Un estudio multidisciplinario y un marco para el diseño y Evaluación de sistemas de IA explicables

SINA MOHSENI y NILOOFAR ZAREI, Universidad Texas A&M, EE. UU .
ERIC D. RAGAN, Universidad de Florida, EE. UU.

La necesidad de sistemas inteligentes interpretables y responsables crece junto con la prevalencia de las aplicaciones de inteligencia artificial (IA) utilizadas en la vida cotidiana. Los sistemas de IA explicables (XAI) están destinados a autoexplicar el razonamiento detrás de las decisiones y predicciones del sistema. Investigadores de diferentes disciplinas trabajan juntos para definir, diseñar y evaluar sistemas explicables. Sin embargo, los académicos de diferentes disciplinas se centran en diferentes objetivos y temas bastante independientes de la investigación de XAI, lo que plantea desafíos para identificar la metodología adecuada de diseño y evaluación y consolidar el conocimiento entre los esfuerzos. Con este fin, este artículo presenta una encuesta y un marco destinados a compartir conocimientos y experiencias sobre los métodos de diseño y evaluación de XAI en múltiples disciplinas. Con el objetivo de respaldar diversos objetivos de diseño y métodos de evaluación en la investigación de XAI, después de una revisión exhaustiva de artículos relacionados con XAI en los campos de aprendizaje automático, visualización e interacción persona-computadora, presentamos una categorización de los objetivos de diseño y métodos de evaluación de XAI. Nuestra categorización presenta el mapeo entre los objetivos de diseño para diferentes grupos de usuarios de XAI y sus métodos de evaluación. A partir de nuestros hallazgos, desarrollamos un marco con pautas de diseño paso a paso combinadas con métodos de evaluación para cerrar los ciclos iterativos de diseño y evaluación en equipos multidisciplinarios de XAI. Además, proporcionamos tablas resumidas listas para usar de métodos de evaluación y recomendaciones para diferentes objetivos en la investigación de XAI.

Conceptos CCS: • Computación centrada en el ser humano → Métodos de diseño y evaluación de HCI; • Metodologías informáticas → Aprendizaje automático

Palabras y frases clave adicionales: inteligencia artificial explicable (XAI), interacción persona-computadora (HCI), aprendizaje automático, explicación, transparencia

Formato de referencia ACM:

Sina Mohseni, Niloofar Zarei y Eric D. Ragan. 2021. Encuesta multidisciplinaria y marco para el diseño y evaluación de sistemas de IA explicables. Transmisión ACM. Interactuar. Intel. Sistema. 11, 3-4, artículo 24 (agosto 2021), 45 páginas.
https://doi.org/10.1145/3387166

La revisión de este artículo estuvo a cargo de los editores asociados del número especial Shixia Liu, Daniel Archambault, Tatiana von Landesberger y Remco ChangCagatay Turkay.
El trabajo de este artículo está respaldado por el programa DARPA XAI según N66001-17-2-4031 y por el premio NSF 1900767.
Direcciones de los autores: S. Mohseni y N. Zarei, B208 Langford Building, 3137 TAMU, College Station, TX 77840; correo electrónico: {sina.mohseni, n.zarei.3001}@tamu.edu; ED Ragan, Universidad de Florida, Edificio E301 CSE, Gainesville, FL 32611; correo electrónico: eragan@ufl.edu.

2160-6455/2021/08-ART24 $15.00
https://doi.org/10.1145/3387166

1. INTRODUCCIÓN

Impresionantes aplicaciones de la Inteligencia Artificial (IA) y el aprendizaje automático se han vuelto predominantes en nuestro tiempo. Gigantes tecnológicos como Google, Facebook y Amazon han recopilado y analizado suficientes datos personales a través de teléfonos inteligentes, dispositivos de asistencia personal y redes sociales que pueden modelar a los individuos mejor que otras personas. La reciente interferencia negativa de los robots de redes sociales en elecciones políticas [91, 212] fueron otra señal de cuán susceptibles son nuestras vidas al mal uso de IA y big data [163]. En estas circunstancias, a pesar de los gigantes tecnológicos y la sed de tecnologías más avanzadas. sistemas, otros sugieren posponer el uso total de la IA para aplicaciones críticas hasta que puedan ser mejor comprendidos por quienes confiarán en ellos. La demanda de servicios predecibles y responsables La IA crece a medida que se le confían cada vez más tareas con mayor sensibilidad e impacto social servicios. Por lo tanto, la transparencia del algoritmo es un factor esencial para responsabilizar a las organizaciones. y responsables de sus productos, servicios y comunicación de información.

Los sistemas de inteligencia artificial explicable (XAI) son una posible solución hacia la responsabilidad IA, que lo hace posible al explicar los procesos y la lógica de toma de decisiones de la IA para los usuarios finales [72]. Específicamente, los algoritmos explicables pueden permitir el control y la supervisión en caso de efectos adversos o no deseados, como una toma de decisiones sesgada o discriminación social. Un sistema XAI puede ser Definido como un sistema inteligente que se explica por sí mismo y que describe el razonamiento detrás de sus decisiones. y predicciones. Las explicaciones de IA (ya sea explicaciones bajo demanda o en forma de descripción del modelo) podrían beneficiar a los usuarios de muchas maneras, como mejorar la seguridad y la equidad al confiar. sobre las decisiones de la IA.

Si bien el impacto cada vez mayor de los sistemas avanzados de aprendizaje automático de caja negra en el sector de big data La era ha atraído mucha atención de diferentes comunidades, la interpretabilidad de los sistemas inteligentes. También se ha estudiado en numerosos contextos [69, 167]. El estudio de agentes personalizados, sistemas de recomendación y tareas críticas de toma de decisiones (por ejemplo, análisis médicos, control de redes eléctricas) ha sumado a la importancia de la explicación del aprendizaje automático y la transparencia de la IA para los usuarios finales. Para Por ejemplo, como un paso hacia este objetivo, se ha establecido el derecho legal a explicaciones en el Comisión del Reglamento General de Protección de Datos (GDPR) de la Unión Europea . Mientras que la corriente estado de la normativa se centra principalmente en la protección y privacidad de los datos de los usuarios, se espera que cubra Más transparencia algorítmica y requisitos de explicaciones de los sistemas de IA [67].

Claramente, abordar una gama tan amplia de definiciones y expectativas para XAI requiere esfuerzos de investigación multidisciplinarios, ya que las comunidades existentes tienen requisitos diferentes y a menudo tienen prioridades y áreas de especialización drásticamente diferentes. Por ejemplo, la investigación en el ámbito de El aprendizaje automático busca diseñar nuevos modelos interpretables y explicar modelos de caja negra con explicadores ad hoc. En la misma línea pero con enfoques diferentes, los investigadores en analítica visual herramientas y métodos de diseño y estudio para que los expertos en datos y dominios visualicen cajas negras complejas modelos y estudian interacciones para manipular modelos de aprendizaje automático. En cambio, la investigación en La interacción persona-computadora (HCI) se centra en las necesidades del usuario final, como la confianza del usuario y la comprensión de las explicaciones generadas por la máquina. La investigación en psicología también estudia los fundamentos de la comprensión humana, la interpretabilidad y la estructura de las explicaciones.

Al observar el amplio espectro de investigaciones sobre XAI, es evidente que los académicos de diferentes disciplinas tienen diferentes objetivos en mente. Aunque diferentes aspectos de la investigación XAI siguen Para alcanzar los objetivos generales de la interpretabilidad de la IA, los investigadores de cada disciplina utilizan diferentes medidas y métricas para evaluar los objetivos de XAI. Por ejemplo, los métodos analíticos numéricos se emplean en los campos del aprendizaje automático para evaluar la interpretabilidad computacional, mientras que la interpretabilidad humana y Las evaluaciones de sujetos humanos son más comúnmente los objetivos principales en HCI y comunidades de visualización. En este sentido, aunque parece haber un desajuste en los objetivos específicos para diseñar y evaluar la explicabilidad y la interpretabilidad, una convergencia en los objetivos es beneficiosa para lograr todo el potencial de XAI. Con este fin, este artículo presenta una encuesta y un marco destinados a compartir

Conocimiento y experiencia de los métodos de diseño y evaluación de XAI en múltiples disciplinas. Para respaldar los diversos objetivos de diseño y métodos de evaluación en la investigación de XAI, después de una revisión exhaustiva de los artículos relacionados con XAI en los campos de aprendizaje automático, visualización y HCI, presentamos una categorización de objetivos de diseño y métodos de evaluación de aprendizaje automático interpretables y mostramos un mapeo. entre los objetivos de diseño para diferentes grupos de usuarios de XAI y sus métodos de evaluación. A partir de nuestros hallazgos, desarrollamos un marco con pautas de diseño paso a paso combinadas con métodos de evaluación para cerrar los ciclos iterativos de diseño y evaluación en equipos multidisciplinarios. Además, proporcionamos métodos de evaluación resumidos y listos para usar para diferentes objetivos en la investigación de XAI. Por último, revisamos recomendaciones para el diseño y la evaluación de XAI extraídas de nuestra revisión de la literatura.

2. FONDO

Hoy en día, los algoritmos analizan los datos de los usuarios y afectan los procesos de toma de decisiones de millones de personas en asuntos como el empleo, las tasas de seguros, las tasas de préstamos e incluso la justicia penal [35]. Sin embargo, estos algoritmos que cumplen funciones críticas en muchas industrias tienen sus propias desventajas que pueden resultar en discriminación [44, 196] y toma de decisiones injustas [163]. Por ejemplo, recientemente, los algoritmos de noticias y publicidad dirigida en las redes sociales han atraído mucha atención por agravar la falta de diversidad de información en las redes sociales [23]. Una parte importante del problema podría deberse a que los sistemas algorítmicos de toma de decisiones, a diferencia de los sistemas de recomendación, no permiten a sus usuarios elegir entre los elementos recomendados, sino que presentan ellos mismos el contenido u opción más relevante. Para abordar esto, Heer [75] sugiere el uso de representaciones compartidas de tareas que se aumentan tanto con modelos de aprendizaje automático como con conocimiento del usuario para reducir los efectos negativos de los sistemas autónomos de IA inmaduros. Presentan estudios de casos de sistemas interactivos que integran soporte computacional proactivo en sistemas interactivos.

Bellotti y Edwards [16] sostienen que los sistemas inteligentes conscientes del contexto no deberían actuar en nuestro nombre. Sugieren el control del usuario sobre el sistema como principio para respaldar la responsabilidad de un sistema y sus usuarios. La transparencia puede proporcionar información esencial para la toma de decisiones que queda oculta para los usuarios finales y evita la fe ciega [218]. Los beneficios clave de la transparencia y la interpretabilidad algorítmica incluyen la conciencia del usuario [9]; detección de prejuicios y discriminación [45, 196]; comportamiento interpretable de sistemas inteligentes [124]; y responsabilidad de los usuarios [46]. Además, considerando el creciente conjunto de ejemplos de discriminación y otros aspectos legales de la toma de decisiones algorítmica, los investigadores exigen e investigan la transparencia y la responsabilidad de la IA ante la ley para mitigar los efectos adversos de la toma de decisiones algorítmica [49, 145, 201] . En esta sección, revisamos los antecedentes de investigación relacionados con los sistemas XAI desde una perspectiva amplia y multidisciplinaria. Al final, relacionamos los resúmenes y posiciones derivadas de nuestra encuesta con otros trabajos en el campo.

2.1 Auditoría de la IA inexplicable Los

investigadores auditan algoritmos para estudiar el sesgo y la discriminación en la toma de decisiones algorítmicas [184] y estudian la conciencia de los usuarios sobre los efectos de estos algoritmos [58]. La auditoría de algoritmos es un mecanismo para investigar la funcionalidad de los algoritmos para detectar sesgos y otros comportamientos no deseados de los algoritmos sin la necesidad de conocer los detalles específicos de su diseño. Los métodos de auditoría se centran en los efectos problemáticos sobre los resultados de los sistemas algorítmicos de toma de decisiones. Para auditar un algoritmo, los investigadores introducen nuevas entradas en el algoritmo y revisan los resultados y el comportamiento del sistema. Los investigadores generan nuevos datos y cuentas de usuario con la ayuda de scripts, bots [44] y crowdsourcing [73] para emular datos reales y usuarios reales en el proceso de auditoría. Para la detección de sesgos entre múltiples algoritmos, la auditoría multiplataforma puede detectar si un algoritmo se comporta de manera diferente a otro algoritmo. Un ejemplo reciente de auditoría multiplataforma es un trabajo de Eslami et al. [59], en el que analizaron las opiniones de los usuarios en tres sitios web de reservas de hoteles para estudiar el conocimiento de los usuarios sobre

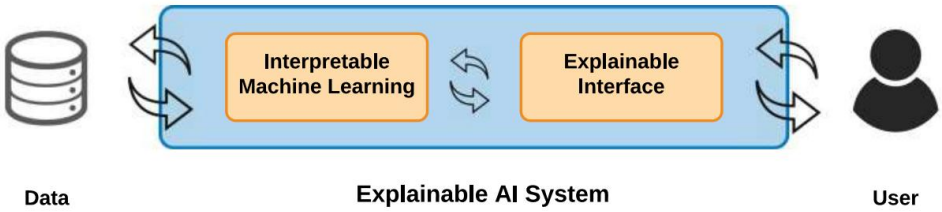


Fig. 1. El usuario interactúa con la interfaz explicable para enviar consultas al aprendizaje automático interpretable y recibir predicciones y explicaciones del modelo. El modelo interpretable interactúa con los datos para generar una explicación o una nueva predicción para la consulta del usuario.

Sesgo en los algoritmos de calificación en línea. Estos ejemplos demuestran que la auditoría es un proceso valioso pero que requiere mucho tiempo y que no se puede ampliar fácilmente a una gran cantidad de algoritmos. Esto requiere nuevas investigaciones para encontrar soluciones más efectivas hacia la transparencia algorítmica.

2.2 IA explicable Junto con

los métodos mencionados anteriormente para respaldar la transparencia, las explicaciones del aprendizaje automático también se han convertido en un enfoque común para lograr la transparencia en muchas aplicaciones, como las redes sociales, el comercio electrónico y la gestión de trabajadores humanos basada en datos [116 , 197, 199]. El sistema XAI, como se ilustra en la Figura 1, es capaz de generar explicaciones y describir el razonamiento detrás de las decisiones y predicciones del aprendizaje automático. Las explicaciones del aprendizaje automático permiten a los usuarios comprender cómo se procesan los datos. Su objetivo es concienciar sobre posibles sesgos y fallos del sistema. Por ejemplo, para medir la percepción de justicia de los usuarios en la toma de decisiones inteligente, Binns et al. [20] estudiaron explicaciones en sistemas para tareas cotidianas, como determinar las tarifas de seguros de automóviles y aprobaciones de solicitudes de préstamos. Sus resultados resaltan la importancia de las explicaciones del aprendizaje automático en la comprensión y la confianza de los usuarios en los sistemas algorítmicos de toma de decisiones.
En un trabajo similar que estudia el conocimiento de los algoritmos de las redes sociales, Radar et al. [170] realizaron un estudio colaborativo para ver cómo los diferentes tipos de explicaciones afectan las creencias de los usuarios sobre la transparencia algorítmica de las noticias en una plataforma de redes sociales. En su estudio, midieron la conciencia, la corrección y la responsabilidad de los usuarios para evaluar la transparencia algorítmica. Descubrieron que todas las explicaciones hacían que los usuarios se volvieran más conscientes del comportamiento del sistema. Stumpf y cols. [194] diseñaron experimentos para investigar explicaciones e interacciones significativas para responsabilizar a los usuarios mediante algoritmos de aprendizaje automático. Muestran explicaciones como un método potencial para respaldar una colaboración más rica entre humanos y computadoras para compartir inteligencia.

Los avances y tendencias recientes para la investigación de IA explicable exigen una amplia gama de objetivos para la transparencia algorítmica que exige investigación en diversas áreas de aplicación. Con este fin, nuestra revisión fomenta una perspectiva interdisciplinaria de los objetivos de inteligibilidad y transparencia.

