discutiré estos dos tipos de opacidad algorítmica. Para facilitar la discusión, llamaré al primer tipo *opacidad jurídica* y al segundo, *opacidad epistémica*.

2.1 La opacidad jurídica

A medida que se extiende el uso de herramientas basadas en inteligencia artificial, tanto en la empresa privada como en las agencias del Estado, los modelos que utilizan se han convertido en bienes valiosos que deben ser protegidos. La IA está siendo utilizada para tomar decisiones importantes que afectan la vida de las personas en ámbitos como la salud, la vivienda, la educación, la asignación de subsidios, el crédito y el acceso al empleo. También comienza a verse su uso en los procesos penales, donde sirven para evaluar el riesgo de fuga o reincidencia de las personas imputadas, y en general en el sistema judicial en aplicaciones como la detección de la evasión fiscal y la vigilancia predictiva. En esta sección vamos a examinar las implicaciones éticas de algunos casos en los que se han utilizado sistemas de IA jurídicamente opacos.

Quizás el caso más conocido de opacidad jurídica es el del algoritmo COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*). Se trata de un algoritmo utilizado principalmente en el sistema judicial estadounidense para evaluar el riesgo de reincidencia de las personas en diferentes momentos del proceso penal. El algoritmo fue creado y le pertenece a una compañía privada llamada Northpointe, rebautizada como Equivant. En una entrevista, su gerente general afirmó: “La clave de nuestro producto son los algoritmos y son de nuestra propiedad. … Nosotros los creamos y no los hacemos públicos porque son una pieza esencial de nuestro negocio. No se trata de conocer los algoritmos. Se trata de conocer los resultados” (citado en Smith, 2016).

El algoritmo genera escalas de riesgo para la reincidencia general y violenta, que son usadas para tomar decisiones sobre la libertad condicional de las personas encarceladas. Otro algoritmo es utilizado para decidir si las personas imputadas pueden esperar su juicio en libertad, o si existe el riesgo de fuga, reincidencia o interferencia en el proceso. La naturaleza secreta del algoritmo se convirtió en el centro de una disputa jurídica que terminó en la decisión de *State v. Loomis* (2016), tomada por la Corte Suprema del estado de Wisconsin, en los Estados Unidos. La demanda contra la opacidad jurídica del algoritmo fue impuesta por un imputado llamado Eric Loomis, quien alegó que utilizar la herramienta de evaluación de riesgo proporcionada por COMPAS en la decisión sobre la pena que se le debía imponer violaba su derecho al debido proceso, pues se estaría violando su derecho a ser condenado con base en información precisa (*Loomis*, p. 757). La opacidad jurídica del algoritmo le impedía cuestionar su precisión y validez científica. También alegaba que COMPAS violaba su derecho al debido proceso porque la corte inconstitucionalmente tuvo en cuenta el género al tomar la decisión puesto que el algoritmo incluye el género como una de sus variables.

La Corte Suprema de Wisconsin reafirmó la decisión en contra de Loomis que había tomado el juez del caso. Los argumentos de Loomis acerca de la violación del debido proceso fueron rechazados porque COMPAS solo usa datos disponibles públicamente y datos proporcionados por el acusado. En consecuencia, concluyó la Corte, Loomis hubiera podido corregir la información, verificando de esa manera su precisión. Sobre el asunto del género, la Corte concluyó que el uso del género como un factor en la evaluación del riesgo sirve al propósito no discriminatorio de promover la precisión (pp. 766-767). La sentencia incluyó una nota de cautela para los jueces que utilicen herramientas de evaluación de riesgo y prescribió cómo se deben presentar las evaluaciones a las cortes y en qué medida las pueden utilizar (pp. 763-765).

Como lo han señalado varios autores (Freeman, 2016; HLR, 2017), la cuestión que *Loomis* nunca resolvió tenía que ver, paradójicamente, con el riesgo involucrado en utilizar algoritmos jurídicamente opacos al tomar decisiones judiciales. COMPAS y otros algoritmos opacos de evaluación de riesgo han sido criticados por reforzar desigualdades preexistentes, violar el derecho a la no discriminación con base en raza (Thomas & Pontón-Núñez, 2022), y disfrazar "discriminación evidente basada en demografía y estatus socioeconómico" (Starr, 2014, p. 806). COMPAS también ha sido acusado de ser menos preciso al evaluar acusados afrodescendientes (Angwin et al., 2016), aunque esta afirmación ha sido puesta en duda (Corbett-Davies et al., 2016).

También se ha encontrado que no se ajusta a algunas métricas de justicia algorítmica (Gursoy & Kakadiaris, 2022). En términos generales, las herramientas actuariales y algorítmicas sufren de problemas en la calidad de los datos y de incertidumbre en sus resultados (Páez, 2016), con lo cual parece factualmente irresponsable basar cualquier decisión en ellas.

Otro ejemplo bastante conocido de opacidad jurídica es el algoritmo conocido como SyRI (*System Risk Indication*), el cual fue desarrollado por el gobierno neerlandés en 2014 como una herramienta para evaluar el riesgo de evasión fiscal y abuso de subsidios. El sistema generaba perfiles individuales basados en datos personales que habían sido recolectados de diversos organismos públicos. En 2020, la Corte de Distrito de La Haya prohibió seguir usando SyRI por violar el Artículo 8 de la Convención Europea de Derechos Humanos (ECHR), el cual protege el derecho al respeto de la vida privada y familiar (Lazcoz & Castillo, 2020). Al igual que COMPAS, el algoritmo era jurídicamente opaco. En 2017, el Ministerio de Asuntos Sociales había decidido que los modelos de riesgo que utilizaba debían ser secretos. La justificación era que los potenciales infractores podrían adaptar su comportamiento si el Estado permitía el acceso al algoritmo de riesgo. Sin embargo, la gran mayoría de los factores en el modelo eran estáticos, es decir, imposibles de cambiar. En consecuencia, la justificación de la opacidad judicial era muy poco convincente. Varios estudios han mostrados que el uso de SyRI genera muchos de los mismos problemas éticos acerca de raza y clase que ya habían sido detectados en el caso de COMPAS debido a su efecto desproporcionado sobre grupos sociales desfavorecidos (Bekker, 2021; Rachovitsa & Johann, 2022).

En Colombia, el Estado usa un sistema algorítmico conocido como Sisbén IV para estimar los ingresos de las personas y determinar su elegibilidad a subsidios y programas sociales. El algoritmo base también es "reservado", es decir, es jurídicamente opaco, según lo establece el Artículo 70 de la Ley 2294 de 2023. La justificación de la reserva es la misma ofrecida por el gobierno neerlandés: evitar que los ciudadanos manipulen el algoritmo reportando información falsa. En este caso la reserva está un poco más justificada porque gran parte de la información que sirve de base para el algoritmo es autorreportada por los potenciales beneficiarios y se han detectado múltiples incidentes de fraude[3].