2.3 Encuestas y directrices relacionadas En los

últimos años, ha habido encuestas y documentos de posición que sugieren direcciones de investigación y destacan los desafíos en la investigación de aprendizaje automático interpretable [48, 78, 127]. Aunque nuestra revisión se limita a la literatura sobre ciencias de la computación, aquí resumimos varias de las encuestas revisadas por pares más relevantes relacionadas con el tema de XAI en disciplinas activas, incluidas las ciencias sociales.
Si bien todas las encuestas, modelos y pautas de esta sección agregan valor a la investigación de XAI, hasta donde sabemos, no existe una encuesta ni un marco integral para los métodos de evaluación de sistemas de aprendizaje automático explicables.

2.3.1 Encuestas de Ciencias Sociales. La investigación en ciencias sociales es particularmente importante para que los sistemas XAI comprendan cómo las personas generan, comunican y comprenden explicaciones mediante

teniendo en cuenta el pensamiento de los demás, los sesgos cognitivos y las expectativas sociales en el proceso de explicación. Hoffman, Mueller y Klein revisaron los conceptos clave de explicaciones de sistemas inteligentes en una serie de ensayos para identificar cómo las personas formulan y aceptan explicaciones, formas de generar autoexplicaciones e identificaron propósitos y patrones para el razonamiento causal [83, 84, 102] .].
Por último, se centran en las redes neuronales profundas (DNN) para examinar sus hallazgos teóricos y empíricos sobre un algoritmo de aprendizaje automático [79]. En otro trabajo, presentaron un modelo conceptual del proceso de explicación en el contexto XAI [85]. Su marco incluye pasos y medidas específicos para la bondad de las explicaciones, la satisfacción del usuario y la comprensión de las explicaciones, la confianza de los usuarios en los sistemas XAI, los efectos de la curiosidad en la búsqueda de explicaciones y el rendimiento del sistema XAI humano.

Miller [142] sugiere que una estrecha colaboración entre los investigadores del aprendizaje automático en el espacio de XAI con las ciencias sociales perfeccionaría aún más la explicabilidad de la IA para las personas. Analiza cómo comprender y replicar cómo las personas generan, seleccionan y presentan explicaciones podría mejorar las interacciones entre humanos y XAI. Por ejemplo, Miller analiza cómo las personas generan y seleccionan explicaciones relacionadas con sesgos cognitivos y expectativas sociales. Otros artículos que revisan aspectos de ciencias sociales de los sistemas XAI incluyen estudios sobre el papel de la transparencia algorítmica y la explicación en la IA legal [49] y de los procesos algorítmicos de toma de decisiones justos y responsables [117].

2.3.2 Encuestas HCI. Muchas encuestas de HCI analizan las limitaciones y desafíos de la transparencia de la IA [208] y el aprendizaje automático interactivo [6]. Otros sugieren un conjunto de principios teóricos y de diseño para respaldar la inteligibilidad del sistema inteligente y la responsabilidad de los usuarios humanos (por ejemplo, [16, 90]). En una encuesta reciente, Abdul et al. [1] presentó un análisis exhaustivo de la literatura para encontrar temas relacionados con XAI y relaciones entre estos temas. Utilizaron visualización de palabras clave, modelos de temas y redes de citas para presentar una visión holística de los esfuerzos de investigación en una amplia gama de dominios relacionados con XAI; desde privacidad y equidad hasta agentes inteligentes y sistemas conscientes del contexto. En otro trabajo, Wang et al. [204] exploraron los fundamentos teóricos de la toma de decisiones humana y propusieron un marco conceptual para construir sistemas XAI basados en teorías de decisiones centradas en el ser humano. Su marco ayuda a elegir mejores explicaciones para presentar, respaldadas por teorías de razonamiento y sesgos cognitivos humanos. Centrados en el diseño de la interfaz XAI, Eiband et al. [56] presentan un proceso participativo basado en etapas para la integración de la transparencia en los sistemas inteligentes existentes utilizando explicaciones. Otro marco de diseño es XAID de Zhu et al. [225], que presenta un enfoque centrado en el ser humano para facilitar a los diseñadores de juegos la co-creación con técnicas de aprendizaje automático. Su estudio investiga la usabilidad de los algoritmos XAI en términos de qué tan bien apoyan a los diseñadores de juegos.

2.3.3 Encuestas de Análisis Visual. Las encuestas relacionadas con XAI en el dominio de la visualización siguen objetivos de análisis visual, como comprender e interactuar con sistemas de aprendizaje automático en diferentes aplicaciones de análisis visual [57, 180]. Choo y Liu [34] revisaron los desafíos y oportunidades de Visual Analytics para un diseño de aprendizaje profundo explicable. En un artículo reciente, Hohman et al. [88] proporcionan una excelente revisión y categorización de herramientas de análisis visual para aplicaciones de aprendizaje profundo. Cubren diversas técnicas de visualización y datos que se utilizan en aplicaciones de análisis visual profundo. Además, Spinner et al. [192] propusieron un proceso XAI que asigna el proceso XAI a un flujo de trabajo iterativo en tres etapas: comprensión del modelo, diagnóstico y refinamiento. Para poner en práctica su marco, diseñaron explAIner, un sistema de análisis visual para aprendizaje automático interactivo e interpretable que crea instancias de todos los pasos de su proceso.

2.3.4 Encuestas de Aprendizaje Automático. En el área del aprendizaje automático, Guidotti et al. [71] presentan una revisión y clasificación exhaustivas de las técnicas de interpretabilidad del aprendizaje automático. Además, Montavon et al. [152] se centran en técnicas de interpretabilidad para modelos DNN. Sobre convolucional

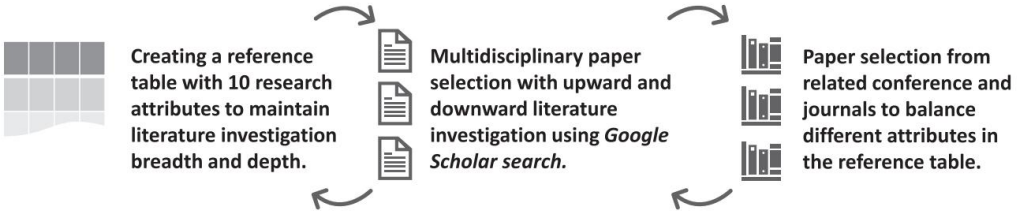


Fig. 2. Un diagrama que resume nuestro proceso de revisión y selección de literatura iterativo y de múltiples pasadas para lograr la amplitud y profundidad deseada de la investigación de la literatura. Comenzamos con 40 artículos para crear la tabla de referencia. Luego agregamos 80 artículos mediante investigación bibliográfica ascendente y descendente para mejorar la amplitud y profundidad de la revisión. Finalmente, agregamos otros 80 artículos de actas de congresos y revistas relacionadas para equilibrar la tabla de referencia.

Red neuronal (CNN), Zhang et al. [221] revisa la investigación sobre técnicas de interpretabilidad en seis direcciones, incluida la visualización de representaciones de CNN, técnicas de diagnóstico para CNN, enfoques para transformar representaciones de CNN en gráficos interpretables, construcción de modelos explicables y aprendizaje a nivel semántico basado en la interpretabilidad del modelo. En otro trabajo, Gilpin et al. [64] revisa las técnicas de interpretabilidad en los algoritmos de aprendizaje automático y categoriza los enfoques de evaluación para cerrar la brecha entre el aprendizaje automático y las comunidades HCI.

Para complementar el trabajo existente, nuestra encuesta proporciona una categorización multidisciplinaria de objetivos de diseño y métodos de evaluación para sistemas XAI. Como resultado de los artículos analizados, proponemos un marco que proporciona un plan de diseño y evaluación paso a paso para un equipo multidisciplinario de diseñadores para construir sistemas XAI del mundo real. A diferencia de Eiband et al. [56], no asumimos la necesidad de agregar transparencia a una interfaz inteligente existente y no limitamos la evaluación de los sistemas XAI al modelo mental de los usuarios. En cambio, caracterizamos tanto los objetivos de diseño como los métodos de evaluación y los compilamos todos en un marco unificado para el trabajo en equipo multidisciplinario. Nuestro marco de diseño tiene similitudes con el marco teórico de Wang et al. [204] que respalda nuestros objetivos de diseño (ver Sección 9.6). Nuestro trabajo multidisciplinario amplía su marco conceptual al (1) incluir el diseño de algoritmos de interpretabilidad como parte del marco y (2) combinar métodos de evaluación con cada paso de diseño para cerrar los ciclos iterativos de diseño y evaluación.

3 MÉTODO DE ENCUESTA

Realizamos una encuesta de la literatura de investigación existente para capturar y organizar la amplitud de diseños y objetivos para la evaluación de XAI. Utilizamos una metodología estructurada e iterativa para encontrar investigaciones relevantes para XAI y categorizar los métodos de evaluación presentados en los artículos de investigación (resumidos en la Figura 2). En nuestro proceso iterativo de selección de artículos, comenzamos seleccionando trabajos existentes de las principales conferencias y revistas de informática en los campos de HCI, visualización y aprendizaje automático. Sin embargo, dado que XAI es un tema de rápido crecimiento, también queríamos incluir preimpresiones de arXiv y discusiones útiles en los artículos del taller. Comenzamos con 40 artículos relacionados con temas de XAI en tres campos de investigación que incluyen, entre otros, investigación sobre técnicas de aprendizaje automático interpretables, visualización de aprendizaje profundo, visualización de modelos interactivos, explicaciones de máquinas en agentes inteligentes y sistemas conscientes del contexto, interfaces de usuario explicables, depuración explicativa y transparencia y equidad algorítmica.

Luego utilizamos codificación selectiva para identificar 10 atributos principales de investigación en esos artículos. Los principales atributos que identificamos incluyen disciplina de investigación (ciencias sociales, HCI, visualización o aprendizaje automático), tipo de artículo (diseño de interfaz, diseño de algoritmos o artículo de evaluación), dominio de aplicación (interpretabilidad del aprendizaje automático, equidad algorítmica, sistemas de recomendación, transparencia de sistemas inteligentes, sistemas y agentes interactivos inteligentes, sistemas inteligentes explicables y

agentes, explicaciones humanas o confianza humana), modelo de aprendizaje automático (p. ej., aprendizaje profundo, toma de decisiones árboles, SVM), modalidad de datos (imagen, texto, datos tabulares), tipo de explicación (p. ej., gráfica, textual, datos visualización), objetivo de diseño (p. ej., depuración del modelo, confianza del usuario, mitigación de sesgos), tipo de evaluación (por ejemplo, cualitativo, computacional, cuantitativo con sujetos humanos), usuario objetivo (novatos en IA, datos expertos, expertos en IA) y medida de evaluación (p. ej., confianza del usuario, desempeño de tareas, modelo mental del usuario).

En la segunda ronda de recopilación de literatura XAI, llevamos a cabo una investigación de literatura ascendente y descendente utilizando el motor de búsqueda Google Scholar para agregar 80 artículos más a nuestra referencia. mesa. Redujimos nuestra búsqueda por temas y palabras clave relacionados con XAI que incluyen, entre otros: interpretabilidad, explicabilidad, inteligibilidad, transparencia, toma de decisiones algorítmica, equidad, confianza, modelo mental y depuración en aprendizaje automático y sistemas inteligentes. Con Esta información, realizamos una codificación axial para organizar la literatura y comenzamos discusiones sobre nuestra propuesta de categorización de diseño y evaluación.

Finalmente, para mantener una cobertura bibliográfica razonable y equilibrar el número de artículos para cada uno de los nuestras categorías de objetivos de diseño y medidas de evaluación, agregamos otros 80 artículos a nuestra tabla de referencias. Las conferencias de las cuales seleccionamos artículos relacionados con XAI incluyen, entre otras, a los siguientes: CHI, IUI, HCOMP, SIGDIAL, UbiComp, AIES, VIS, ICWSM, IJCAI, KDD, AAAI, Conferencias CVPR y NeurIPS. Las revistas incluyeron Trends in Cognitive Science, Transactions sobre sistemas cognitivos y de desarrollo, Cognition Journal, Transactions on Interactive Intelligence Systems, Revista internacional de estudios humanos-computadores, Transactions on Visualization y gráficos por computadora y transacciones en redes neuronales y sistemas de aprendizaje.

Luego de una revisión de 226 artículos, nuestra categorización de los objetivos de diseño y métodos de evaluación de XAI está respaldada por referencias de artículos que realizan el diseño o la evaluación de sistemas XAI. Nuestro La tabla de referencia 1 está disponible en línea para la comunidad de investigación para proporcionar más información más allá nuestras discusiones en este documento. La Tabla 2 muestra un resumen de nuestros artículos encuestados que contiene 42 artículos con diseño y evaluación del sistema XAI. Más adelante en la Sección 7, proporcionamos una serie de tablas para organizar diferentes métodos de evaluación de artículos de investigación con referencias de ejemplo para cada uno, documentando nuestro análisis en profundidad de 69 artículos en total.

4 TERMINOLOGÍA XAI

Para familiarizar a los lectores con conceptos y terminologías comunes de XAI a los que se hace referencia repetidamente en esta revisión, las siguientes cuatro subsecciones resumen caracterizaciones de alto nivel de explicaciones del modelo. Muchas encuestas relacionadas (p. ej., [2, 207]) e informes (p. ej., [38, 200]) también proporcionan compilaciones útiles de terminología y conceptos en informes completos. Por ejemplo, Abdul et al. [1] presentan un gráfico de citas de diversos dominios relacionados con explicaciones, incluidos sistemas inteligentes inteligibles, sistemas conscientes del contexto y capacidad de aprendizaje de software. Posteriormente, Arrieta et al. [11] presentar una revisión exhaustiva de los conceptos y taxonomías de XAI y llegar al concepto de IA responsable como una variedad de múltiples principios de IA que incluyen la equidad del modelo, la explicabilidad y la privacidad. De manera similar, el concepto de IA segura ha sido revisado por Amodei et al. [8], que es un interés en aplicaciones inteligentes críticas para la seguridad, como vehículos autónomos [147]. La Tabla 1 presenta descripciones de 14 términos comunes relacionados con el tema de esta encuesta y organiza su relación con Inteligible. Temas de sistemas e IA transparente. Consideramos los sistemas de IA transparentes como la clase de inteligencia artificial basada en IA. Sistemas Inteligibles. Por tanto, propiedades y objetivos previamente establecidos para los Sistemas Inteligibles. Lo ideal es que sean transferibles a sistemas de IA transparentes. Sin embargo, los desafíos y limitaciones para lograr la transparencia en algoritmos complejos de aprendizaje automático plantean problemas (por ejemplo, garantizar la equidad de un algoritmo) que no eran necesariamente problemáticos en sistemas inteligibles basados en reglas, pero ahora requieren una mayor atención por parte de las comunidades de investigación.

1https://github.com/SinaMohseni/Awesome-XAI-Evaluación.

Tabla 1. Tabla de terminología común relacionada con sistemas inteligibles e IA transparente

Concepto	Categoría	Descripción Un
principal del sistema inteligible	a través de	sistema que es comprensible y predecible para sus usuarios . Concepto transparencia o explicaciones [1, 16, 207].
Comprensibilidad	Deseado	Los sistemas inteligibles apoyan la comprensión del usuario (inteligibilidad) de las funciones subyacentes del sistema [11, 123].
Previsibilidad	Properties	Intelligibility apoya la construcción de un modelo mental del sistema que permite al usuario predecir el comportamiento del sistema [207].
Integridad	Deseado Resultados	Permitir una actitud positiva del usuario hacia el sistema que surja del conocimiento, la experiencia y la emoción [82, 85].
Fiabilidad		Apoyar la confianza del usuario para confiar y seguir los consejos del sistema para un mayor rendimiento [82, 85].
Seguridad		Mejorar la seguridad al reducir el uso indebido involuntario del usuario debido a percepciones erróneas y desconocimiento [147].
Concepto principal de IA transparente		Un sistema basado en IA que proporciona información sobre sus procesos de toma de decisiones [38, 127].
IA interpretable	Práctico	Modelos inherentemente interpretables por humanos debido a la baja complejidad de los algoritmos de aprendizaje automático [151].
IA explicable	Enfoques	Apoyar la comprensión del usuario de modelos complejos proporcionando explicaciones para las predicciones [204].
Interpretabilidad	Deseado	La capacidad de respaldar la comprensión y la comprensión del usuario del proceso de toma de decisiones y las predicciones del modelo [11, 127].
Explicabilidad	Propiedades	La capacidad de explicar el modelo subyacente y su razonamiento. con explicaciones precisas y comprensibles para el usuario [11, 127].
IA responsable	Desired	Permitir auditorías y documentación para responsabilizar a las organizaciones de sus productos y servicios basados en IA [49, 117].
IA justa	Resultados	Permitir el análisis ético y justo de los modelos. y datos utilizados en los procesos de toma de decisiones [11, 117].

Los conceptos principales de nivel superior se muestran en gris, mientras que los términos relacionados con los conceptos principales se enumeran a continuación y se clasifican como resultado, propiedad o enfoque práctico deseado. La IA explicable es un enfoque práctico particular para que los sistemas inteligibles permitan una mayor transparencia. Tenga en cuenta que las definiciones e interpretaciones pueden variar según la literatura y esta tabla pretende servir como referencia rápida.

Las descripciones presentadas en la Tabla 1 pretenden ser una introducción a estos términos, aunque las definiciones e interpretaciones exactas pueden depender del contexto de uso y la disciplina de investigación. En consecuencia, investigadores de diferentes disciplinas suelen utilizar estos términos indistintamente, sin tener en cuenta las diferencias de significado [2]. Quizás los dos términos genéricos de modelo de caja negra y modelo transparente estén en el centro de la ambigüedad terminológica de XAI. El término caja negra se refiere a modelos complejos de aprendizaje automático que no son interpretables por humanos [127] , a diferencia de modelos transparentes que son lo suficientemente simples como para ser interpretables por humanos [11]. Consideramos que es más preciso y consistente separar la transparencia de un sistema XAI (como se describe en la Figura 1) de la interpretabilidad de sus modelos internos de aprendizaje automático. Específicamente, la Tabla 1 muestra que la IA transparente podría lograrse mediante enfoques de IA interpretable o IA explicable. Otros ejemplos de ambigüedad terminológica incluyen los términos Interpretabilidad y Explicabilidad que a menudo se utilizan como sinónimos en el campo del aprendizaje automático. Por ejemplo, la frase "técnica de aprendizaje automático interpretable" a menudo se refiere a técnicas ad-hoc para generar explicaciones para modelos no interpretables como las DNN [151]. Otro ejemplo es el caso ocasional de utilizar los términos Sistema Transparente y Sistema Explicable indistintamente en la investigación de HCI [56], mientras que otros aclaran que explicabilidad no equivale a transparencia porque no requiere conocer el flujo de los bits en la toma de decisiones de la IA. proceso [49].

4.1 Explicaciones globales y locales

Una forma de clasificar las explicaciones es según su escala de interpretación. Por ejemplo, una explicación podría ser tan completo como describir todo el modelo de aprendizaje automático. Alternativamente, podría explicar sólo parcialmente el modelo, o podría limitarse a explicar una instancia de entrada individual. Global La explicación (o explicación del modelo) es un tipo de explicación que describe cómo funciona la máquina en general. El modelo de aprendizaje funciona. La visualización de modelos [130, 131] y las reglas de decisión [113] son ejemplos de explicaciones que caen en esta categoría. En otros casos, aproximaciones interpretables de modelos complejos. sirva como modelo de explicación. La regularización de árboles [213] es un ejemplo reciente de un modelo complejo regularizado para aprender límites de decisión en forma de árbol. La complejidad del modelo y el diseño de explicación son los principales factores utilizados para elegir entre diferentes tipos de explicaciones globales.

Por el contrario, las explicaciones locales (o explicaciones de instancia) tienen como objetivo explicar la relación entre pares de entrada-salida específicos o el razonamiento detrás de los resultados de una consulta de usuario individual. Se cree que este tipo de explicación es menos abrumadora para los principiantes y puede ser adecuada para investigar casos extremos para el modelo o depurar datos. Las explicaciones locales a menudo hacen uso de métodos de prominencia [14, 219] o aproximación local del modelo principal [172, 173]. Métodos de prominencia (también conocidos como mapas de atribución o mapas de sensibilidad) utilizan diferentes enfoques (por ejemplo, métodos basados en perturbaciones, métodos basados en gradientes) para mostrar qué características de la entrada influyen fuertemente en la predicción del modelo. La aproximación local del modelo, por otro lado, entrena un interpretable modelo (aprendido del modelo principal) para representar localmente el comportamiento del modelo complejo.

4.2 Modelos interpretables frente a explicadores ad hoc

La interpretabilidad humana de un modelo de aprendizaje automático es inversamente proporcional a la capacidad del modelo. tamaño y complejidad. Los modelos complejos (por ejemplo, DNN) con alto rendimiento y robustez en aplicaciones del mundo real no son interpretables por usuarios humanos debido a su gran espacio variable. Lineal Los modelos de regresión o árboles de decisión ofrecen una mejor interpretabilidad pero tienen un rendimiento limitado en datos de alta dimensión, mientras que un modelo de bosque aleatorio (conjunto de cientos de árboles de decisión) Puede tener un rendimiento mucho mayor pero es menos comprensible. Esta compensación entre la interpretabilidad del modelo y el rendimiento llevó a los investigadores a diseñar métodos ad hoc para explicar cualquier tipo de caja negra. Algoritmo de aprendizaje automático como DNN. Los explicadores ad hoc (p. ej., [134, 172]) son independientes algoritmos que pueden describir las predicciones del modelo explicando "por qué" se ha tomado una determinada decisión. hecho en lugar de describir el modelo completo. Sin embargo, existen limitaciones a la hora de explicar modelos de caja negra con explicadores ad hoc, como la incertidumbre sobre la fidelidad del propio explicador. Discutiremos más sobre la fidelidad de las explicaciones en la Sección 7.5. Además, aunque los explicadores ad hoc generalmente describen "por qué" se hace una predicción, estos métodos no explican "cómo" se toma la decisión.

4.3 Qué explicar

Cuando los usuarios se enfrentan a un sistema inteligente complejo, pueden exigir diferentes tipos de explicaciones. La información y cada tipo de explicación pueden requerir su propio diseño. Aquí, revisamos seis comunes tipos de explicaciones utilizadas en los diseños de sistemas XAI.

Cómo las explicaciones demuestran una representación holística del algoritmo de aprendizaje automático para Explique cómo funciona el modelo. Para representaciones visuales, los gráficos modelo [113] y los límites de decisión [135] son ejemplos de diseño comunes para explicaciones de cómo. Sin embargo, la investigación muestra que los usuarios También puede ser capaz de desarrollar un modelo mental del algoritmo basado en una colección de explicaciones. de múltiples instancias individuales [133].

Las explicaciones de por qué describen por qué se realiza una predicción para una entrada en particular. Tales explicaciones objetivo comunicar qué características en los datos de entrada [172] o qué lógica en el modelo [113, 173] ha llevado a una predicción dada por el algoritmo. Este tipo de explicación puede tener cualquiera de los dos modelos. soluciones agnósticas [134, 172] o dependientes del modelo [188] .

Las explicaciones Why-nNot ayudan a los usuarios a comprender las razones por las que un resultado específico no estaba en la salida del sistema [202]. Las explicaciones de por qué no (también llamadas explicaciones contrastivas) caracterizan las razones de las diferencias entre la predicción de un modelo y el resultado esperado del usuario.

La importancia de la característica (o atribución de características) se usa comúnmente como una técnica de interpretabilidad para Explicaciones de por qué y por qué no.

Las explicaciones hipotéticas implican la demostración de cómo los diferentes cambios algorítmicos y de datos afectan la salida del modelo dadas nuevas entradas [29], la manipulación de entradas [125] o el cambio de parámetros del modelo [103]. El sistema puede recomendar automáticamente diferentes escenarios hipotéticos o puede ser elegido para la exploración a través del control interactivo del usuario. Dominios con datos de alta dimensión (p. ej., imagen y texto) y los modelos complejos de aprendizaje automático (p. ej., DNN) tienen menos parámetros para usuarios sintonizar y examinar directamente modelos entrenados en comparación con datos más simples (por ejemplo, de baja dimensión). datos tabulares) y modelos.

Las explicaciones prácticas detallan ajustes hipotéticos a la entrada o modelo que resultarían. en una salida diferente [125, 126], como una salida de interés especificada por el usuario. Las técnicas para generar explicaciones prácticas (o contrafactuales) son ad hoc e independientes del modelo, considerando que el modelo La estructura y los valores internos no son parte de la explicación [203]. Estos métodos pueden funcionar de forma interactiva con la curiosidad del usuario y la concepción parcial del sistema para permitir una evolución mental. modelo del sistema mediante pruebas iterativas.

Las explicaciones What-Else presentan a los usuarios instancias similares de entrada que generan el mismo o resultados similares del modelo. También llamado Explicación por ejemplo, Explicaciones de qué más seleccionar muestras del conjunto de datos de entrenamiento del modelo que sean similares a la entrada original en el modelo espacio de representación [30]. Aunque son muy populares y fáciles de lograr, las investigaciones muestran que las explicaciones basadas en ejemplos podrían ser engañosas cuando los conjuntos de datos de entrenamiento carecen de una distribución uniforme de los datos. datos [98].

4.4 Cómo explicar

En todo tipo de explicaciones de aprendizaje automático, el objetivo es revelar nueva información sobre el sistema subyacente. En esta encuesta, nos centramos principalmente en explicaciones comprensibles para los humanos, aunque Observamos que la investigación sobre el aprendizaje automático interpretable también ha estudiado otros propósitos como transferencia de conocimiento, localización de objetos y detección de errores [61, 162].

Las explicaciones se pueden diseñar utilizando una variedad de formatos para diferentes grupos de usuarios [216]. Visual Las explicaciones utilizan elementos visuales para describir el razonamiento detrás de los modelos de aprendizaje automático. Los mapas de atención y la prominencia visual en forma de mapas de calor de prominencia [190, 219] son ejemplos de explicaciones que se utilizan ampliamente en la literatura sobre aprendizaje automático. Las explicaciones verbales describen modelo de máquina o razonamiento con palabras, frases o lenguaje natural. Las explicaciones verbales son Popular en aplicaciones como explicaciones de preguntas y respuestas y listas de decisiones [113]. Esta forma de La explicación también se ha implementado en sistemas de recomendación [17, 77] y robótica [176]. Las interfaces explicables comúnmente utilizan múltiples modalidades (por ejemplo, visual, verbal y numérica). elementos) para obtener explicaciones que respalden la comprensión del usuario [156]. La explicación analítica es otra enfoque para ver y explorar los datos y las representaciones de los modelos de aprendizaje automático [88]. Las explicaciones analíticas suelen basarse en métricas numéricas y visualizaciones de datos. Analítica visual Las herramientas también permiten a los investigadores revisar las estructuras del modelo, las relaciones y sus parámetros en modelos profundos complejos. Las visualizaciones de mapas de calor [193], gráficos y redes [66] y las visualizaciones jerárquicas (árboles de decisión) se utilizan comúnmente para visualizar explicaciones analíticas de interpretables.

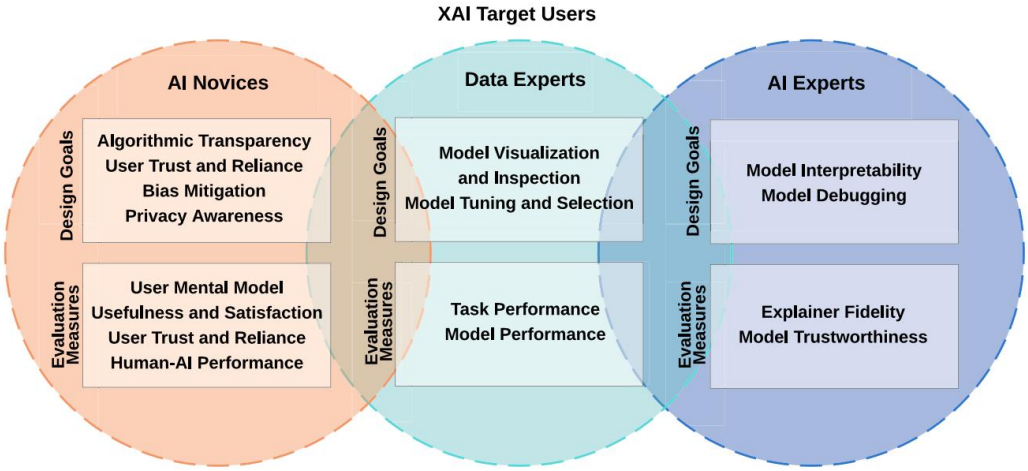


Fig. 3. Un resumen de nuestra categorización de los objetivos de diseño de XAI y las medidas de evaluación entre grupos de usuarios. Arriba: Diferentes objetivos de diseño del sistema para cada grupo de usuarios. Abajo: Medidas de evaluación comunes utilizadas en cada grupo de usuarios. Tenga en cuenta que objetivos XAI similares para diferentes grupos de usuarios requieren diferentes objetivos de investigación, métodos de diseño y rutas de implementación.

algoritmos. Recientemente, Hohman et al. [87] implementaron una combinación de visualización y verbalización para comunicar o resumir aspectos clave de un modelo.

Desde una perspectiva diferente, Chromik et al. [36] extiende la idea de "patrones oscuros" del diseño de interfaz de usuario interactiva [68] a explicaciones de aprendizaje automático. Revisan posibles formas en que la redacción de las explicaciones y su implementación en la interfaz podrían engañar a los usuarios en beneficio de otras partes. Revisan efectos negativos como la falta de atención del usuario a las explicaciones y la formación de un modelo mental incorrecto, e incluso la ansiedad algorítmica [93] podría estar entre las consecuencias de tales presentaciones e interacciones engañosas de explicaciones de aprendizaje automático.

5 CATEGORIZACIÓN DE METAS DE DISEÑO XAI Y MÉTODOS DE EVALUACIÓN

Si bien un sistema XAI ideal debería poder responder todas las consultas de los usuarios y cumplir con todos los objetivos del concepto XAI [72], los esfuerzos de investigación individuales se centran en diseñar y estudiar sistemas XAI con respecto a objetivos de interpretabilidad específicos y usuarios específicos. De manera similar, evaluar las explicaciones puede demostrar y verificar la efectividad de los sistemas explicables para los objetivos previstos.

Después de una cuidadosa revisión y análisis de los objetivos de XAI y sus métodos de evaluación en la literatura, reconocimos los dos atributos siguientes como los más importantes para nuestros propósitos de organización interdisciplinaria de los métodos de diseño y evaluación de XAI:

- Objetivos de diseño. El primer atributo de nuestra categorización es el objetivo de diseño de algoritmos interpretables e interfaces explicables en la investigación XAI. Obtenemos objetivos de diseño XAI de múltiples disciplinas de investigación: aprendizaje automático, visualización de datos y HCI. Para comprender mejor las diferencias entre los distintos objetivos de XAI, organizamos los objetivos de diseño de XAI con sus tres grupos de usuarios: principiantes en IA (es decir, usuarios finales de productos de IA en general), expertos en datos (expertos en análisis de datos y expertos en dominios) y Expertos en IA (diseñadores de modelos de aprendizaje automático). • Medidas de Evaluación. Revisamos los métodos de evaluación y discutimos las medidas utilizadas para evaluar las explicaciones del aprendizaje automático. Las medidas incluyen el modelo mental del usuario, la confianza del usuario, la utilidad y satisfacción de la explicación, el desempeño de la tarea humano-má

medidas computacionales. En nuestra revisión, prestaremos más atención a las medidas de evaluación de XAI ya que los autores creen que esta categoría está relativamente menos explorada.