No existe un remedio fácil para la opacidad jurídica cuando ésta es innecesaria y/o éticamente problemática. Es poco probable que haya cambios drásticos en las leyes mercantiles y

[3] Agradezco a Juan David Gutiérrez por señalarme la existencia de este caso local de opacidad jurídica.

de protección de derechos de autor, y muchos sectores están interesados en diseñar e implementar este tipo de algoritmos. Desde mi perspectiva, el principal riesgo de la utilización indiscriminada de este tipo de algoritmos es que cada vez más reemplazarán la capacidad de juzgar de los humanos, pues es fácil caer en la tentación de aceptar acríticamente los puntajes y recomendaciones que generan. Varios estudios empíricos respaldan esta conclusión. Algunos resultados en economía comportamental y psicología social muestran que es psicológicamente difícil, y muy poco frecuente, actuar en contra de las recomendaciones de los algoritmos (Christin et al., 2015; Thaler, 1999). Estos mismos estudios muestran que los resultados generados por los algoritmos actúan como anclas para las decisiones humanas, eliminando la libre discreción en el proceso de evaluación. No hay ninguna garantía de que el acceso a los datos sobre el funcionamiento del algoritmo pueda contrarrestar esta tendencia, pero al menos habría una base para la atribución de responsabilidades en la medida en que se conocería el peso dado a cada una de las variables. Las personas encargadas de tomar las decisiones asumirían la responsabilidad de tomar decisiones con base en las variables potencialmente sesgadas o discriminatorias del algoritmo.

Finalmente, algunos autores han señalado la necesidad de que existan auditorías independientes de los datos de entrenamiento y de la estructura de los algoritmos jurídicamente opacos. Quizás la propuesta más elaborada en este sentido es la de Langenkamp et al. (2020), que solo puedo esbozar brevemente aquí[4]. Los autores introducen la idea de "reportes de transparencia algorítmica" que cubren cuatro categorías: (i) *Intención*: ¿cuál es el propósito del modelo?; (ii) *Dataset*: información sobre demografía, características y *dataset* de prueba; (iii) *Métricas*: medidas del desempeño del modelo, umbrales de prueba y definiciones de "justicia"; y (iv) *Aplicaciones*: cómo va a ser usado el modelo en la toma de decisiones. Estos reportes, que serían confidenciales, serían la base fáctica para cualquier reclamo acerca de la ausencia de medidas precautelares de parte de las compañías para evitar la discriminación (Páez, 2021a).

2.2 La opacidad epistémica

El segundo tipo de opacidad algorítmica no es el resultado de las acciones o decisiones de ningún individuo. Es, más bien, la consecuencia de la forma en que funcionan los algoritmos más

[4] Otras contribuciones importantes en esta misma dirección son: Gebru et al. (2018) y Mitchell et al. (2019).

complejos de inteligencia artificial. No todos los modelos de IA son epistémicamente opacos. Algunos usan arquitecturas simples, como árboles de decisión o funciones lineales sencillas, que no requieren conocimiento técnico para ser entendidas. Pero el funcionamiento de los algoritmos más sofisticados, con las capacidades predictivas más poderosas, escapan a la comprensión humana, como veremos a continuación.

Consideremos el caso de las redes neuronales profundas (DNNs, por sus siglas en inglés), que son el tipo más común de algoritmo epistémicamente opaco. Las DNNs son diseñadas para identificar correlaciones y patrones en los datos, muchos de los cuales no son simbólicos, lo cual los hace incomprensibles para los humanos. La red usa esos patrones y correlaciones para hacer las predicciones y clasificaciones para los cuales ha sido entrenada. Dentro de la red, los *inputs* pasan a través de múltiples capas ocultas de nodos o "neuronas" interconectados, cada uno de los cuales transforma los datos de diferentes maneras antes de pasarlos a la siguiente capa de nodos. Estas transformaciones a menudo incluyen operaciones no lineales, las cuales, al combinarse con las interacciones entre las neuronas a lo largo de las diferentes capas, le permiten a la red neuronal modelar límites de decisión complejos y multidimensionales. Incluso si pudiéramos ver todos los pesos y sesgos en la red[5], es decir, los parámetros que el modelo aprendió durante el entrenamiento, no sería claro cómo interpretarlos a la luz de las características del *input* original. Así, la opacidad de las DNNs no surge de las funciones no lineales mismas, sino de la forma en que son usadas en la red. Cada *input* de la red pasa por una serie de transformaciones complejas e interconectadas que hace imposible entender cómo el *input* se relaciona con el *output*. La naturaleza no lineal de las funciones de activación aumenta esa complejidad. En un sistema lineal, el efecto de cada *input* sobre el *output* puede ser considerado independientemente de los demás, pero en un sistema no lineal el efecto de cambiar un *input* depende del valor de todos los demás *inputs*. Esto hace que sea epistémicamente imposible entender cómo cada *input* influencia el *output*.

El segundo problema es que no hay manera de verificar cuáles parámetros están siendo usados en las capas ocultas de la DNN, y por lo tanto, cuál es el modelo que resultó del entrenamiento. Los modelos profundos a menudo tienen un número muy grande de óptimos con un grado de precisión predictiva semejante. A este estado de cosas se le conoce como el *problema*

[5] Los pesos determinan la fuerza de la conexión entre neuronas. Los sesgos son constantes asociadas a cada neurona que sirven como una forma de umbral, permitiendo que las neuronas se activen incluso cuando la suma ponderada de sus *inputs* no es suficiente para hacerlo por sí sola.

de la identificabilidad de modelos: "Un modelo es identificable si un conjunto de datos de entrenamiento lo suficientemente grande puede descartar todas las configuraciones posibles de los parámetros del modelo excepto una. Los modelos con variables latentes a menudo no son identificables porque podemos obtener modelos equivalentes intercambiando variables latentes entre sí" (Goodfellow et al., 2016, p. 284). Por lo tanto, es imposible verificar cuál de los muchos modelos equivalentes es el que generó un output en un caso particular. Sin poder identificar el modelo utilizado, es imposible "explicar" las predicciones del modelo. Por supuesto, existe una descripción verdadera del modelo, pero es inaccesible al conocimiento humano.

La opacidad epistémica es una característica problemática de los algoritmos por varias razones. Desde el punto de vista ético, plantea muchas de las mismas preguntas que la opacidad jurídica con respecto a la discriminación oculta y la violación de los derechos humanos. Desde el punto técnico, es un obstáculo para los desarrolladores que quieran mejorar el desempeño del modelo, y detectar y resolver sesgos y otros riesgos semejantes. Desde el punto de vista regulatorio, como vimos más arriba, el Reglamento General de Protección de Datos de la Unión Europea (GDPR), y otras regulaciones más recientes en otras jurisdicciones, requieren que la lógica de las decisiones automatizadas que afecten la vida de las personas de manera significativa debe ser conocida por los usuarios. Sigue siendo una pregunta abierta si los desarrolladores van a poder cumplir este mandato legal. La opacidad epistémica también es considerada a menudo un obstáculo para la generación de confianza en los modelos por parte de los usuarios, tesis que examinaré en detalle en la sección 5. Por estas y muchas otras razones, se han desarrollado diferentes métodos para eliminar, o al menos disminuir, la opacidad epistémica. En la siguiente sección estudiaremos la efectividad de estos métodos.