La Figura 3 presenta el emparejamiento entre los objetivos de diseño de XAI y sus medidas de evaluación. Tenga en cuenta que los grupos de usuarios se utilizan como dimensión auxiliar para enfatizar la importancia de los usuarios finales para los objetivos del sistema. La superposición entre los grupos de usuarios de XAI muestra similitudes en los métodos de diseño y evaluación entre diferentes grupos de usuarios objetivo. Sin embargo, los objetivos XAI similares en diferentes grupos de usuarios requieren diferentes objetivos de investigación, métodos de diseño y rutas de implementación. Para ayudar a resumir nuestra caracterización junto con literatura de ejemplo, la Tabla 2 presenta una tabla de referencias cruzadas de la literatura de evaluación XAI para enfatizar la importancia de los objetivos de diseño, las medidas de evaluación y los tipos de usuarios. Primero revisamos los detalles de la investigación centrada en los objetivos de diseño de XAI en la Sección 6 , incluidos ocho objetivos organizados por sus grupos de usuarios. Luego revisamos las medidas y métodos de evaluación en la Sección 7 , incluidas seis medidas principales y sus métodos recopilados de la literatura analizada.

6 OBJETIVOS DE DISEÑO XAI

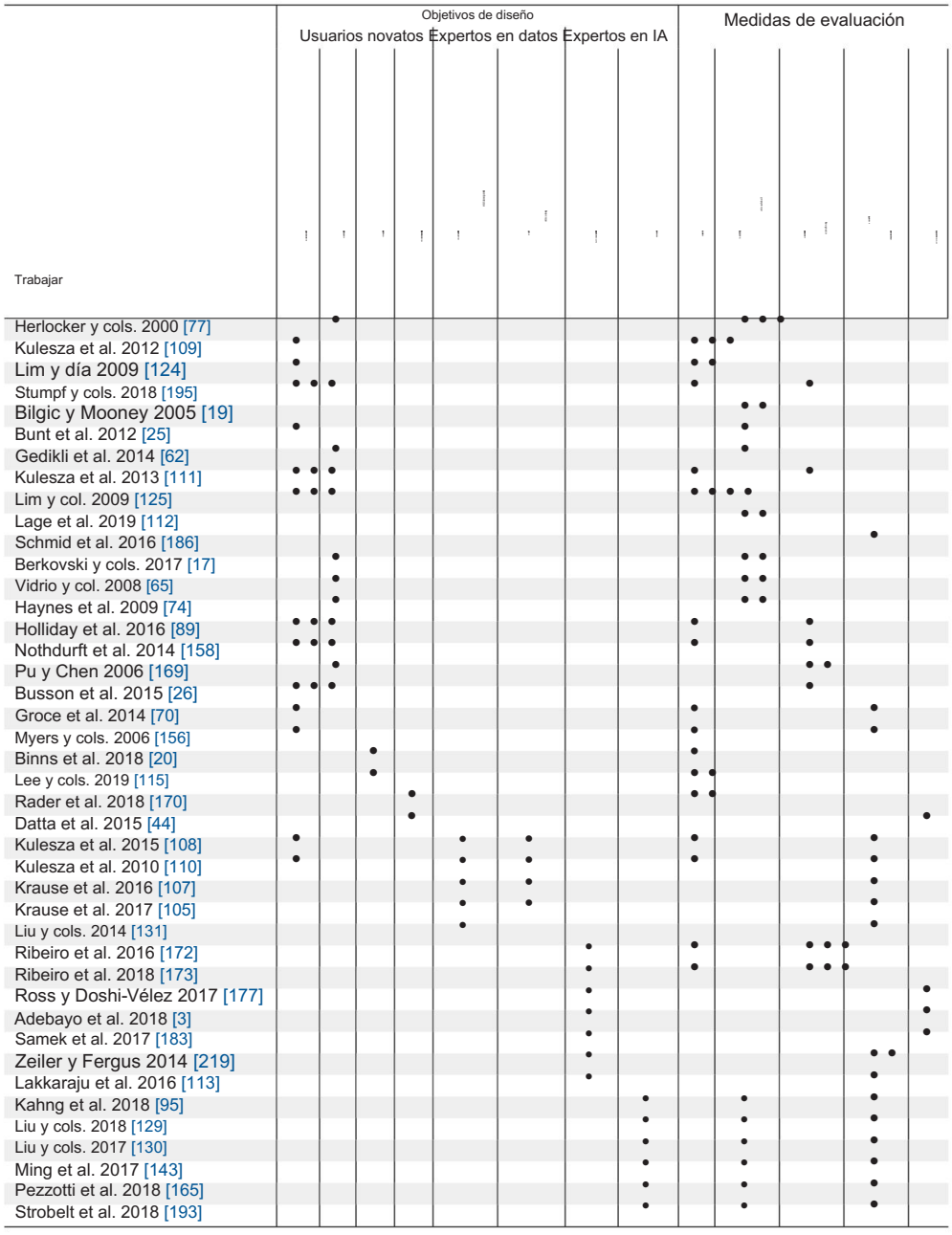
Los esfuerzos de investigación han explorado muchos objetivos para los sistemas XAI. Doshi-Velez y Kim [48] revisaron múltiples prioridades de alto nivel para los sistemas XAI con ejemplos que incluyen seguridad, ética, confianza del usuario y comprensión científica. Posteriormente, Arrieta et al. [11] presentó una revisión exhaustiva de las oportunidades de XAI en diferentes dominios de aplicación. En consecuencia, las diferentes opciones de diseño, como el tipo de explicación, el alcance y el nivel de detalle, se verán afectadas por el dominio de la aplicación, el objetivo del diseño y el tipo de usuario. Por ejemplo, si bien los expertos en aprendizaje automático pueden preferir visualizaciones muy detalladas de modelos profundos para ayudarlos a optimizar y diagnosticar algoritmos, los usuarios finales de productos de inteligencia artificial de uso diario no esperan explicaciones completamente detalladas para cada consulta por parte de un agente personalizado. Por lo tanto, se espera que los sistemas XAI proporcionen el tipo correcto de explicaciones para el grupo correcto de usuarios, lo que significa que será más eficiente diseñar un sistema XAI de acuerdo con las necesidades y niveles de experiencia del usuario.

Con este fin, distinguimos los objetivos de diseño de XAI en función del usuario final designado y los sujetos de evaluación, que clasificamos en tres grupos generales de expertos en IA, expertos en datos y novatos en IA.
Hacemos hincapié en que esta separación de grupos se presenta principalmente por conveniencia organizacional, ya que los objetivos no son mutuamente excluyentes entre los grupos y las prioridades específicas dependen de cada caso para cualquier proyecto en particular. Los objetivos de diseño de XAI también se extienden al objetivo más amplio de una IA responsable al mejorar la transparencia y la explicabilidad de los sistemas inteligentes. Tenga en cuenta que, aunque existen superposiciones en los métodos utilizados para lograr estos objetivos, los objetivos de la investigación y los enfoques de diseño son sustancialmente diferentes entre los distintos campos de investigación y sus grupos de usuarios. Por ejemplo, aunque aprovechar modelos interpretables para reducir el sesgo del modelo de aprendizaje automático es un objetivo de investigación para los expertos en IA, la mitigación del sesgo también es un objetivo de diseño para que los principiantes en IA eviten los efectos adversos de la toma de decisiones algorítmica en sus respectivos entornos de dominio. Sin embargo, las técnicas de interpretabilidad para los expertos en IA y las herramientas de mitigación de sesgos para los principiantes en IA requieren diferentes métodos y elementos de diseño. En las siguientes subsecciones, revisamos ocho objetivos de diseño para sistemas XAI organizados por sus grupos de usuarios.

6.1 Novatos en IA

Los novatos en IA se refieren a usuarios finales que utilizan productos de IA en la vida diaria pero que no tienen (o tienen muy poca) experiencia en sistemas de aprendizaje automático. Estos incluyen usuarios finales de aplicaciones inteligentes como agentes personalizados (por ejemplo, dispositivos de asistencia doméstica), redes sociales y sitios web de comercio electrónico. En la mayoría de los sistemas inteligentes, los algoritmos de aprendizaje automático sirven como funciones internas y API para habilitar funciones específicas integradas en interfaces inteligentes y sensibles al contexto. Investigaciones anteriores muestran que el diseño intuitivo de interfaz e interacción puede mejorar la experiencia de los usuarios con el sistema al mejorar la comprensión de los usuarios finales y su confianza en los sistemas inteligentes [154]. A este respecto,

Tabla 2. Resumen tabular de nuestras dimensiones de medidas de evaluación XAI y tipos de usuarios objetivo



La tabla incluye 42 artículos que representan un subconjunto de la literatura encuestada organizada por las dos dimensiones.

crear representaciones comprensibles para los humanos pero precisas del complicado aprendizaje automático explicaciones para usuarios finales novatos es una compensación de diseño desafiante en los sistemas XAI. Tenga en cuenta que aunque existen superposiciones entre los objetivos de los principiantes en IA y los expertos en IA que construyen interpretables algoritmos, cada grupo de usuarios requiere un conjunto diferente de métodos y objetivos de diseño que sean siendo estudiado en diferentes comunidades de investigación.

Los principales objetivos de diseño para los usuarios finales novatos en IA de un sistema XAI se pueden detallar de la siguiente manera.

G1: Transparencia algorítmica: un objetivo inmediato para un sistema XAI, en comparación con un sistema inteligente inexplicable, es ayudar a los usuarios finales a comprender cómo funciona el sistema inteligente. Las explicaciones del aprendizaje automático mejoran el modelo mental de los usuarios de los algoritmos inteligentes subyacentes al proporcionar una transparencia comprensible para los complejos algoritmos inteligentes [208]. Además, la transparencia de un sistema XAI puede mejorar la experiencia del usuario a través de una mejor comprensión del resultado del modelo [123], mejorando así las interacciones del usuario con el sistema [108].

G2: Confianza del usuario: un sistema XAI puede mejorar la confianza de los usuarios finales en el algoritmo inteligente al proporcionar explicaciones. Un sistema XAI permite a los usuarios evaluar la confiabilidad del sistema y calibrar su percepción de la precisión del sistema. Como resultado, la confianza de los usuarios en el algoritmo conduce a su dependencia del sistema. Ejemplos de aplicaciones en las que XAI pretende mejorar la confianza del usuario a través de su diseño transparente incluyen sistemas de recomendación [17], sistemas autónomos [209] y sistemas críticos de toma de decisiones [26].

G3: Mitigación de sesgos: la toma de decisiones algorítmicas injustas y sesgadas es un efecto secundario crítico de los sistemas inteligentes. El sesgo en el aprendizaje automático tiene muchas fuentes, incluidos datos de entrenamiento sesgados y aprendizaje de características que podrían resultar en discriminación en la toma de decisiones algorítmicas [137].
Las explicaciones del aprendizaje automático pueden ayudar a los usuarios finales a inspeccionar si los sistemas inteligentes están sesgados en su toma de decisiones. Ejemplos de casos en los que se utiliza XAI para mitigar el sesgo y evaluar la equidad son la evaluación del riesgo penal [20, 115] y la predicción de tasas de préstamos y seguros [32]. Vale la pena mencionar que existe una superposición entre el objetivo de mitigación de la toma de decisiones sesgada para los principiantes en IA y el objetivo del sesgo del conjunto de datos para los expertos en IA (Sección 6.2), lo que resulta en técnicas de implementación compartidas. Sin embargo, los dos grupos de usuarios distintos requieren sus propios conjuntos de objetivos y procesos de diseño de XAI.

G4: Concientización sobre la privacidad: otro objetivo al diseñar sistemas XAI es proporcionar un medio para que los usuarios finales evalúen la privacidad de sus datos. Las explicaciones del aprendizaje automático pueden revelar a los usuarios finales qué datos de usuario se utilizan en la toma de decisiones algorítmicas. Ejemplos de aplicaciones de IA en las que la conciencia de la privacidad es principalmente importante incluyen anuncios personalizados que utilizan publicidad en línea de los usuarios [44] y noticias personalizadas en las redes sociales [58, 170].

Además de los principales objetivos de XAI, también se han desarrollado herramientas de visualización interactiva para ayudar a los principiantes en IA a aprender conceptos y modelos de aprendizaje automático interactuando con datos simplificados y representaciones de modelos. Ejemplos de estas herramientas educativas incluyen TensorFlow Play-Ground [191] para enseñar conceptos elementales de redes neuronales y Adversarial Playground [157] para aprender el concepto de ejemplos contradictorios en DNN. Estas metas menores cubren objetivos del sistema XAI que tienen un alcance limitado en comparación con las metas principales.

6.2 Expertos en datos Los

expertos en datos incluyen científicos de datos y expertos en dominios que utilizan el aprendizaje automático para análisis, toma de decisiones o investigación. Las herramientas de análisis visual pueden respaldar el aprendizaje automático interpretable de muchas maneras, como visualizar la arquitectura de red de un modelo entrenado y el proceso de capacitación de modelos de aprendizaje automático. Los investigadores han implementado varios diseños de visualización y técnicas de interacción para comprender mejor y mejorar los modelos de aprendizaje automático.

Los expertos en datos analizan datos en formas y dominios especializados, como ciberseguridad [18, 66], medicina [31, 107], texto [128, 131] y análisis de imágenes satelitales [174]. Estos usuarios pueden ser expertos.

de ciertas áreas de dominio o expertos en áreas generales de la ciencia de datos, pero en nuestra categorización, consideramos que los usuarios en la categoría de expertos en datos generalmente carecen de experiencia en los detalles técnicos de los algoritmos de aprendizaje automático. En cambio, este grupo de usuarios suele utilizar herramientas inteligentes de análisis de datos o sistemas de análisis visual para obtener información a partir de los datos. Tenga en cuenta que existen superposiciones entre los objetivos de XAI en diferentes disciplinas y que tanto los diseñadores de modelos como los analistas de datos podrían utilizar herramientas de análisis visual diseñadas para expertos en datos. Sin embargo, las necesidades de diseño y los enfoques para estos sistemas XAI pueden ser diferentes entre las comunidades de investigación. Los principales objetivos de diseño para los usuarios expertos en datos de un sistema XAI son los siguientes.

G5: Visualización e inspección de modelos: al igual que los principiantes en IA, los expertos en datos también se benefician de la interpretabilidad del aprendizaje automático para inspeccionar la incertidumbre y la confiabilidad del modelo [181]. Por ejemplo, las explicaciones del aprendizaje automático ayudan a los expertos en datos a visualizar modelos [86] e inspeccionar problemas como el sesgo [4]. Otro aspecto importante de la visualización e inspección de modelos para expertos en el dominio es identificar y analizar casos de falla de modelos y sistemas de aprendizaje automático [144]. Por lo tanto, el principal desafío para los sistemas de análisis de datos y soporte de decisiones es mejorar la transparencia del modelo mediante técnicas de visualización e interacción para expertos en el dominio [216].

G6: Ajuste y selección de modelos: los enfoques de análisis visual pueden ayudar a los expertos en datos a ajustar los parámetros de aprendizaje automático para sus datos específicos de una manera visual interactiva [131]. El elemento de interpretabilidad en los sistemas de análisis visual XAI aumenta la capacidad de los expertos en datos para comparar múltiples modelos [5] y seleccionar el modelo correcto para los datos específicos. Como ejemplo, Du et al. [51] presentan EventAction, un enfoque de recomendación de secuencia de eventos que permite a los usuarios seleccionar de forma interactiva registros que comparten los valores de atributos deseados. En el caso de ajustar redes DNN, las herramientas de análisis visual mejoran la capacidad de los diseñadores para modificar redes [165], mejorar la capacitación [129] y comparar diferentes redes [211].

6.3 Expertos en IA En

nuestra categorización, los expertos en IA son científicos e ingenieros de aprendizaje automático que diseñan algoritmos de aprendizaje automático y técnicas de interpretabilidad para sistemas XAI. Las técnicas de interpretabilidad del aprendizaje automático proporcionan interpretación de modelos o explicaciones de instancias. Ejemplos de técnicas de interpretación de modelos incluyen modelos inherentemente interpretables [205], simplificación profunda del modelo [213] y visualización de las partes internas del modelo [215]. Sin embargo, las técnicas de explicación de instancias presentan características importantes para instancias individuales, como el mapa de prominencia en datos de imágenes y la atención en datos textuales [43]. Los ingenieros de IA también se benefician de las herramientas de visualización y análisis visual para inspeccionar interactivamente las variables internas del modelo [129] para detectar fallas de arquitectura y entrenamiento o monitorear y controlar el proceso de entrenamiento [95], lo que indica posibles superposiciones entre los objetivos de diseño. Enumeramos los principales objetivos de diseño para los expertos en IA en los dos elementos siguientes.

G7: Interpretabilidad del modelo: la interpretabilidad del modelo es a menudo un objetivo principal de XAI para los expertos en IA. La interpretabilidad del modelo permite obtener nuevos conocimientos sobre cómo los modelos profundos aprenden patrones a partir de los datos [162]. En este sentido, se han propuesto varias técnicas de interpretabilidad para diferentes dominios para satisfacer la necesidad de explicación [99, 188]. Por ejemplo, Yosinski et al. [215] crearon una caja de herramientas interactiva para explorar las capas de activación de la CNN en tiempo real que le da al usuario una intuición sobre "cómo funciona la CNN".

G8: Depuración de modelos: los investigadores de IA utilizan técnicas de interpretabilidad de diferentes maneras para mejorar la arquitectura del modelo y el proceso de capacitación. Por ejemplo, Zeiler y Fergus [219] presentan un caso de uso en el que la visualización de filtros y mapas de características en CNN conduce a la revisión del entrenamiento.

Tabla 3. Medidas y métodos de evaluación utilizados en el estudio de modelos mentales de usuario en sistemas XAI

Medidas del modelo mental	Métodos de evaluación
Comprensión del usuario del modelo	Entrevista ([40]) y Autoexplicación ([20, 47, 164])
	Cuestionario escala Likert ([99, 111, 113, 124, 133, 171])
Predicción de salida del modelo	Predicción del usuario de la salida del modelo ([96, 172, 173])
Predicción de fallas del modelo	Predicción del usuario del fallo del modelo ([15, 161])

hiperparámetros y, por tanto, mejora del rendimiento del modelo. En otro trabajo, Ribeiro et al. [172] utilizaron explicaciones de instancias de modelos y revisión humana de explicaciones para mejorar el rendimiento del modelo a través de la ingeniería de características.

Además de los objetivos principales de XAI para los expertos en IA, las explicaciones del aprendizaje automático se utilizan para otros fines, incluida la detección de sesgos en conjuntos de datos [220], la detección de ataques adversarios [61] y la predicción de fallas del modelo [146]. Además, los mapas de prominencia visual y los mecanismos de atención se han utilizado como técnicas de localización de objetos débilmente supervisadas [190], reconocimiento de múltiples objetos [12] y transferencia de conocimiento [122] .

7 MEDIDAS DE EVALUACIÓN DEL XAI

Las medidas de evaluación para sistemas XAI son otro factor importante en el proceso de diseño de sistemas XAI. Las explicaciones están diseñadas para responder a diferentes objetivos de interpretabilidad y, por lo tanto, se necesitan diferentes medidas para verificar la validez de la explicación para el propósito previsto. Por ejemplo, el diseño experimental con estudios de sujetos humanos es un enfoque común para realizar evaluaciones con usuarios finales novatos en IA. Se han utilizado varios estudios controlados en laboratorio y en línea para la evaluación de XAI. Además, los estudios de caso tienen como objetivo recopilar comentarios de los usuarios expertos en el dominio mientras realizan tareas cognitivas de alto nivel con herramientas de análisis. Por el contrario, las medidas computacionales están diseñadas para evaluar la precisión e integridad de las explicaciones de algoritmos interpretables.

En esta sección, revisamos y categorizamos las principales medidas de evaluación para sistemas y algoritmos XAI. La Tabla 2 muestra una lista de cinco medidas de evaluación asociadas con sus objetivos de diseño.
Además, proporcionamos medidas y métodos de evaluación XAI resumidos y listos para usar extraídos de la literatura en las Tablas 3 a 7.

7.1 M1: Modelo Mental

Siguiendo las teorías de la psicología cognitiva, un modelo mental es una representación de cómo los usuarios entienden un sistema. Los investigadores de HCI estudian los modelos mentales de los usuarios para determinar su comprensión de los sistemas inteligentes en diversas aplicaciones. Por ejemplo, Costanza et al. [40] estudiaron cómo los usuarios entienden un sistema de red inteligente, y Kay et al. [96] estudiaron cómo los usuarios entienden y se adaptan a la incertidumbre en la predicción del aprendizaje automático de los tiempos de llegada de los autobuses.

En el contexto de XAI, las explicaciones ayudan a los usuarios a crear un modelo mental de cómo funciona la IA.
La explicación del aprendizaje automático es una forma de ayudar a los usuarios a construir un modelo mental más preciso. Estudiar los modelos mentales de los usuarios de los sistemas XAI puede ayudar a verificar la efectividad de la explicación al describir el proceso de toma de decisiones de un algoritmo. La Tabla 3 resume los diferentes métodos de evaluación utilizados para medir el modelo mental de los usuarios de los modelos de aprendizaje automático.

La investigación en psicología sobre las interacciones entre humanos y IA también ha explorado la estructura, los tipos y las funciones de las explicaciones para encontrar ingredientes esenciales de una explicación ideal para una mejor comprensión del usuario y modelos mentales más precisos [97, 132]. Por ejemplo, Lombrozo [133] estudió cómo diferentes tipos de explicaciones pueden ayudar a estructurar la representación conceptual. Para descubrir cómo un sistema inteligente debería explicar su comportamiento a los no expertos, se investigan explicaciones del aprendizaje automático

ha estudiado cómo los usuarios interpretan los agentes inteligentes [47, 164] y los algoritmos [171] para descubrir qué los usuarios esperan de las explicaciones de la máquina. En relación con esto, Lim y Dey [124] obtienen tipos de explicaciones que los usuarios podrían esperar en cuatro aplicaciones del mundo real. Estudian específicamente qué tipos de explicaciones que los usuarios exigen en diferentes escenarios, como recomendaciones del sistema, eventos críticos, y comportamiento inesperado del sistema. Al medir el modelo mental del usuario a través de la predicción de fallos del modelo, Bansal et al. [15] diseñó un juego en el que los participantes reciben incentivos monetarios basados en su puntuación final de rendimiento. Aunque los experimentos se realizaron en un simple tridimensional tarea, sus resultados indican una disminución en la capacidad de los usuarios para predecir fallas del modelo a medida que los datos y el modelo volverse más complicado.

Una forma útil de estudiar la comprensión de los usuarios sobre los sistemas inteligentes es preguntarles directamente sobre el proceso de toma de decisiones del sistema inteligente. Analizar las entrevistas de los usuarios, pensar en voz alta, y las autoexplicaciones proporcionan información valiosa sobre los procesos de pensamiento y los modelos mentales de los usuarios [110]. Al estudiar la comprensión del usuario, Kulesza et al. [111] estudiaron el impacto de la solidez y la integridad de la explicación en la fidelidad del modelo mental de los usuarios finales en una interfaz de recomendación musical. Sus resultados encontraron que la exhaustividad (amplitud) de la explicación tenía un efecto más efecto significativo en la comprensión del usuario sobre el agente en comparación con la solidez de la explicación. En otro ejemplo, Binns et al. [20] estudió la relación entre las explicaciones de las máquinas y las de los usuarios. percepción de justicia en la toma de decisiones algorítmicas con diferentes conjuntos de estilos de explicación. Usuario La atención y las expectativas también pueden considerarse durante los ciclos de diseño de la interfaz interpretable. para sistemas inteligentes [195].

El interés en desarrollar y evaluar explicaciones comprensibles para los humanos también ha llevado a modelos interpretables y explicadores ad hoc para medir modelos mentales. Por ejemplo, Ribeiro et al. [172] evaluó la comprensión de los usuarios del algoritmo de aprendizaje automático con explicaciones visuales. Ellos mostró cómo las explicaciones mitigan la sobreestimación humana de la precisión de un clasificador de imágenes y ayudar a los usuarios a elegir un mejor clasificador según las explicaciones. En un trabajo de seguimiento, comparó las explicaciones globales de un modelo clasificador con las explicaciones de instancia del mismo El modelo y las explicaciones globales encontradas fueron soluciones más efectivas para encontrar las debilidades del modelo [173]. En otro artículo, Kim et al. [99] realizaron un estudio colaborativo para evaluar la comprensibilidad de las explicaciones basadas en funciones para los usuarios finales. Al abordar la comprensión de las representaciones de modelos, Lakkaraju et al. [113] presentó conjuntos de decisiones interpretables, una clasificación interpretable modelo y midió los modelos mentales de los usuarios con diferentes métricas, como la precisión del usuario al predecir la salida de la máquina y la duración de las autoexplicaciones de los usuarios.

7.2 M2: Explicación Utilidad y Satisfacción

La satisfacción del usuario final y la utilidad de la explicación de la máquina también son importantes al evaluar explicaciones en sistemas inteligentes [19]. Los investigadores utilizan diferentes subjetivos y objetivos. Medidas de comprensibilidad, utilidad y suficiencia de detalles para evaluar el valor explicativo. para los usuarios [142]. Aunque existen métodos implícitos para medir la satisfacción del usuario [80], una parte considerable de la literatura sigue la evaluación cualitativa de la satisfacción en las explicaciones, tales como como cuestionarios y entrevistas. Por ejemplo, Gedikli et al. [62] evaluaron 10 tipos diferentes de explicaciones con calificaciones de los usuarios en cuanto a satisfacción y transparencia de la explicación. Sus resultados mostraron una Fuerte relación entre la satisfacción del usuario y la transparencia percibida. De manera similar, Lim et al. [125] explorar la utilidad y eficiencia de la explicación en su sistema interpretable y consciente del contexto mediante Presentar diferentes tipos de explicaciones, como explicaciones de “por qué”, “por qué no” y “y si”. tipos y medir el tiempo de respuesta de los usuarios.

Otra línea de investigación estudia si los sistemas inteligibles siempre son apreciados por los usuarios. o tiene un valor condicional. Uno de los primeros trabajos de Lim y Dey [124] estudió la comprensión del usuario. y satisfacción de diferentes tipos de explicaciones en cuatro aplicaciones conscientes del contexto del mundo real. Su

Tabla 4. Medidas de satisfacción del usuario y métodos de estudio utilizados para medir al usuario Satisfacción y utilidad de las explicaciones en los estudios XAI

Medidas de satisfacción	Métodos de evaluación
Satisfacción del usuario	Entrevista y Autoinforme ([25, 62, 124, 125])
	Cuestionario escala Likert ([39, 62, 112, 124, 125])
	Estudio de caso de expertos ([95, 106, 128, 130, 193])
Explicación Utilidad	Compromiso con explicaciones ([39])
	Duración de la tarea y carga cognitiva ([62, 112, 125])

Los hallazgos muestran que, al considerar escenarios relacionados con la criticidad, los usuarios quieren más información que explique el proceso de toma de decisiones y experimentan mayores niveles de satisfacción después de recibir estas explicaciones. De manera similar, Bunt et al. [25] consideró si las explicaciones son siempre necesario para los usuarios en todo sistema inteligente. Sus resultados muestran que, en algunos casos, el costo de ver explicaciones en entradas del diario como recomendaciones de Amazon y YouTube podría superar sus beneficios. Para estudiar el impacto de la complejidad de las explicaciones en la comprensión de los usuarios, Lage et al. [112] estudiaron cómo la longitud y la complejidad de la explicación afectan el tiempo de respuesta, la precisión y la precisión de los usuarios. y satisfacción subjetiva. También observaron que la creciente complejidad de las explicaciones daba como resultado Disminución de la satisfacción subjetiva del usuario. En un estudio reciente, Coppers et al. [39] también muestran que agregar inteligibilidad no necesariamente mejora la experiencia del usuario en un estudio con traductores expertos. Su El experimento sugiere que los expertos prefieren un sistema inteligible cuando las explicaciones adicionales no forman parte del conocimiento fácilmente disponible del traductor. En otra obra, Curran et al. [41] midieron la comprensión y preferencia de las explicaciones de los usuarios en un reconocimiento de imágenes. tarea clasificando y codificando las transcripciones de los usuarios. Proporcionan tres tipos de explicaciones de instancia para participantes y muestran que, aunque todas las explicaciones provenían del mismo modelo, los participantes tenían diferentes niveles de confianza en la exactitud de las explicaciones, según la claridad de las explicaciones. y comprensibilidad.

La Tabla 4 resume los métodos de estudio utilizados para medir la satisfacción del usuario y la utilidad de las explicaciones del aprendizaje automático. Tenga en cuenta que el objetivo principal de las evaluaciones del sistema XAI para dominio y Los expertos en IA se realizan mediante la evaluación directa de la satisfacción del usuario del diseño explicativo durante el diseño. ciclo. Por ejemplo, los estudios de casos y el diseño participativo son enfoques comunes para incluir directamente a usuarios expertos como parte de los procesos de diseño y evaluación del sistema.

7.3 M3: Confianza del usuario

La confianza del usuario en un sistema inteligente es un factor afectivo y cognitivo que influye positiva o percepciones negativas de un sistema [82, 136]. Confianza inicial del usuario y desarrollo de la confianza con el tiempo Se han estudiado y presentado con diferentes términos, como confianza rápida [141], confianza predeterminada [139], y confianza sospechosa [21]. Los conocimientos y creencias previos son importantes para dar forma al estado inicial. de confianza; Sin embargo, la confianza y la confianza pueden cambiar en respuesta a explorar y desafiar las sistema con casos extremos [81]. Por tanto, el usuario puede tener diferentes sentimientos de confianza y desconfianza. durante diferentes etapas de experiencia con cualquier sistema determinado.

Los investigadores definen y miden la confianza de diferentes maneras. Conocimiento del usuario, competencia técnica, familiaridad, confianza, creencias, fe, emociones y apegos personales son términos comunes utilizados analizar e investigar la confianza [94, 136]. Para estos resultados, la confianza del usuario se puede medir preguntando explícitamente sobre las opiniones de los usuarios durante y después de trabajar con un sistema, lo que Se puede realizar a través de entrevistas y cuestionarios. Por ejemplo, Yin et al. [214] estudiaron la importancia de la precisión del modelo en la confianza del usuario. Sus hallazgos muestran que la confianza del usuario en el sistema fue

Tabla 5. Medidas y métodos de evaluación utilizados para medir la confianza del usuario en estudios XAI

Medidas de confianza	Métodos de evaluación
Medidas subjetivas	Autoexplicación y entrevista ([26, 28])
	Cuestionario escala Likert ([17, 26, 28, 160])
Medidas objetivas	Competencia del sistema percibida por el usuario ([160, 169, 214])
	Cumplimiento del usuario con el sistema ([55])
	Comprensibilidad percibida por el usuario ([158, 214])

afectada tanto por la precisión declarada del sistema como por la precisión percibida por los usuarios a lo largo del tiempo. De manera similar, Nourani et al. [160] exploraron cómo la inclusión de explicaciones y el nivel de significado afectarían la percepción de precisión del usuario. Los resultados de sus experimentos controlados muestran que el hecho de que las explicaciones sean significativas para los humanos puede afectar significativamente la percepción de la precisión del sistema, independientemente de la precisión real observada en el uso del sistema. Además, las escalas de evaluación de la confianza podrían ser específicas del contexto de la aplicación de los sistemas y de los propósitos del diseño de XAI. Por ejemplo, múltiples escalas evaluarían por separado la opinión de los usuarios sobre la confiabilidad, previsibilidad y seguridad de los sistemas. En relación con esto, en el artículo de Cahour y Forzy [28] se presenta una configuración detallada de medición de la confianza, que mide la confianza del usuario con múltiples escalas de confianza (constructos de confianza), grabación de video y entrevistas de autoconfrontación para evaluar tres modos de sistema. presentación. Además, para comprender mejor los factores que influyen en la confianza en los agentes adaptativos, Glass et al. [65] estudiaron qué tipos de preguntas a los usuarios les gustaría poder hacerle a un asistente adaptativo. Otros han analizado los cambios en la conciencia del usuario a lo largo del tiempo mostrando la confianza del sistema y la incertidumbre de los resultados del aprendizaje automático en aplicaciones con diferentes grados de criticidad [10, 96].