3. La XAI y los esfuerzos para restaurar la transparencia

La inteligencia artificial explicable (XAI, por sus siglas en inglés) es un programa de investigación en las ciencias de la computación que busca desarrollar métodos que provean algún grado de comprensión del funcionamiento de los modelos de aprendizaje automático. Las aproximaciones más comunes a la XAI son: (i) intentar explicar una predicción particular de un modelo encontrando los elementos del *input* responsables de ese *output*, o (ii) proporcionar una explicación global de cómo funciona el modelo y cuáles son sus capacidades a través de un modelo más simple. La primera aproximación usa métodos locales de interpretación *post hoc*, es decir, posteriores al

entrenamiento del modelo. Estos incluyen los sondeos contrafácticos (Wachter et al., 2018; Mothilal et al., 2020), y diferentes tipos de métodos de perturbación del *input*, como LIME (Ribeiro et al., 2016;), Grad-CAM (Selvaraju et al., 2017; Ancona et al., 2019), SHAP (Lundberg & Lee, 2017), TCAV (Kim et al, 2018), entre muchos otros (véase Ivanovs et al., 2021 para un estudio comprehensivo). La segunda aproximación está basada en el uso de modelos proxy, interpretativos o sustitutos. Las clases de modelos sustitutos más usados son las aproximaciones lineares o de gradiente, las reglas de decisión y los árboles de decisión. (Frosst & Hinton, 2017; Wu et al., 2018). Ninguna de las dos aproximaciones logra eliminar completamente la opacidad epistémica, como veremos en esta sección.

Durante mucho tiempo, los métodos locales de interpretación *post hoc* eran considerados —al menos dentro de la comunidad de desarrolladores— como el camino más prometedor para abrir la caja negra de la IA. Más recientemente, sin embargo, han sido objeto de muchas críticas debido a sus limitaciones y debilidades intrínsecas (Páez, 2024), y debido a la inescrutabilidad de las "explicaciones" que producen para no expertos y usuarios finales (Ehsan & Riedl, 2020). Quizás el problema más grave de los métodos locales es su bajo desempeño en diversas métricas de robustez. Lo ideal es que una alteración mínima del *input* no resulte en una explicación muy diferente del mismo *output*. Sin embargo, una transformación simple del *input*, o repetir el proceso de muestreo, pueden generar explicaciones muy diferentes. Kindermans et al. (2019) revelan que añadir un cambio (*shift*) constante a los datos de entrada, que es un paso simple y común de preprocesamiento que no afecta el desempeño del modelo, hace que muchos métodos locales de interpretación arrojen resultados equivocados. Slack et al. (2020) descubrieron la vulnerabilidad de LIME y SHAP a los ataques adversariales. Y Ghorbani et al. (2019) muestran cómo generar perturbaciones adversariales que producen *inputs* perceptualmente indistinguibles a los que se les asigna la misma etiqueta predictiva, y que sin embargo arrojan interpretaciones muy diferentes usando métodos locales de interpretación *post hoc*.

Otra limitación de los métodos locales de interpretación *post hoc* tales como los mapas de calor o de prominencia (*saliency maps*) es que carecen de precisión. Por ejemplo, Rajpurkar et al. (2017) propusieron un mapa de calor para explicar las predicciones de una red neuronal convolucional de uso médico. El método resalta en rojo las áreas de una placa de rayos equis que son más relevantes en el diagnóstico positivo de neumonía, y en azul las menos relevantes. Sin embargo, algunos autores han cuestionado la utilidad del método. Ghassemi et al. (2021), por

ejemplo, arguyen que incluso las partes más calientes del mapa contienen información útil e inútil, desde la perspectiva de un agente humano, y que simplemente localizar la región más caliente de la placa no revela exactamente cuál elemento en esa región fue el que el modelo consideró importante:

> Un médico clínico no puede saber si el modelo estableció apropiadamente que la presencia de una opacidad en un conducto de aire fue importante en la decisión, si la forma del borde del corazón o de la arteria pulmonar izquierda fueron el factor decisivo, o si el modelo se basó una característica inhumana, tal como el valor o la textura de un píxel que pudo estar más relacionado con el proceso de toma de la imagen que con la enfermedad subyacente (p. e746).

Más aún, la información proporcionada en el área caliente debe ser interpretada, abriendo la puerta de ese modo a las creencias previas del médico y al riesgo de que se cuele el sesgo de confirmación. La explicación también carece de cualquier tipo de justificación de por qué esa área en particular era más relevante que otras porque no existe conocimiento causal que le dé sustento a la explicación. Finalmente, el sesgo de automatización (Lyell & Coiera, 2017) puede llevar a una sobreestimación del desempeño del sistema y a un abandono de una actitud crítica frente a los resultados.

Los métodos contrafácticos no son inmunes al problema de la robustez. Al igual que los métodos de perturbación del input, los métodos contrafácticos pueden ser manipulados y pueden converger hacia explicaciones drásticamente diferentes usando pequeñas perturbaciones (Slack at al., 2021). Los métodos contrafácticos también dependen críticamente de métricas de cercanía pero no hay una forma fundamentada para decidir cuál métrica usar en un caso particular. Y al igual que los métodos de prominencia, la falta de una base causal adecuada puede generar explicaciones subóptimas e incluso completamente equivocadas (Chou et al., 2021).

Esta es apenas una pequeña muestra de los problemas a los que se enfrentan los métodos locales de interpretación. Su fragilidad e imprecisión son lo suficientemente graves como para recomendar combinarlos con los métodos globales de explicación. Sin embargo, estos tampoco son la panacea. Por ejemplo, para crear un modelo lineal que se asemeje funcionalmente a un modelo opaco, se necesita conocimiento experto para seleccionar las características que deben ser incluidas. Solo aquellas características que excedan un cierto umbral de correlación con las

predicciones deseadas deben ser usadas, pero existe el riesgo de que muchas características no muestren correlación alguna cuando son examinadas individualmente, y/o que su contribución sólo se pueda apreciar en combinación con otras características. La ventaja de los modelos lineares es que son usados frecuentemente en las ciencias sociales y naturales, incluyendo la medicina, lo cual los hace una herramienta conocida y aceptada por la mayoría de sus usuarios. Sin embargo, no siempre es posible encontrar modelos lineales sustitutos, especialmente cuando el modelo está basado en datos subsimbólicos, como en los modelos de visión por computador.

Los árboles de decisión, por otra parte, son usados como sustitutos en aquellos casos en los que la relación entre las características y las predicciones son lineales o cuando las características interactúan entre sí. También pueden ser expresados como reglas de decisión. Sin embargo, su naturaleza escalonada no los hace muy eficientes. También son muy sensibles a cualquier cambio en los datos de entrenamiento o a cualquier cambio en las características escogidas: un cambio en una bifurcación al comienzo de un árbol afecta al árbol completo.