Múltiples esfuerzos han estudiado el impacto de XAI en el desarrollo de una confianza justificada en los usuarios en diferentes dominios. Por ejemplo, Pu y Chen [169] propusieron un marco organizativo para generar explicaciones y midieron la competencia percibida y la intención del usuario de regresar como medidas de confianza del usuario. Otro ejemplo comparó la confianza del usuario con explicaciones para diferentes objetivos como la transparencia y la explicación de la justificación [158]. Consideraron la comprensibilidad percibida para medir la confianza del usuario y demostraron que las explicaciones transparentes pueden ayudar a reducir los efectos negativos de la pérdida de confianza en situaciones inesperadas.

Al estudiar la confianza del usuario en aplicaciones del mundo real, Berkovsky et al. [17] evaluaron la confianza con varias interfaces de recomendación y estrategias de selección de contenido. Midieron la confianza de los usuarios en un sistema de recomendación de películas con seis construcciones distintas de confianza. También sobre algoritmos de recomendación, Eiband et al. [55] repite el experimento de Langer et al. [114] sobre el papel de las explicaciones “placebicas” (es decir, explicaciones que no transmiten información) en la falta de atención al comportamiento del usuario. Estudiaron si proporcionar explicaciones placebicas aumentaría la confianza del usuario en el sistema de recomendación.
Sus resultados sugieren que el trabajo futuro sobre explicaciones de sistemas inteligentes puede considerar el uso de explicaciones placebicas como base para la comparación con explicaciones generadas por el aprendizaje automático.
También preocupados por la confianza del usuario experto, Bussone et al. [26] midieron la confianza mediante la escala Likert y el pensamiento en voz alta y descubrieron que las explicaciones de los hechos conducen a una mayor confianza del usuario y a una mayor confianza en un sistema de apoyo a las decisiones clínicas. La Tabla 5 resume una lista de métodos de evaluación subjetivos y objetivos para medir la confianza del usuario en los sistemas de aprendizaje automático y sus explicaciones.

Muchos estudios evalúan la confianza del usuario como una propiedad estática. Sin embargo, es esencial tener en cuenta la experiencia y el aprendizaje del usuario a lo largo del tiempo cuando se trabaja con sistemas complejos de IA. Recopilar medidas repetidas a lo largo del tiempo puede ayudar a comprender y analizar la tendencia de los usuarios a desarrollar confianza con la progresión de la experiencia. Por ejemplo, en su estudio, Holliday et al. [89]

Tabla 6. Medidas y métodos de evaluación utilizados para medir el desempeño de tareas hombre-máquina en estudios XAI

Medidas de desempeño	Métodos de evaluación
Rendimiento del usuario	Rendimiento de tareas ([70, 95, 110, 125])
	Rendimiento de la tarea ([110, 113, 125])
	Predicción de fallas del modelo ([70, 105, 194])
Rendimiento del modelo	Precisión del modelo ([108, 130, 165, 172, 194])
	Ajuste y selección de modelo ([131])

evaluó la confianza en múltiples etapas del trabajo con un sistema de minería de texto explicable.
Demostraron que el nivel de confianza del usuario en el sistema variaba con el tiempo a medida que el usuario adquiría más experiencia y familiaridad con el sistema.

Observamos que, aunque nuestra revisión de la literatura no encontró una medición directa de la confianza que comúnmente se priorice en las herramientas de análisis para expertos en datos y aprendizaje automático, la dependencia de los usuarios en las herramientas y la tendencia a continuar usándolas a menudo se consideran parte de la evaluación. canalización durante los estudios de caso. En otras palabras, nuestro resumen no pretende afirmar que los expertos en datos no consideren la confianza, sino que no encontramos que sea un resultado central medido explícitamente en la literatura para este grupo de usuarios.

7.4 M4: Desempeño de tareas entre humanos y IA

Un objetivo clave de XAI es ayudar a los usuarios finales a tener más éxito en sus tareas relacionadas con sistemas de aprendizaje automático [90]. Por lo tanto, el desempeño de las tareas humanas-IA es una medida relevante para los tres grupos de tipos de usuarios. Por ejemplo, Lim et al. [125] midieron el desempeño de los usuarios en términos de tasa de éxito y tiempo de finalización de la tarea para evaluar el impacto de diferentes tipos de explicaciones. Utilizan una interfaz genérica que se puede aplicar a varios tipos de sistemas sensibles al contexto basados en sensores, como la predicción del tiempo. Además, las explicaciones pueden ayudar a los usuarios a ajustar el sistema inteligente a sus necesidades. Kulesza et al. [109] un estudio de explicaciones para un agente recomendador de música encontró un efecto positivo de las explicaciones en la satisfacción de los usuarios con la producción del agente, así como en la confianza de los usuarios en el sistema y su experiencia general.

Otro caso de uso de las explicaciones del aprendizaje automático es ayudar a los usuarios a juzgar la exactitud de la salida del sistema [70, 105, 194]. Las explicaciones también ayudan a los usuarios a depurar programas interactivos de aprendizaje automático según sus necesidades [108, 110]. En un estudio de usuarios finales que interactúan con un sistema clasificador de correo electrónico, Kulesza et al. [108] midieron el rendimiento del clasificador para mostrar que la depuración explicativa beneficia el rendimiento del usuario y de la máquina. Asimismo, Ribeiro et al. [172] descubrieron que los usuarios podían detectar y eliminar explicaciones incorrectas en la clasificación de textos, lo que resultaba en la capacitación de mejores clasificadores con mayor rendimiento y calidad de explicaciones. Para apoyar estos objetivos, Myers et al. [156] diseñaron un marco en el que los usuarios pueden preguntar por qué y por qué no y esperar explicaciones de las interfaces inteligentes. La Tabla 6 resume una lista de métodos de evaluación para medir el desempeño de tareas en escenarios de colaboración entre humanos e IA y ajuste de modelos.

Las herramientas de análisis visual también ayudan a los expertos en el dominio a realizar mejor sus tareas al proporcionar interpretaciones de modelos. Visualizar la estructura del modelo, los detalles y la incertidumbre en los resultados de la máquina puede permitir a los expertos en el campo diagnosticar modelos y ajustar los hiperparámetros a sus datos específicos para un mejor análisis. La investigación en análisis visual ha explorado la necesidad de interpretación de modelos en tareas de análisis de texto [92, 128, 210] y multimedia [24, 33] . Este conjunto de trabajos demuestra la importancia de integrar los comentarios de los usuarios para mejorar los resultados del modelo. Un ejemplo de una herramienta de análisis visual para el análisis de texto es TopicPanaroma [131], que modela un corpus textual como un gráfico temático e incorpora aprendizaje automático y selección de funciones para permitir a los usuarios modificar el gráfico de forma interactiva.

En su procedimiento de evaluación, realizaron estudios de caso con dos expertos en el campo: un especialista en relaciones públicas
Un gerente utilizó la herramienta para encontrar un conjunto de patrones relacionados con la tecnología en los medios de comunicación, y un
profesor analizó el impacto de los medios de comunicación en el público durante una crisis de salud. En el análisis de la transmisión de datos,
Los enfoques automatizados son propensos a errores y requieren que los usuarios expertos revisen los detalles del modelo y la incertidumbre
para una mejor toma de decisiones [18, 179]. Por ejemplo, Goodall et al. [66] presentó a Situ, un visual
Sistema de análisis para descubrir comportamientos sospechosos en los datos de la red cibernética. El objetivo era hacer
Los resultados de la detección de anomalías eran comprensibles para los analistas, por lo que realizaron múltiples estudios de casos.
con expertos en ciberseguridad para evaluar cómo el sistema podría ayudar a los usuarios a mejorar el desempeño de sus tareas. Ahn y Lin
[4] presentan un marco y un diseño analítico visual para ayudar a la equidad basada en datos.
Toma de decisiones. Propusieron FairSight, un sistema de análisis visual para lograr diferentes nociones de
equidad en la clasificación de decisiones mediante la visualización, medición, diagnóstico y mitigación de sesgos.

Además de los expertos en el dominio que utilizan herramientas de análisis visual, los expertos en aprendizaje automático también utilizan herramientas visuales.
análisis para encontrar deficiencias en la arquitectura del modelo o fallas de entrenamiento en DNN para mejorar el
rendimiento de clasificación y predicción [130, 165]. Por ejemplo, Kahng et al. [95] diseñaron un sistema para visualizar el nivel de instancia
y el nivel de subconjunto de activación neuronal en una investigación a largo plazo.
y desarrollo con ingenieros de aprendizaje automático. En sus estudios de caso, entrevistaron a tres
ingenieros de aprendizaje automático y científicos de datos que utilizaron la herramienta e informaron las observaciones clave. De manera
similar, Hohman et al. [86] presentan un sistema interactivo que resume escalablemente y
visualiza qué características ha aprendido un modelo DNN y cómo esas características interactúan en la instancia
predicciones Su sistema analítico visual presenta agregación de activación para descubrir importantes
neuronas y agregación de influencia neuronal para identificar interacciones entre neuronas importantes.
En el caso de las redes neuronales recurrentes (RNN), LSTMVis [193] y RNNVis [143] son herramientas
Interpretar modelos RNN para tareas de procesamiento del lenguaje natural. En otro artículo reciente, Wang
et al. [206] presentaron DNN Genealogy, una herramienta de visualización interactiva que ofrece un resumen visual de las representaciones
DNN.

Otra función fundamental del análisis visual para los expertos en aprendizaje automático es visualizar los procesos de entrenamiento de
modelos [224]. Un ejemplo de herramienta de analítica visual para diagnosticar el proceso de formación de
un modelo generativo profundo es DGMTracker [129], que ayuda a los expertos a comprender el proceso de capacitación al representar
visualmente la dinámica del entrenamiento. Se llevó a cabo una evaluación de DGMTracker en
dos estudios de caso con expertos para validar la eficiencia de la herramienta para respaldar la comprensión del
proceso de formación y diagnosticar un proceso de formación fallido.

7.5 M5: Medidas Computacionales

Las medidas computacionales son comunes en el campo del aprendizaje automático para evaluar la exactitud y la integridad de las técnicas
de interpretabilidad en términos de explicar lo que el modelo ha aprendido.
Herman [78] señala que confiar en la evaluación humana de las explicaciones puede conducir a resultados persuasivos.
explicaciones en lugar de sistemas transparentes debido a la preferencia del usuario por explicaciones simplificadas.
Por lo tanto, este problema lleva al argumento de que la fidelidad de las explicaciones al modelo de caja negra
deben evaluarse mediante métodos computacionales en lugar de estudios con sujetos humanos. La fidelidad de un explicador ad-hoc se
refiere a la exactitud de la técnica ad-hoc al generar explicaciones verdaderas.
(p. ej., corrección de un mapa de prominencia) para predicciones de modelos. Esto conduce a una serie de procesos computacionales.
métodos para evaluar la exactitud de las explicaciones generadas, la coherencia de los resultados de las explicaciones y
fidelidad de las técnicas de interpretabilidad ad-hoc al modelo de caja negra original [175].

En muchos casos, los investigadores del aprendizaje automático suelen considerar la coherencia en los resultados de la explicación,
interpretabilidad computacional y autointerpretación cualitativa de los resultados como evidencia de la corrección de la explicación [162, 215,
217, 226]. Por ejemplo, Zeiler y Fergus [219] analizan la fidelidad de
La visualización de una red CNN por su validez para encontrar debilidades del modelo dio como resultado resultados de predicción mejorados.
En otros casos, comparar una nueva técnica de explicación con la existente

Tabla 7. Medidas de evaluación y métodos utilizados para evaluar la fidelidad de las técnicas de interpretabilidad y la confiabilidad de los modelos entrenados

Medidas computacionales	Métodos de evaluación
Fidelidad explicativa	Experimentos simulados ([172, 173])
	Comprobación de cordura ([101, 162, 177, 215, 217, 226]
	Evaluación comparativa ([178, 183])
Modelo de confiabilidad	Modelo de depuración y entrenamiento ([219])
	Evaluación basada en el ser humano ([43, 134, 149, 187])

Este conjunto de métodos de evaluación es utilizado por expertos en datos y aprendizaje automático para evaluar la exactitud de los métodos de interpretabilidad o evaluar la calidad de la capacitación de los modelos entrenados más allá de las métricas de desempeño estándar.

Se utilizan técnicas de explicación de última generación para verificar la calidad de la explicación [37, 134, 189]. Por ejemplo, Ross et al. [178] diseñaron un conjunto de evaluaciones empíricas y compararon la consistencia y el costo computacional de sus explicaciones con la técnica LIME [172]. En una configuración integral, Samek et al. [183] propusieron un marco para evaluar explicaciones de prominencia para datos de imágenes que cuantifican la importancia de la característica con respecto a la predicción del clasificador. Compararon tres técnicas diferentes de explicación de la prominencia para datos de imágenes (basada en sensibilidad [190], deconvolución [219] y propagación de relevancia por capas [13]) e investigaron la correlación entre la calidad del mapa de prominencia y el rendimiento de la red en diferentes conjuntos de datos de imágenes bajo entrada. perturbación. Por el contrario, Kindermans et al. [101] muestran que las técnicas de interpretabilidad tienen inconsistencias en transformaciones de imágenes simples, por lo que sus mapas de prominencia pueden ser engañosos. Definen una propiedad de invariancia de entrada para la confiabilidad de las explicaciones de los métodos de prominencia. Para ampliar una idea similar, Adebayo et al. [3] proponen tres pruebas para medir la idoneidad de las técnicas de interpretabilidad para tareas que son sensibles a los datos o al modelo mismo.

Otros métodos de evaluación incluyen evaluar la fidelidad de la explicación en comparación con modelos inherentemente interpretables (por ejemplo, regresión lineal y árboles de decisión). Por ejemplo, Ribeiro et al. [172] compararon las explicaciones generadas por el explicador ad-hoc LIME con explicaciones de un modelo interpretable. Crearon explicaciones estándar de oro directamente a partir de modelos interpretables (regresión logística dispersa y árboles de decisión) y las utilizaron para comparaciones en su estudio. Una desventaja de este enfoque es que la evaluación se limita a generar un estándar de oro mediante un modelo interpretable. La evaluación simulada por el usuario es otro método para realizar una evaluación computacional de las explicaciones del modelo. Ribeiro et al. [172] simuló la confianza del usuario en explicaciones y modelos definiendo explicaciones y modelos "no confiables". Probaron una hipótesis sobre cómo los usuarios reales preferirían explicaciones más fiables y elegirían mejores modelos. Posteriormente, los autores repitieron evaluaciones simuladas de usuarios similares en el enfoque de explicación de Anchors [173] para informar la precisión y cobertura de los usuarios simulados para encontrar el mejor clasificador mirando solo las explicaciones.

Schmidt y Biessmann [187] adoptaron un enfoque diferente para cuantificar la calidad de las explicaciones con la intuición humana al definir una métrica de calidad de las explicaciones basada en el tiempo de finalización de las tareas del usuario y la concordancia de las predicciones. Otro ejemplo es el trabajo de Lundberg y Lee [134], quienes compararon el modelo explicativo ad-hoc SHAP con LIME y DeepLIFT [189] con el supuesto de que las buenas explicaciones del modelo deben ser consistentes con las explicaciones de los humanos que entienden el modelo. Lertvittayakumjorn y Toni [118] también presentan tres tareas de usuario para evaluar técnicas de explicación local para la clasificación de textos revelando el comportamiento del modelo a los usuarios humanos, justificando las predicciones y ayudando a los humanos a investigar predicciones inciertas. Se ha implementado una idea similar en [149] mediante la comparación de características de una verdad fundamental

y explicación del modelo. Proporcionan un punto de referencia anotado por el usuario para evaluar las explicaciones de las instancias de aprendizaje automático. Posteriormente, Poerner et al. [166] utilizan este punto de referencia como verdad fundamental anotada por humanos en comparación con la evaluación de explicaciones de contexto pequeño (nivel de palabra) y contexto grande (nivel de oración). Los puntos de referencia anotados por el usuario pueden ser valiosos al considerar el significado humano de las explicaciones, aunque la discusión de Das et al. [43] implica que los modelos de aprendizaje automático (modelos de atención de respuesta a preguntas visuales en su caso) no parecen observar las mismas regiones que los humanos. Introducen un conjunto de datos de atención humana [42] (colección de datos de seguimiento del ratón) y evalúan mapas de atención generados por modelos de última generación en comparación con humanos.

Las técnicas de interpretabilidad también permiten medidas cuantitativas para evaluar la confiabilidad del modelo (por ejemplo, equidad, confiabilidad y seguridad del modelo) a través de sus explicaciones. La confiabilidad de un modelo representa un conjunto de objetivos específicos de un dominio, como la equidad (mediante el aprendizaje justo de características), la confiabilidad y la seguridad (mediante el aprendizaje sólido de características). Por ejemplo, Zhang et al. [220] presentan un caso de uso de explicaciones de aprendizaje automático para encontrar fallas de aprendizaje de representación causadas por posibles sesgos en el conjunto de datos de entrenamiento. Su técnica explora las relaciones entre pares de atributos según sus patrones de inferencia. Además, Kim et al. [99] presentaron pruebas cuantitativas de modelos de aprendizaje automático mediante sus explicaciones. En su técnica de vectores de activación de conceptos, el modelo se puede probar para conceptos específicos (por ejemplo, patrones de imágenes) y una puntuación vectorial muestra si el modelo está sesgado hacia ese concepto. Posteriormente ampliaron su explicación global basada en conceptos del aprendizaje de la representación de modelos para el descubrimiento sistemático de conceptos que son significativos para los humanos e importantes para la predicción del modelo [63]. Utilizaron experimentos con sujetos humanos para evaluar los conceptos ap
La Tabla 7 resume una lista de métodos de evaluación para medir la fidelidad de las técnicas de interpretabilidad y la confiabilidad del modelo con técnicas computacionales.

8 MARCO DE DISEÑO Y EVALUACIÓN DE XAI

La variedad de diferentes objetivos de diseño de XAI (Sección 6) y métodos de evaluación (Sección 7) de nuestra revisión sugiere la necesidad de diversos conjuntos de técnicas para construir sistemas XAI de extremo a extremo. Sin embargo, generalmente es insuficiente tomar por separado las prácticas de diseño y los métodos de evaluación. Una ventaja holística y más práctica requerirá la consideración de las dependencias entre los objetivos de diseño y los métodos de evaluación e informará cuándo elegir entre ellos durante los ciclos de diseño. Anteriormente, se propusieron varios modelos y pautas para el diseño y evaluación de interfaces de usuario interactivas con IA [7, 54] y sistemas de análisis visual [155] para ayudar a los diseñadores en el proceso de diseño. Sin embargo, los desafíos para generar explicaciones útiles de aprendizaje automático y presentarlas a través de una interfaz adecuada que se alinee con los resultados objetivo requieren un marco de flujo de trabajo multidisciplinario.

Por lo tanto, con base en nuestro análisis de trabajos anteriores, proponemos un marco de diseño y evaluación para sistemas XAI. El impulso de este marco es el deseo de organizar y relacionar el conjunto diverso de objetivos de diseño y métodos de evaluación existentes en un modelo unificado. El marco tiene como objetivo brindar orientación sobre qué medidas de evaluación son apropiadas para usar en cada etapa de diseño del sistema XAI. La Figura 4 resume el marco como un modelo anidado para el diseño y evaluación del sistema XAI de un extremo a otro. La formulación del modelo como capas se relaciona con los objetivos centrales de diseño y los intereses de evaluación de las diferentes comunidades de investigación (como se identifican en la revisión de la literatura) para ayudar a promover el progreso interdisciplinario en la investigación XAI. El modelo está estructurado para respaldar los pasos de diseño del sistema comenzando desde la capa externa (Objetivos del sistema XAI), luego abordando las necesidades del usuario final en la capa intermedia (Interfaz explicable) y finalmente enfocándose en algoritmos interpretables subyacentes en la capa más interna (Algoritmos interpretables).). El modelo anidado está organizado con un Polo de Diseño que se centra en los objetivos y opciones de diseño, y un Polo de Evaluación que presenta métodos y medidas de evaluación apropiados para cada capa. Nuestro marco sugiere ciclos iterativos de diseño y evaluación para cubrir aspectos algorítmicos y humanos.

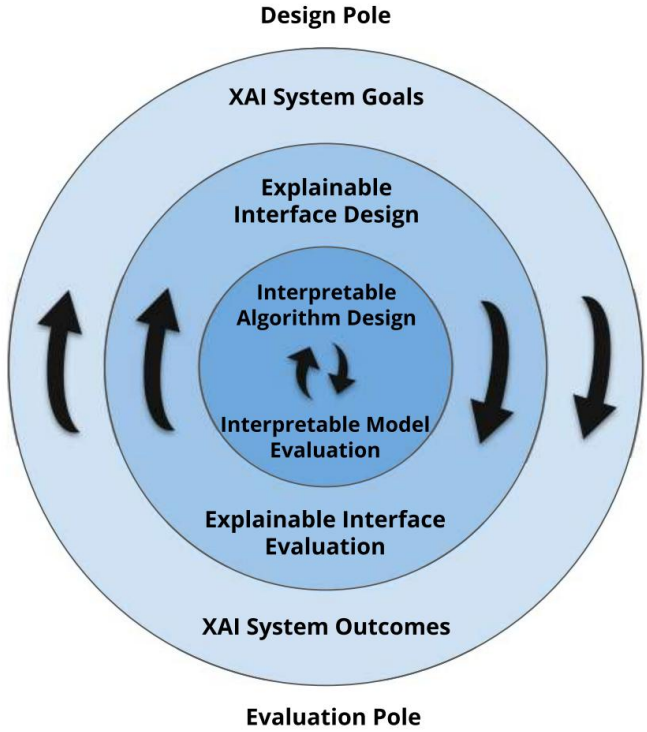


Fig. 4. Marco de diseño y evaluación de XAI: nuestro modelo anidado para el diseño y evaluación de sistemas de aprendizaje automático explicables. La capa exterior demuestra los objetivos de diseño a nivel del sistema que se combinan con la evaluación de resultados XAI de alto nivel. La capa intermedia muestra una interfaz de usuario explicable y un paso de diseño de visualización junto con medidas apropiadas de evaluación de satisfacción y comprensión del usuario. La capa más interna presenta el diseño y la evaluación de algoritmos de aprendizaje automático interpretables y confiables.

de los sistemas XAI. En esta sección, detallamos los detalles del marco anidado y brindamos pautas sobre su uso para el diseño de sistemas XAI multidisciplinarios.

Ejemplo de estudio de caso: para mostrar un ejemplo práctico del uso del marco, también incluimos un estudio de caso de un esfuerzo colaborativo de diseño y desarrollo para un sistema XAI. En el escenario del estudio de caso, un equipo multidisciplinario de investigadores diseñó un sistema XAI para la detección de noticias falsas para lectores diarios de noticias no expertos (ni expertos en IA ni analistas de noticias). El equipo de diseño planeó agregar una función de Asistente XAI a un sitio web de lectura e intercambio de noticias para realizar la detección de noticias falsas. El diseño del sistema consistió en una interfaz de lectura de noticias equipada con el asistente de noticias XAI (asistente de noticias) para ayudar al usuario a identificar noticias falsas mientras revisa noticias y artículos. El caso presentado es el resultado de una investigación en curso realizada durante un período de un año por un equipo de ocho investigadores universitarios con experiencia en HCI, visualización e inteligencia artificial. En las siguientes subsecciones, cada directriz marco va seguida de un ejemplo de su aplicación en nuestro estudio de caso.

8.1 Capa de objetivos del sistema XAI

Como los diseñadores de un equipo multidisciplinario tienen diferentes roles y prioridades en la construcción de un sistema XAI, sugerimos comenzar el ciclo de diseño del sistema desde la capa de objetivos XAI (la capa exterior de la Figura 4) para caracterizar los objetivos del diseño y las expectativas del sistema. En concreto, este paso implica

identificar el propósito de la explicación y elegir qué explicar al usuario final objetivo y
aplicación dedicada. Los refinamientos iterativos entre el objetivo XAI (polo superior) y la evaluación del sistema (polo inferior)
presentan cómo las medidas de evaluación pareadas ayudan a mejorar el diseño del sistema. Nosotros
Organice las siguientes pautas para la capa de objetivos XAI.

Al comienzo del proceso de diseño del sistema, el equipo deberá especificar la explicabilidad.
requisitos para cada capa del marco en función de las métricas de evaluación. La explicabilidad
Los requisitos están destinados a satisfacer los objetivos generales del sistema definidos por las necesidades del usuario (o cliente), y
a veces regulaciones, leyes y estándares de seguridad. Posteriormente, el paso de evaluación en cada ciclo de diseño.
Hará que el equipo revise los requisitos iniciales del sistema XAI. La suficiencia de la evaluación
Los resultados en comparación con los requisitos de diseño iniciales sirven como un indicador clave de si
para detener o continuar la iteración del diseño. Sin embargo, dado que se utilizan muchas medidas subjetivas en el
proceso, es importante elegir una línea de base de evaluación adecuada (ver Sección 9.4) para realizar un seguimiento
progreso durante los ciclos de diseño.

Directriz 1: Determinar los objetivos del sistema XAI: Identificar y establecer objetivos y expectativas claros de un sistema XAI es el
primer paso en el proceso de diseño. Los objetivos de diseño de XAI podrían ser
impulsado por muchas motivaciones, como mejorar la experiencia del usuario en un sistema existente, promover hallazgos científicos
[107, 120] o adherirse a nuevas regulaciones [198]. En la Sección 6, revisamos ocho
Objetivos principales (G1–G8) para los sistemas XAI. Además, ordenar la prioridad de objetivos en casos con múltiples
Los objetivos de diseño pueden ser beneficiosos en los siguientes pasos del proceso (ver Directriz 2). Dado el hecho de que
diferentes tipos de usuarios y aplicaciones XAI están interesados en diversos objetivos de diseño, es importante
Establecer estos objetivos al principio del proceso de diseño para identificar y alinearlos con el diseño apropiado.
principios. Un error en esta etapa es elegir los objetivos de XAI sin considerar el grupo de usuarios finales, las limitaciones algorítmicas
y las preferencias del usuario en el contexto de la aplicación. Sobrepasando XAI
Los objetivos podrían perjudicar los resultados de la evaluación en el avance del proceso de diseño.

> Aplicación en estudio de caso: en el primer paso de nuestro estudio de caso con una aplicación de curación de noticias,
> el equipo comenzó identificando los principales objetivos y expectativas para la XAI.
> asistente de noticias. El diseño se centró en usuarios finales novatos sin ninguna experiencia particular.
> El objetivo del diseño de XAI era mejorar la confianza del usuario y el modelo mental de predicciones de noticias.
> a través de un diseño explicable. El equipo planteó la hipótesis de que los usuarios finales confiarían y confiarían
> en el asistente de detección de noticias falsas, dado que el nuevo XAI es capaz de proporcionar
> Explicaciones para cada noticia. Además, el equipo esperaba que los usuarios pudieran utilizar
> las explicaciones para aprender las debilidades y fortalezas del modelo para proporcionar retroalimentación al
> equipo de desarrolladores.

Directriz 2: Decidir qué explicar: El segundo paso en el diseño del sistema XAI es identificar
"qué explicar" al usuario para lograr los objetivos XAI iniciales (ver Directriz 1) del sistema. Revisamos múltiples técnicas de
interpretabilidad del aprendizaje automático y tipos de explicaciones en
Secciones 4.1, 4.2 y 4.3 que pueden proporcionar diferentes tipos de información al usuario. A pesar de
Los marcos de diseño basados en teoría discuten los mecanismos de explicación impulsados por el razonamiento humano.
semántica [126], métodos centrados en el usuario para identificar explicaciones útiles, como encuestas en línea, entrevistas y
observaciones de usuarios (p. ej., [26, 138]) para comprender cuándo y qué es necesario explicar
para que los usuarios comprendan mejor y confíen en los sistemas inteligentes. Los experimentos preliminares son valiosos en los
primeros pasos del ciclo de diseño para identificar y reducir las opciones de explicación para el
usuario para satisfacer los objetivos de diseño. Un enfoque típico para evaluar la efectividad y utilidad de la elección de explicación
en experimentos centrados en el usuario es comparar el modelo mental del usuario sobre la explicación.

Sistema con y sin componentes explicativos. Sobre este tema, Lim y Dey [124] realizaron experimentos para descubrir qué tipo de información les interesa a los usuarios, en diferentes escenarios de aplicaciones conscientes del contexto del mundo real. Stumpf y cols. [195] también realizaron entrevistas a usuarios finales para identificar las percepciones y expectativas de los usuarios a partir de una interfaz interpretable, así como para encontrar los principales puntos de decisión en los que los usuarios pueden necesitar explicaciones. En otro trabajo, Haynes et al. [74] proporcionan una revisión y estudios que incorporan diferentes explicaciones (explicaciones operativas, ontológicas, mecanicistas y de fundamentos de diseño) en sistemas inteligentes. De manera similar, el diseño de análisis visual implica entrevistas con expertos y grupos focales en el camino del diseño para identificar los objetivos del diseño [155].

El proceso de diseño en este paso implica restricciones de implementación algorítmica como "qué se puede explicar" al usuario. Por ejemplo, las explicaciones globales de una DNN pueden no ser factibles ni comprensibles debido a la gran cantidad de variables en el gráfico. Además, las investigaciones muestran que las explicaciones de instancias de una DNN carecen de integridad y pueden no presentar características destacadas en los casos [3]. Estas limitaciones y puntos de decisión podrían resolverse mediante grupos enfocados, lluvias de ideas y entrevistas entre los diseñadores de modelos y los diseñadores de interfaces del equipo. Por lo tanto, un error de diseño a la hora de elegir una explicación es no tener en cuenta las limitaciones de las técnicas de interpretabilidad.

> Aplicación en un caso de estudio: en nuestro escenario, la curación de noticias eficiente requería la detección de noticias falsas con la ayuda de nuestro asistente XAI. En el análisis de lo que debería explicar el sistema, el equipo de diseño decidió identificar opciones de explicación candidatas útiles e impactantes. Comenzamos revisando la investigación de aprendizaje automático sobre la detección de información falsa (p. ej., rumores, engaños, noticias falsas, clickbait), así como la investigación de HCI sobre fuentes de noticias y sistemas de búsqueda de noticias para identificar atributos clave para la verificación de la veracidad de las noticias [148] . Dado que los usuarios finales no son expertos, se necesita información explicativa para limitar los detalles técnicos. A continuación, los diseñadores de interfaces de usuario y de aprendizaje automático del equipo discutieron las opciones de explicación de los candidatos y las limitaciones algorítmicas en las técnicas de interpretabilidad. Es decir, algunas opciones sobre qué explicar pueden no ser del todo posibles dada la interpretabilidad de los modelos existentes, y el equipo necesitaba considerar si técnicas de aprendizaje alternativas podrían proporcionar mejores explicaciones o si el equipo de diseño necesitaría encontrar formas significativas de explicarlo. Explique la información que estaba disponible en el modelo.