Finalmente, muchos métodos globales de explicación que intentan preservar la precisión del modelo opaco original terminan generando "cajas grises". Por ejemplo, Xu et al. (2018), comprimieron una red neuronal profunda en una red neuronal superficial, pero esta última sigue siendo completamente opaca para un usuario no experto. Y cuando los modelos sustitutos son fáciles de entender —e.g., los árboles de decisión de Bastani & Bastani (2019) para valorar el riesgo de diabetes—, sufren de sobreajuste y pérdida de precisión en comparación con el modelo original. Por supuesto, se podría argüir que el propósito principal del modelo sustituto no es alcanzar un nivel de precisión similar al del modelo original —pues en ese caso el modelo original sería innecesario—, sino ayudar a los usuarios finales a obtener algún grado de comprensión de su funcionamiento. Una explicación funcional sencilla puede ayudar a los usuarios a entender las capacidades y limitaciones del modelo original para que puedan ajustar de ese modo sus expectativas. Por ejemplo, una explicación sencilla de cómo funciona un modelo de IA generativa como ChatGPT ayuda a sus usuarios a entender por qué no es confiable cuando se trata de información fáctica. Los árboles de decisión simples, las listas de decisión, los métodos basados en ejemplos, e incluso las explicaciones dialógicas pueden ser más útiles para hacer comprensible y transparente, en algún grado, el funcionamiento de un sistema de IA. Pero ¿qué significa exactamente que un método de XAI haga que un modelo sea *comprensible*? La siguiente sección estará dedicada a responder esta pregunta.

4. Explicabilidad y comprensión

Para explorar el concepto de comprensión debemos recurrir a la literatura filosófica al respecto. A primera vista, comprender o entender (usaré los dos términos indistintamente) una decisión específica de un sistema de IA a través de un método local de interpretación *post hoc*, y entender un modelo como un todo a través de un modelo sustituto, son dos estados mentales diferentes que requieren un análisis independiente. El primero parece corresponder a los que la literatura filosófica llama "entender por qué", mientras que el segundo parece corresponder a una "comprensión objetual", es decir, a entender el objeto como un todo. Ambos tipos de comprensión han sido ampliamente discutidos en la epistemología y la filosofía de la ciencia. A continuación examinaremos cada uno en más detalle.

La caracterización de la comprensión objetual en la literatura epistemológica coincide con el propósito de las explicaciones globales a través de modelos sustitutos. Zagzebski, por ejemplo, afirma que la comprensión "implica lograr ver relaciones de unas partes con otras partes y quizás la relación de las partes con el todo" (2009, p. 144). El tipo de relaciones que ella tiene en mente pueden ser espaciales, temporales o causales. Para Grimm (2011), por su parte, la comprensión global de un objeto complejo como el sistema metro de Nueva York es un caso de "saber cómo":

> Si "saber cómo" implica una aprehensión de cómo funciona una cosa, entonces parece seguirse de ello que el objeto del "saber cómo" debe estar constituido por una estructura que pueda ser manejada, esto es, que pueda ser manipulada para determinar cómo los diversos elementos de la cosa se relacionan entre sí y dependen los unos de los otros (p. 86).

Ambas descripciones asumen que la comprensión objetual requiere identificar las diversas partes de un objeto, poder describir sus interdependencias funcionales, y usar esa información para hacer inferencias útiles. Esto es justamente lo que ofrecen los modelos sustitutos, ya sea a través de ecuaciones lineales o reglas de asociación, o directamente en un árbol de decisión. Estos modelos pretenden proporcionar una versión simplificada del modelo original mostrando sus características (*features*) principales y las relaciones funcionales entre ellas y las decisiones del sistema. La meta es encontrar un proxy que restaure algún grado de transparencia, en el sentido descrito por Paul Humphreys: "En gran parte de los modelos estáticos, nuestra comprensión está basada en la habilidad de descomponer el proceso entre los *inputs* y los *outputs* del modelo en pasos modulares,

cada uno de los cuales es metodológicamente aceptable tanto individualmente como en combinación con los demás" (2004, p. 148). El nivel de descomposición en el caso de los modelos proxy o sustitutos está determinado por consideraciones pragmáticas. Una vez el usuario ha entendido las relaciones funcionales que le interesan para sus metas prácticas o epistémicas, se puede afirmar que el modelo se ha vuelto transparente para él. La transparencia es un concepto asociado con el éxito práctico y/o epistémico, y depende de lograr ver la estructura funcional del modelo opaco a través del modelo sustituto.

Por otra parte, la descripción filosófica de "entender por qué" encaja muy bien con los métodos locales de interpretación *post hoc*. *Entender* por qué pasó *p* no es equivalente a *saber* por qué *p*. Saber que un sistema de reconocimiento de imágenes clasificó correctamente una imagen como un perro porque le fue mostrada una foto de un perro claramente no es suficiente para entender la decisión del sistema. La persona debe poder responder una amplia variedad de preguntas contrafácticas del tipo: ¿qué hubiera pasado si las cosas fueran de otro modo? (Woodward, 2003). ¿Qué hubiera pasado si las orejas del perro no hubieran sido visibles? ¿O si la luz hubiera sido más tenue? ¿O si la hubiéramos mostrado una imagen espejo de la imagen original? Los métodos de interpretación local deben permitirles a los usuarios visualizar variaciones del input para poder resolver estas preguntas. Sólo su capacidad de responderlas puede demostrar que han entendido por qué el sistema hizo la clasificación correcta.

De hecho, muchos autores han argüido que no solo el "entender por qué", sino la comprensión *en general,* requieren de la habilidad de visualizar diferentes configuraciones de las partes de un objeto e inferir sus estados resultantes. Es decir, la comprensión en general requiere pensar contrafácticamente (de Regt & Dieks 2005; Wilkenfeld 2013). Como afirma Knuuttila, "la comprensión de lo posible es la manera de entender por qué emergió lo real y cómo funciona" (2011, p. 269). A pesar de las apariencias, yo considero que los métodos locales de interpretabilidad *post hoc* no proveen las herramientas para pensar contrafácticamente. Unas pocas manipulaciones del *input* solo les pueden dar a los usuarios una idea básica acerca de algunas correlaciones con las decisiones del modelo, pero éstas no pueden ser generalizadas fácilmente a casos similares. Lo que impide que estos métodos sean genuinamente contrafácticos es la falta de información causal acerca de cómo funciona el modelo como un todo, esto es, la falta de comprensión objetual. Los modelos sustitutos proporcionan las reglas generales que han sido extraídas de los datos o directamente del modelo, proporcionando de esa manera el soporte

funcional necesario para poder pensar contrafácticamente acerca de cualquier predicción. Por ejemplo, en un árbol de decisión uno puede seguir una rama u otra, y cada una de ellas será un caso contrafáctico cuyo resultado estará totalmente determinado por la estructura funcional estática representada en el árbol. El comprender por qué siempre requiere algún grado de comprensión objetual (Páez, 2019).

Karimi et al. (2020) ofrecen un argumento similar, pero desde una perspectiva más práctica. Ellos se enfocan en el problema del recurso algorítmico: cuando una persona ha sido afectada desfavorablemente por una decisión automatizada (e.g., le fue negado un crédito bancario), los métodos de XAI deberían poder sugerirle acciones posibles para mejorar o cambiar la decisión del sistema. Los autores se enfocan en uno de estos métodos, las explicaciones contrafácticas más cercanas desarrolladas por Wachter et al. (2018). Muestran que tales contrafácticos "no resultan en un conjunto de acciones óptimas o factibles que podrían cambiar favorablemente la predicción de *h* si fueran implementadas. Este defecto se debe principalmente a no considerar las relaciones causales que gobiernan el mundo" (Karimi et al., 2020, p. 359). La información causal faltante es parte del conocimiento teórico que hace parte de la comprensión objetual del sistema. Si bien es cierto que un modelo sustituto no proporciona por sí solo la información causal faltante, al menos proporcionan un conjunto robusto de correlaciones funcionales simbólicas que pueden ser investigadas y validadas empíricamente. Estas correlaciones pueden ser vistas como un paso inicial hacia la obtención del conocimiento causal requerido para diseñar un sistema de decisión verdaderamente operable (Buijsman, 2023).

En suma, los métodos locales de interpretabilidad *post hoc* por sí solos no pueden proporcionar una comprensión adecuada de un sistema de IA. Al proporcionar las causas de predicciones específicas, estos métodos pueden contribuir a establecer las interconexiones entre características y decisiones, pero solo la habilidad de un agente para razonar contrafácticamente sobre el modelo, solo su capacidad de usarlo y manipularlo, puede ser vista como evidencia de que lo ha comprendido, y por ende, de que se ha vuelto transparente en alguna medida para ese usuario. Los modelos sustitutos son quizás la mejor herramienta epistémica para que una amplia variedad de partes interesadas pueda entender el funcionamiento de los sistemas epistémicamente opacos. La pregunta final que quiero discutir es si la comprensión de un modelo obtenida a través de un método de XAI tiene como resultado un mayor nivel de confianza en sus predicciones.

5. Transparencia, explicabilidad y confianza en la IA

Existe la creencia generalizada de que la transparencia y la explicabilidad son condiciones necesarias para que los usuarios puedan confiar en los sistemas de IA. La *UNESCO Recommendation on the Ethics of Artificial Intelligence* (2021), por ejemplo, le dedica un capítulo entero a la transparencia y la explicabilidad, y afirma que estas "tienen como objetivo proveer de información apropiada a sus destinatarios respectivos para permitirles entender y para promover la confianza" (III, §39). En esa misma línea, el marco ético *AI4People* propuesto por Floridi et al. (2018) afirma que "es especialmente importante que la IA sea explicable, pues la explicabilidad es una herramienta crítica para generar confianza y comprensión hacia la tecnología" (p. 701). Los ejemplos de afirmaciones semejantes son numerosos[6]. Pero a pesar de la popularidad de la idea de que la transparencia promueve la confianza, tanto la evidencia disponible como la naturaleza misma de la confianza complican esta imagen tan sencilla, como veremos en esta sección final.

Antes de examinar la conexión entre explicabilidad y confianza, es importantes examinar la naturaleza de la segunda. En las relaciones humano-humano, la honestidad, la competencia y los valores compartidos son esenciales para establecer confianza tanto *cognitiva* como *emocional* (Gambetta, 1991).[7] La confianza cognitiva está basada en buenas razones racionales (Lewis & Weigert, 1985), en qué tanto conocemos a la persona en quien depositamos nuestra confianza, y en la evidencia sobre su confiabilidad. La confianza emocional, por su parte, está basada en los sentimientos positivos generados por nuestras interacciones con los demás. Es altamente contextual y depende de características sociales y culturales que no son fáciles de codificar. Los dos tipos de confianza son independientes entre sí. La gente a menudo confía en personas con las que no tiene ninguna conexión emocional si tiene evidencia clara de su competencia y habilidades. En otras ocasiones confía en personas que le generan sensaciones positivas, quizás con base en claves sociales compartidas, sin conocer sus capacidades cognitivas.

La honestidad, la competencia y los valores compartidos solo pueden ser atribuidos a los demás si les atribuimos además las intenciones y creencias de las cuales es posible inferir estos rasgos. En el caso de la IA, la honestidad y los valores compartidos son irrelevantes en gran

[6] Véase Kästner et al. (2021, p. 169) para una revisión de las opiniones favorables acerca de la conexión entre explicabilidad y confianza en la IA.

[7] En la literatura sobre la interacción entre humanos y robots, la confianza cognitiva y la emocional son llamadas "objetiva" y "subjetiva", respectivamente (Witkowski & Pitt, 2000; Witkowski et al., 2001; Tong et al., 2013).

medida, excepto quizás en el estudio de las relaciones humano-robot, donde es importante determinar si los usuarios les atribuyen a los robots estados mentales de los cuales se pueden inferir rasgos como la honestidad y otras intenciones (Páez, 2021b). En los demás contextos, la confianza en la IA se reduce a la competencia y la confiabilidad, esto es, a la confianza cognitiva. La literatura sobre la confianza en la IA la define como "el grado en el que la persona que confía cree que el sistema automatizado se comportará como se espera" (Papenmeier et al., 2022). Otra definición popular se enfoca en el desempeño del sistema: la confianza es "la voluntad de una de las partes de hacerse vulnerable a las acciones de la contraparte con base en la expectativa de que la segunda llevará a cabo una acción de importancia para la primera, independientemente de su habilidad para monitorearlo o controlarlo" (Mayer et al., 1995, p. 712).

Hay otro contexto en el que la honestidad y los valores compartidos son importantes, pero no como rasgos atribuidos a los sistemas de IA, sino más bien como contrapeso a las decisiones impersonales del sistema. Hay varios estudios que muestran que en ciertos contextos, como el ámbito médico, las personas tienden a preferir las decisiones tomadas por humanos, incluso cuando son menos precisas y confiables que las tomadas por sistemas automatizados (Ferrario & Loi, 2022; Longoni et al., 2019). En el contexto de los vehículos autónomos, la gente tiende a desconfiar de ellos incluso cuando las estadísticas indican que generan un menor número de accidentes (Hutson, 2017; Brenan, 2018). Dietvorst et al. (2014) llaman a este fenómeno "aversión algorítmica".

Una de las posibles explicaciones de la aversión algorítmica es que los seres humanos juzgamos de manera diferente a las máquinas y a nuestros congéneres. En un estudio reciente, Hidalgo y sus colaboradores compararon las reacciones de las personas a una amplia variedad de acciones llevadas a cabo por humanos y por máquinas. Concluyeron que, en general, "los humanos son juzgados por sus intenciones, mientras que las máquinas son juzgadas por sus resultados" (Hidalgo et al., 2021, p. 139). Un meta-análisis anterior de factores que afectan la confianza en las interacciones humano-robot (HRI) también reveló que "las características de los robots, en particular los factores relacionados con su desempeño, son la influencia más grande actualmente sobre la confianza percibida en HRI" (Hancock et al., 2011, p. 523). Este resultado se alinea muy bien con las definiciones de confianza que se encuentran en la literatura sobre sistemas multiagente. En el campo de la IA médica, Hatherley (2020) arguye que es un error usar categorías consideradas relevantes para la confianza interpersonal en las interacciones entre humanos y la IA

médica. Es posible decir que uno depende de estos sistemas, pero no parecen ser el tipo de objetos en los que uno confía. En esta misma línea, Ferrario et al. (2020) afirman que la confianza que los médicos tienen en los sistemas de IA no requiere monitorearlos con respecto a propiedades que solo los humanos pueden tener. El meta-análisis de Hancock et al. (2011) también encontró que los factores relacionados con actitudes humanas hacia los robots tenían un papel menor en la construcción de confianza.

Parece, por lo tanto, que el desempeño del sistema es el principal factor en la construcción de confianza hacia las máquinas, y que a menudo no es suficiente, como lo muestra la evidencia en el campo de la IA médica. La pregunta es si la explicabilidad puede ser agregada como un factor que complemente al desempeño como una fuente de confianza. La respuesta es que la evidencia acerca de la utilidad de la explicabilidad para este propósito no es concluyente. Algunos estudios parecen indicar que tiene un efecto positivo. Shin (2021) hizo un estudio con 350 individuos que usaban regularmente servicios automatizados de noticias. Sus resultados indican que la transparencia y la explicabilidad impactan positivamente la confianza de los usuarios. En el área de la IA de apoyo en decisiones clínicas, Liu et al. (2022) y Wysocki et al. (2023) reportaron que la transparencia y la explicabilidad fueron efectivas en la construcción de confianza entre el personal médico, aunque acentuaba el sesgo de confirmación y el exceso de confianza en el modelo.

A pesar de estos resultados positivos, la evidencia que muestra la inefectividad de la explicabilidad es mucho más extensa y convincente, incluso en los mismos sectores y en contextos similares. Papenmeier et al. (2019) encontraron que las explicaciones de alta fidelidad de hecho *disminuían* la confianza de los usuarios en varios algoritmos de clasificación altamente precisos utilizados en redes sociales. Schmidt et al. (2020) también reportaron que un mayor grado de transparencia en un algoritmo de clasificación de texto puede tener un impacto negativo sobre la confianza. Más aún, los autores afirman que "este efecto ocurre predominantemente en aquellos casos en los que las predicciones del sistema de aprendizaje automático son correctas, mostrando de este modo que el uso descuidado de la transparencia en herramientas de asistencia basadas en IA puede de hecho desmejorar el desempeño humano" (p. 261). Estos son algunos de los muchos ejemplos en los que la investigación empírica no ha encontrado ningún soporte para la hipótesis

de la relación positiva entre explicabilidad y confianza[8]. Kästner et al. (2021) creen que hay tres razones por las que las explicaciones fallan en la promoción de la confianza:

> 1) Si la confianza que una persona le tiene al sistema ya está en su grado máximo, una explicación no puede aumentarla; 2) si la explicación revela un problema en el sistema, la explicación puede disminuir en lugar de aumentar la confianza; 3) Si una persona no puede comprender la explicación o no puede usarla para evaluar el sistema, es posible que la explicación no cambie su confianza en el sistema (p. 2).

Esta revisión de literatura indica que existe una aguda controversia en torno a la efectividad de la explicabilidad en la construcción de confianza. Hay una premisa implícita en esta discusión, y es la creencia de que la confianza en la IA es un fin deseable. Es evidente que no confiar en un sistema de decisión automatizado que sea útil, explicable y de gran desempeño parece irracional, incluso antiético, si la falta de confianza impide que la gente reciba sus beneficios sin un riesgo significativo. Sin embargo, hay voces que invitan a la cautela. Peters y Visser (2023) nos advierten acerca del exceso de confianza en el modelo y recomiendan una dosis saludable de desconfianza. Ghassemi et al. (2021) también advierten acerca del peligro de usar explicaciones superficiales o poco confiables en las aplicaciones en el sector salud, que pueden generar falsas esperanzas en el poder de la IA y pueden llevar al sesgo de automatización. Los autores llegan al punto de advertir que la explicabilidad no debería ser un requisito de los modelos utilizados en el ámbito clínico. Lakkaraju y Bastani (2020) también advierten que no se deben usar métodos de XAI que solo optimicen la fidelidad, esto es, que solo se preocupen por encontrar explicaciones que dupliquen correctamente las predicciones del modelo de caja negra, porque la fidelidad se puede obtener incluso si las explicaciones usan características completamente diferentes a las utilizadas por el modelo original. Las explicaciones de alta fidelidad "pueden incluso engañar al tomador de decisiones y hacerlo confiar en una caja negra problemática" (p. 79). Incluso cuando los métodos de XAI mejoran los reportes de confianza y comprensión de un sistema de IA, hay evidencia de que esos reportes no se traducen en una mejora en el desempeño en tareas que se apoyan en el sistema (Kandul et al., 2023; Papenmeier et al., 2022).

[8] Otros ejemplos incluyen: Chen et al. (2019), Cheng et al. (2019), Kizilsec (2016), Langer et al. (2018), y Langer et al. (2021), y las referencias incluidas en cada uno.

Finalmente, existe el riesgo ético de promover métodos de explicabilidad que promuevan la confianza *emocional* a través de interfaces amigables que provean racionalizaciones *post hoc* falsas, pero persuasivas, de decisiones complejas. Esto tiende a ocurrir en el campo de la robótica social, donde la transparencia algorítmica puede interferir con la integración del robot en su entorno social. Danaher (2020) ha llamado a estos sistemas autónomos, "robots engañosos". ¿Deberíamos descalificarlos como herramientas útiles debido a su naturaleza engañosa? Los seres humanos frecuentemente inventamos racionalizaciones *post hoc* falsas para justificar las cosas que hacemos (Nisbett & Wilson, 1977). ¿Por qué las aceptamos en el caso de los humanos pero no en el de las máquinas? Zerilli et al. (2019) han expresado su preocupación de que "las decisiones automatizadas están siendo sometidas a estándares altos poco realistas, probablemente debido a un estimado muy alto y poco realista del grado de transparencia que es posible alcanzar en el caso de los decisores humanos" (p. 661). Isaac y Bridewell (2017) defienden la idea de que el engaño de los robots es aceptable cuando lo hacen en aras de un bien superior, incluyendo el bien de la integración social fluida. ¿Deberíamos entonces tolerar el mismo grado de falta de sinceridad que encontramos en las relaciones entre humanos? Los robots frecuentemente tienen que tomar decisiones con consecuencias significativas que requieren de explicaciones complejas que no pueden ser empaquetadas en formatos estandarizados. El uso de outputs gráficos en tiempo real para representar los estados internos del proceso de toma de decisiones que ocurre dentro del robot es una forma prometedora para alcanzar la transparencia robótica (Wortham et al., 2017; Edmonds et al., 2019). Esta aproximación no requiere apelar a la confianza emocional de las personas. Al contrario; al hacer explícita la arquitectura jerárquica del software del robot, es más fácil pensar en él como un ser sin estados mentales humanos. En suma, en situaciones de alto riesgo para las personas es deseable que la confianza cognitiva prime sobre la emocional.

6. Comentarios finales

A menudo se da por sentado que diseñar sistemas transparentes o explicables debe ser una meta del desarrollo responsable de la IA. Una de las principales razones que se aducen para ese desiderátum es que es deseable que la gente confíe en esos sistemas. El análisis presentado en este capítulo muestra no solo que la conexión entre explicabilidad y confianza no es obvio, sino también que el grado de transparencia al que podemos aspirar en este momento, dado el estado actual de los métodos de explicabilidad, es muy limitado. Esto ha llevado a que algunos investigadores

adopten una actitud escéptica acerca de la posibilidad de comprender los sistemas de IA. Humphreys, por ejemplo, arguye que "debemos abandonar la insistencia en la transparencia epistémica para las ciencias de la computación". En lugar de transparencia, continúa Humphreys, podemos alcanzar las virtudes de los modelos computacionales "a través de procedimientos de prueba y error, tratando las conexiones entre la plantilla computacional y sus soluciones como una caja negra" (2004, p. 150). Otros se preguntan si el uso cada vez más extendido de la IA en las ciencias "ejemplifica un cambio de paradigma que nos aleja del propósito explicativo tradicional de la ciencia, y nos acerca hacia el reconocimiento de patrones y la predicción" (Boge & Poznic, 2021, p. 171). Yo quisiera resistir estos llamados a abandonar el proyecto de hacer que la IA sea comprensible. Hay razones epistémicas, éticas y jurídicas —es decir, razones normativas— para continuar desarrollando la IA explicable. En consecuencia, la discusión acerca de las posibilidades de la XAI no debe estar limitada a sus limitaciones técnicas. También es una discusión filosófica acerca de la transparencia de una tecnología utilizada para tomar decisiones que afectan en gran medida la vida de las personas, y en ese sentido es una discusión ética de comienzo a fin.

Referencias

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2019). Gradient-based attribution methods. En Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.), *Explainable AI: interpreting, explaining, and visualizing deep learning* (169-191). Springer Nature.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May). *Machine bias*. ProPublica. Retrieved from https://www.propublica.org/article/ machine-bias-risk-assessments-in-criminal-sentencing

Bastani, O., Kim, C., & Bastani, H. (2017). Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504*.

Bekker, S. (2021). Fundamental rights in digital welfare states: The case of SyRI in the Netherlands. En O. Spijkers, W. G. Werner, & R. A. Wessel (Eds.), *Netherlands Yearbook of International Law 2019* (pp. 289-307). Springer.

Boge, F. J., & Poznic, M. (2021). Machine learning and the future of scientific explanation. *Journal for General Philosophy of Science*, *52*(1), 171-176.

Brenan, M. (2018). Driverless cars are a hard sell to Americans. *Gallup*, May 15, 2018. Retrieved from https://news.gallup.com/poll/234416/driverless-cars-tough-sell-americans.aspx.

Buijsman, S. (2023). Causal scientific explanations from machine learning. *Synthese*, *202*(6), 202.

Chen, L., Yan, D., & Wang, F. (2019). User evaluations on sentiment-based recommendation explanations. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *9*(4), 1-38.

Cheng, H. F., Wang, R., Zhang, Z., O'connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019, May). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. En *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-12). New York: ACM.

Christin, A., Rosenblat, A., & Boyd, D. (2015). Courts and predictive algorithms. *Data and Civil Rights: A New Era of Policing and Justice.* http://www.datacivilrights.org/pubs/2015-1027/Courts_and_Predictive_Algorithms.pdf

Coddou, A., & Smart, S. (2021). La transparencia y la no discriminación en el Estado de bienestar digital. *Revista Chilena de Derecho y Tecnología*, *10*(2), 301-332.

Corbett-Davies, S., Pierson, E., Feller, A., & Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post*, October 17, 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas

Danaher, R. (2020). Robot betrayal. A guide to the ethics of robot deception. *Ethics and Information Technology*, *22*, 117–128.

De Regt, H. W., & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, *144*, 137–170.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126.

Edmonds, M., Gao, F., Liu, H., Xie, X., Qi, S., Rothrock, B., Zhu, Y., Wu, Y.N., Lu, H., & Zhu, S. C. (2019). A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics*, *4*(37), eaay4663.

Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable AI: Towards a reflective sociotechnical approach. En *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22* (pp. 449-466). Springer.

Ferrario, A., & Loi, M. (2022, June). How explainability contributes to trust in AI. En *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1457-1466). New York: ACM.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*, 689–707.

Freeman, K. (2016). Algorithmic injustice: How the Wisconsin Supreme Court failed to protect due process rights in State v. Loomis. *North Carolina Journal of Law & Technology*, *18*(5), 75–106.

Frosst, N., & Hinton, G. (2017). Distilling a neural network into a soft decision tree. *arXiv preprint:1711.09784*.

Gambetta, D. (1988) (Ed.). *Trust: Making and breaking cooperative relations*. Oxford: Basil Blackwell.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745-e750.

Ghorbani, A., Abid, A., & Zou, J. (2019, July). Interpretation of neural networks is fragile. En *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 3681-3688).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

GPAI (2024). *Algorithmic transparency in the public sector: A state-of-the-art report of algorithmic transparency instruments*. Report, May 2024, Global Partnership on Artificial Intelligence.

Grimm, S. R. (2011). Understanding. En S. Bernecker, & D. Pritchard (Eds.), *The Routledge companion to epistemology* (pp. 84–94). New York: Routledge.

Gursoy, F., & Kakadiaris, I. A. (2022, November). Equal confusion fairness: Measuring group-based disparities in automated decision systems. En *2022 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 137-146). IEEE.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527.

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics*, *46*(7), 478–481.

Hidalgo, C. A., Orghiain, D., Canals, J. A., De Almeida, F., & Martín, N. (2021). *How humans judge machines*. Cambridge: MIT Press.

HLR (Harvard Law Review). (2017). State v. Loomis: Wisconsin Supreme Court requires warning before use of algorithmic risk assessments in sentencing. *Harvard Law Review*, *130*, 1530-1537.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. New York: Oxford University Press.

Hutson, M. (2017). A matter of trust. *Science*, *358*, 1375-1377.

Isaac, A. M. C., & Bridewell, W. (2017). White lies on silver tongues: Why robots need to deceive (and how). En P. Lin, R. Jenkins, & K. Abney (Eds,), *Robot ethics 2.0: From autonomous cars to artificial intelligence*. Oxford: Oxford University Press.

Ivanovs, M., Kadikis, R., & Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, *150*, 228-234.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Kandul, S., Micheli, V., Beck, J., Kneer, M., Burri, T., Fleuret, F., & Christen, M. (2023). Explainable AI: A Review of the Empirical Literature. *Available at SSRN 4325219*.

Karimi, A. H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. En *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 353-362). New York: ACM.

Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., & Sterz, S. (2021, September). On the relation of trust and explainability: Why to engineer for trustworthiness. En *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)* (pp. 169-175). IEEE.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018, July). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). En *International Conference on Machine Learning* (pp. 2668-2677). PMLR.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S, Erhan, D., & Kim, B. (2019). The (un)reliability of saliency methods. En W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K. R. Müller (Eds.), *Explainable AI: Interpreting, explaining, and visualizing deep learning* (pp. 267–280). Cham: Springer.

Kizilcec, R. F. (2016, May). How much information? Effects of transparency on trust in an algorithmic interface. En *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2390-2395). New York: ACM.

Knuuttila, T. (2011). Modelling and representing: An artefactual approach to model-based representation. *Studies in History and Philosophy of Science Part A*, *42*(2), 262-271.

Lakkaraju, H., & Bastani, O. (2020, February). "How do I fool you?" Manipulating user trust via misleading black box explanations. En *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 79-85). New York: ACM.

Langenkamp, M., Costa, A., & Cheung, C. (2020). Hiring fairly in the age of algorithms. *arXiv preprint arXiv:2004.07132*.

Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D., & Speith, T. (2021). Spare me the details: How the type of information about automated interviews influences applicant reactions. *International Journal of Selection and Assessment*, *29*(2), 154-169.

Langer, M., König, C. J., & Fitili, A. (2018). Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection. *Computers in Human Behavior*, *81*, 19-30.

Lazcoz, G., & Castillo, J. A. (2020). Valoración algorítmica ante los derechos humanos y el Reglamento General de Protección de Datos: el caso SyRI. *Revista Chilena de Derecho y Tecnología*, *9*(1), 207-225.

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social Forces*, *63*(4), 967-985.

Liu, C. F., Chen, Z. C., Kuo, S. C., & Lin, T. C. (2022). Does AI explainability affect physicians' intention to use AI?. *International Journal of Medical Informatics*, *168*, 104884.

Llamas, Z. J., Mendoza, O. A., & Graff, M. (2022). Enfoques regulatorios para la Inteligencia Artificial (IA). *Revista Chilena de Derecho*, *49*(3), 31-62.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, *24*(2), 423-431.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. En *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). New York: ACM.

Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. En *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 607-617).

Nisbett, R. E., & Wilson T. D. (1977): Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231-259.

Páez, A. (2016). The prediction of future behavior: The empty promises of expert clinical and actuarial testimony. *Teoria Jurídica Contemporânea*, *1*, 75-101.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), 441–459.

Páez, A. (2021a). Negligent algorithmic discrimination. *Law and Contemporary Problems*, *84*(3), 19-33.

Páez, A. (2021b). Robot mindreading and the problem of trust. *AISB Convention 2021: Communication and Conversation* (pp. 140-143). AISB.

Páez, A. (2024). Understanding with toy surrogate models in machine learning. *Minds and Machines*, *34*(4), 45.

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*.

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *29*(4), 1-33.

Peters, T. M., & Visser, R. W. (2023). The importance of distrust in AI. En L. Longo (Ed.), *Explainable artificial intelligence: First world conference, xAI 2023* (pp. 301-317). Cham: Springer.

Rachovitsa, A., & Johann, N. (2022). The human rights implications of the use of AI in the digital welfare state: Lessons learned from the Dutch SyRI case. *Human Rights Law Review*, *22*(2), 1–15.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). New York: ACM.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206-215.

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260-278.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. En *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020, February). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. En *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180-186).

Slack, D., Hilgard, A., Lakkaraju, H., & Singh, S. (2021). Counterfactual explanations can be manipulated. *Advances in Neural Information Processing Systems*, *34*, 62-75.

Smith, M. (2016). In Wisconsin, a backlash against using data to foretell defendants' futures. *The New York Times*, June 22, 2016. https://www.nytimes.com/2016/06/23/us/backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, *66*, 803-872.

State v. Loomis (2016). 881 N.W.2d 749 (Wis. 2016), cert. denied, 137 S. Ct. 2290 (2017).

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, *12*(3), 183-206.

Thomas, C., & Pontón-Núñez, A. (2022). Automating judicial discretion: How algorithmic risk assessments in pretrial adjudications violate equal protection rights on the basis of race. *Minnesota Journal of Law & Inequality*, *40*(2), 371-407.

Tong X., Zhang W., Long Y., & Huang H. (2013). Subjectivity and objectivity of trust. En: *International Workshop on Agents and Data Mining Interaction. ADMI 2012* (pp. 105-114). Berlin: Springer.

UNESCO (2021). *Recommendation on the ethics of artificial intelligence.* Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000381137.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841–887.

Wilkenfeld, D. (2013). Understanding as representation manipulability. *Synthese*, *190*, 997–1016.

Witkowski, M., & Pitt, J. (2000). Objective trust-based agents: Trust and trustworthiness in a multi-agent trading society. *Proceedings of the Fourth International Conference on MultiAgent Systems* (pp. 463-464). Boston: IEEE.

Witkowski, M., Artikis, A., & Pitt, J. (2001). Experiments in building experiential trust in a society of objective-trust based agents. In *Trust in Cyber-societies* (pp. 111-132). Berlin: Springer.

Woodward, J. (2003). *Making things happen. A theory of causal explanation*. New York: Oxford University Press.

Wortham, R. H., Theodorou, A., & Bryson, J. J. (2017). Robot transparency: Improving understanding of intelligent behaviour for designers and users. En In Y. Gao, S. Fallah, Y. Jin, & C. Lakakou (Eds.), *Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017, Proceedings* (pp. 274-289). Berlin: Springer.

Wu, M., Hughes, M. C., Parbhoo, S., Zazzi, M., Roth, V., & Doshi-Velez, F. (2018). Beyond sparsity: Tree regularization of deep models for interpretability. *arXiv preprint: 1711.06178v1.*

Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2023). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*, *316*, 103839.

Xu, K., Park, D. H., Yi, C., & Sutton, C. (2018). Interpreting deep classifier by visual distillation of dark knowledge. *arXiv preprint arXiv:1803.04042*.

Zagzebski, L. (2009). *On epistemology*. Belmont: Wadsworth.

Zerilli, J., Knott, A., Maclaurin, J., & Gavagahn, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, 32, 661-683.