Directriz 3: Evaluar los resultados del sistema: La evaluación de los resultados del sistema XAI es el paso final en el proceso de evaluación. La Figura 4 muestra cómo la evaluación final del resultado del sistema se combina con los objetivos de diseño iniciales en la capa exterior de nuestro marco. El objetivo principal de esta etapa es evaluar cuantitativa y cualitativamente la efectividad del sistema XAI para los objetivos XAI a nivel del sistema inicialmente establecidos. Claramente, la evaluación de los resultados finales del sistema podría verse influenciada por el diseño de la interfaz de usuario explicable (capa intermedia) y el diseño de algoritmos interpretables (capa más interna). Por ejemplo, evaluar el resultado de un algoritmo de aprendizaje automático interpretable recién nacido utilizando sujetos humanos a través de un débil estudio de usuario en el laboratorio o de colaboración abierta puede no ser significativo o productivo para los resultados del sistema XAI si los cambios computacionales centrales aún están en progreso y, en última instancia, podrían cambiar toda la interpretabilidad del modelo. y formato de explicación más adelante. Además, los cambios en el usuario objetivo podrían afectar los resultados de la evaluación en esta etapa. Por ejemplo, un sistema diseñado para principiantes puede no satisfacer las necesidades de un usuario experto y, por tanto, no mejoraría el rendimiento como se esperaba. Las medidas de evaluación en esta capa dependen de los objetivos del diseño, el dominio de la aplicación y los usuarios objetivo. Ejemplos de medidas de evaluación para los resultados finales del sistema incluyen la confianza del usuario [169] y la confianza en el sistema [17], el desempeño de tareas humano-máquin

conciencia del usuario [96] y comprensión del usuario de sus datos personales [170]. Un proceso eficaz para la evaluación de los resultados de la XAI de alto nivel es dividir el objetivo de la evaluación en múltiples medidas y métricas bien definidas. De esta manera, el equipo puede realizar estudios de evaluación en diferentes pasos utilizando métodos válidos en una configuración controlada. Por ejemplo, en la evaluación de la confiabilidad de los sistemas XAI, se podrían medir varios factores de la confianza humana durante y después de un período de experiencia del usuario con el sistema XAI. Además, se utilizan medidas computacionales (Sección 7.5) para examinar la fidelidad de los métodos de interpretabilidad y la confiabilidad del modelo con métricas objetivas. Un posible error en la evaluación de los resultados del sistema XAI es realizar la evaluación sin considerar la confiabilidad del modelo y la exactitud de las explicaciones desde la capa del modelo interpretable (ver Directriz 7) y la comprensibilidad y utilidad de la explicación desde la capa de la interfaz de usuario (ver Directriz 5).

Aplicación en el estudio de caso: en nuestro estudio de caso con revisión y curación de noticias, necesitábamos evaluar nuestro asistente de noticias XAI con usuarios no expertos que recopilarían noticias mientras marcaban artículos de noticias falsas. En el paso de evaluación, el equipo realizó múltiples estudios a gran escala con sujetos humanos con participantes novatos reclutados a través de Amazon Mechanical Turk para trabajar con nuestro sistema de lectura de noticias. Tenga en cuenta que tanto la interfaz explicable como el algoritmo interpretable pasaron múltiples iteraciones de diseño y prueba antes de este paso de evaluación. Una decisión importante para esta evaluación fue cómo estructurar la duración y la complejidad de la tarea del usuario mientras se prueba adecuadamente toda la gama de funcionalidades del sistema. La tarea se diseñó con preguntas integradas para ayudar a recopilar datos subjetivos además de los datos objetivos de rendimiento del usuario. Se eligieron múltiples medidas de evaluación para los resultados del sistema, incluida la confianza subjetiva del usuario en el asistente de noticias, la tasa de acuerdo del usuario con el asistente de noticias, la veracidad de las noticias compartidas por los usuarios y la precisión del usuario al adivinar el resultado del asistente de noticias. Tanto el análisis cualitativo como cuantitativo de los comentarios de los usuarios y los datos de interacción fueron valiosos para la evaluación de los resultados del sistema. Los resultados y análisis de estas evaluaciones ayudaron al equipo a comprender la efectividad de los elementos XAI (tanto en el algoritmo como en la interfaz) para los objetivos iniciales del sistema.

8.2 Capa de diseño de interfaz de usuario

La capa intermedia de nuestro marco se ocupa de diseñar y evaluar una interfaz o visualización explicable para que el usuario interactúe con el sistema XAI. El diseño de interfaz para explicaciones consiste en presentar explicaciones del modelo a partir de algoritmos interpretables a los usuarios finales en términos de su formato de explicación y diseño de interacción. La importancia de esta capa es satisfacer los requisitos de diseño y las necesidades deben determinarse en la capa de diseño del sistema XAI (ver Directriz 2). Una traducción elegante de explicaciones generadas por máquinas (por ejemplo, explicaciones verbales, numéricas o visuales) necesita explicaciones satisfactorias y comprensibles para los humanos cuidadosamente diseñadas en la interfaz de usuario. En la Sección 4.4, revisamos múltiples tipos de formatos de explicación para integrar elementos XAI en la interfaz de usuario. El movimiento iterativo entre el polo de Diseño y el polo de Evaluación en esta capa presenta un refinamiento del diseño en la búsqueda de un estado objetivo deseado.

Directriz 4: Decidir cómo explicar: identificar los formatos de explicación candidatos para el sistema objetivo y el grupo de usuarios es el primer paso para ofrecer explicaciones de aprendizaje automático a los usuarios finales. El proceso de diseño puede tener en cuenta diferentes niveles de complejidad, duración, estado de presentación (por ejemplo, permanente o bajo demanda) y opciones de interactividad dependiendo de la aplicación y el tipo de usuario. El formato de las explicaciones en la interfaz es particularmente importante para mejorar la comprensión del usuario.

algoritmos subyacentes. Los estudios demuestran que si bien las representaciones interactivas detalladas y complejas puede tener como objetivo comunicar las explicaciones a los usuarios expertos, usuarios novatos en IA de un sistema XAI Prefieren interfaces de explicación y representación más simplificadas [112]. La satisfacción del usuario con el diseño de la interfaz también es otro factor crítico en la participación del usuario en los componentes de la interfaz [154]. Además, el diseño de interacción para interfaces explicables puede permitir que un usuario se comunique con el sistema para ajustar las explicaciones y podría respaldar mejor la inspección del sistema por parte del usuario [110]. La investigación sobre el diseño de interfaces inteligentes presenta múltiples métodos de diseño, como el wireframing y creación de prototipos de baja fidelidad (por ejemplo, [26, 138]) que también podrían adaptarse al diseño de interfaz explicable. Además, las pautas de diseño existentes y el conocimiento de las mejores prácticas para interfaces infundidas con IA (por ejemplo, [7]) y visualizaciones (por ejemplo, [140]) podrían usarse en esta etapa para aprovechar sistemas similares para un diseño de interfaz explicable. Además de las explicaciones del modelo, proporcionar incertidumbre en la predicción también ha sido identificado como un factor importante tanto para los usuarios finales generales como para los usuarios expertos en datos [181]. Por ejemplo, Kay et al. [96] presentó el ciclo de diseño completo para una interfaz de visualización de incertidumbre en una aplicación de tiempo de llegada de autobús. Su proceso de diseño incluyó encuestas para identificar el uso. requisitos, desarrollar diseños alternativos, ejecutar pruebas de usuario y evaluación final del usuario comprensión de los resultados del aprendizaje automático.

> Aplicación en estudio de caso: determinar cómo explicar los resultados de la clasificación de noticias a usuarios finales no expertos, el equipo de diseño de la interfaz de usuario inició el proceso revisando los objetivos iniciales del sistema y los tipos de explicación. Luego, el equipo continuó con múltiples bocetos de interfaz que coincidían con la aplicación prevista y las tareas del usuario. Durante la inicial pasos de diseño, el equipo trató de mantener un equilibrio entre la complejidad de la interfaz y la utilidad de la explicación eligiendo entre los tipos de explicación disponibles de nuestra interfaz interpretable.
> Algoritmos de aprendizaje automático. A continuación, se implementaron maquetas de los tres diseños principales para probarlas con un pequeño número de participantes. Cada maqueta tenía una diferente disposición de los datos, flujo de tareas del usuario y formato de explicación para la interfaz del asistente de noticias. Nuestros experimentos con sujetos humanos en esta etapa se basaron en observaciones de usuarios y entrevistas posteriores al uso para recopilar comentarios cualitativos sobre la comprensión de los participantes y la satisfacción subjetiva de los componentes de la explicación y los arreglos de la interfaz.
> Las entrevistas dieron como resultado la selección del diseño más comprensible y concluyente. entre las opciones disponibles para continuar (ver Lineamientos 5).

Directriz 5: Evaluar Explicación Utilidad: Este paso de evaluación de capa intermedia se puede utilizar junto con varias medidas para ayudar a evaluar la comprensión del usuario del XAI subyacente inteligente algoritmos. Una serie de evaluaciones centradas en el usuario de una interfaz explicable con múltiples objetivos y Se podrían realizar niveles de granularidad para medir lo siguiente:

(1) Comprensión del usuario de la explicación.
(2) Satisfacción del usuario con la explicación.
(3) Modelo mental de usuario del sistema inteligente.

Las evaluaciones en la capa intermedia son particularmente importantes debido al impacto en los resultados del sistema XAI (capa externa) y al verse afectadas por los resultados interpretables del modelo (capa más interna). Específicamente, las medidas de evaluación en esta etapa pueden informar qué tan bien los usuarios entienden lo interpretable. sistema; sin embargo, la validez del diseño en este paso también puede reflejarse en resultados XAI de nivel superior (es decir, evaluación de la capa externa), como la confianza del usuario y el desempeño de la tarea. Tenga en cuenta que la comprensión del usuario de un sistema XAI podría limitarse a partes del sistema en lugar de a todo el sistema; similarmente, La comprensión puede limitarse a un subespacio de escenarios en lugar de a todos los escenarios posibles.

Las tres medidas de evaluación introducidas para este paso podrían usarse en múltiples ciclos iterativos para mejorar el diseño general de la interfaz explicable. Por ejemplo, Saket et al. [182] estudia la comprensión de los usuarios sobre la codificación de visualización y la eficacia de la codificación gráfica interactiva para el usuario final.
Por otro lado, la satisfacción del usuario con el tipo y formato de explicación depende de factores como la criticidad de la aplicación dirigida y la carga cognitiva preferida del usuario [48]. La evaluación del modelo mental del usuario también es una forma eficaz de medir la utilidad de las interfaces explicables. Las tablas 3 y 4 presentan una lista de medidas para evaluar interfaces explicables en este paso. La elección de la línea de base es otro factor importante en la evaluación de interfaces explicables. Normalmente, se utiliza una combinación de análisis cualitativos y cuantitativos para medir los efectos de los componentes de la explicación (en comparación con un sistema no explicable) o para comparar múltiples tipos de explicaciones. Sin embargo, la elección de explicaciones de lugar bic se ha propuesto como base de evaluación para una medición más precisa del contenido de la explicación [55]. En el caso de la revisión de expertos, la evaluación del modelo mental de un experto en el dominio comúnmente implica una comparación con el modelo mental del experto en IA y una descripción de "cómo funciona el modelo". La Sección 9.4 revisa las opciones comunes de líneas de base de verdad sobre el terreno en los estudios de evaluación XAI. En todos los enfoques, las actualizaciones de los componentes explicativos de la interfaz requieren una evaluación de su impacto en la experiencia y la comprensibilidad del usuario. Sin embargo, las métricas y la profundidad de la evaluación varían durante los ciclos de evaluación a medida que el equipo reduce necesidades específicas.
Finalmente, un posible error en la evaluación de interfaces explicables es buscar medidas amplias de los resultados de XAI (ver Directriz 3) en lugar de centrarse en un alcance más limitado de componentes e interacciones de explicación.

> Aplicación en el estudio de caso: en nuestro estudio de caso, los diseñadores de interfaces comenzaron la evaluación de los componentes de explicación candidatos mediante una serie de pequeños estudios con un diseño de medidas repetidas para que el mismo participante del estudio pudiera experimentar diferentes diseños de explicación en una sesión. A continuación, analizamos los datos cuantitativos y cualitativos recopilados de los usuarios finales para elegir diseños y rutas candidatos para mejorar aún más la interfaz de los componentes explicables. Las discusiones con el equipo de aprendizaje automático también ayudaron a encontrar fuentes de limitaciones en la técnica de interpretabilidad que posiblemente podrían afectar la satisfacción del usuario. Después de los ciclos iniciales de revisión, recopilamos una ronda de revisiones de expertos externos e internos para actualizar la metodología del estudio y los detalles de la recopilación de datos de acuerdo con el progreso del proyecto.

8.3 Capa de diseño de algoritmos interpretables La capa

más interna de nuestro marco implica el diseño de algoritmos interpretables que sean capaces de generar explicaciones para los usuarios. El último paso de diseño en el marco de nuestro sistema XAI es la elección de la técnica de interpretabilidad (polo de diseño) para generar los tipos de explicación descritos. Sin embargo, evaluar la explicación generada (polo de evaluación) es el primer paso de evaluación antes de las evaluaciones humano-sujeto en la interfaz explicable. Idealmente, las técnicas de interpretabilidad deberían generar explicaciones de acuerdo con los requisitos en el paso de diseño de interfaz explicable (ver Directriz 4); sin embargo, la elección de la técnica de interpretabilidad depende del dominio y conlleva limitaciones de implementación. Por ejemplo, si bien se desean modelos superficiales por su alta interpretabilidad, estos modelos generalmente no funcionan bien en casos de datos complejos y de alta dimensión como imágenes y texto. Por otro lado, las predicciones altamente precisas en modelos de caja negra (por ejemplo, DNN y modelos de bosque aleatorios) requieren posprocesamiento y algoritmos ad-hoc para generar explicaciones. El enfoque ad hoc también tiene limitaciones tanto en la elección del tipo de explicación como en la necesidad de una validación de integridad [3] y fidelidad [172] en comparación con el modelo original. Esto muestra que los diseñadores de aprendizaje automático no solo deben considerar la

equilibrio entre la interpretabilidad y el rendimiento del modelo, pero también debe considerar la fidelidad del explicador ad hoc al modelo de caja negra. Sugerimos las dos siguientes pautas de diseño y evaluación para esta capa.

Directriz 6: Técnica de interpretabilidad del diseño: El diseño de algoritmos de toma de decisiones interpretables comienza con la elección del modelo de aprendizaje automático. Los modelos de aprendizaje automático superficial (por ejemplo, modelos lineales y árboles de decisión) tienen interpretabilidad intrínseca debido al bajo número de variables y la simplicidad del modelo. Para modelos más complejos (por ejemplo, bosque aleatorio y DNN), se necesitan técnicas explicativas ad hoc (consulte la Sección 4.2) para generar explicaciones. Sin embargo, la elección del modelo de aprendizaje automático (es decir, superficial versus profundo) está limitada por el desempeño de un modelo en el dominio de datos.
En segundo lugar, las técnicas de explicación ad hoc tienen ciertas limitaciones en su tipo de explicación. La importancia de elegir la combinación correcta de modelo y explicador radica en su impacto a la hora de proporcionar explicaciones útiles (ver Directriz 4) y confiables para los usuarios finales.

La investigación sobre aprendizaje automático ha propuesto varias explicaciones ad hoc para generar explicaciones de "por qué" (p. ej., atribución de características [99, 134]), explicaciones de "cómo" (p. ej., lista de reglas [119, 205]), "qué más" explicación (p. ej., instancia de entrenamiento similar [98, 135]) y tipos de explicación "¿Qué pasaría si?" (p. ej., análisis de sensibilidad [219]) . Sin embargo, a pesar de una investigación sustancial sobre técnicas interpretables de aprendizaje automático, una cuestión central en las explicaciones de los modelos es la diferencia entre la lógica de toma de decisiones del modelo de aprendizaje automático y la percepción humana como receptor [100, 223] .

> Aplicación en el estudio de caso: en nuestro estudio de caso de detección de noticias falsas, el equipo de diseño de interfaz explicable había discutido previamente las opciones de explicación de los candidatos con el equipo de diseño de aprendizaje automático (consulte las Directrices 2 y 4). Por lo tanto, en este paso se realizó una revisión final de las explicaciones generadas por el modelo y una evaluación de las limitaciones de implementación. Por ejemplo, eliminar características similares al ruido de los mapas de prominencia, normalizar las puntuaciones de atribuciones y resolver explicaciones contradictorias entre un conjunto de modelos fueron los principales obstáculos a la implementación que se resolvieron en este paso. Específicamente, como punto de decisión para equilibrar la claridad y la fidelidad de las explicaciones, el equipo decidió utilizar filtros heurísticos para eliminar características con una puntuación de atribución muy baja en aras de la simplicidad de la presentación.

Directriz 7: Evaluar la confiabilidad del modelo: Evaluar el aprendizaje automático interpretable es el primer paso de evaluación en nuestro marco debido a su impacto en la medida de evaluación de la capa externa.
La gran importancia de este paso de evaluación surge de la posibilidad de que cualquier falta de confiabilidad de la interpretabilidad en esta capa interna se propague a todas las demás capas externas. Esta propagación involuntaria de errores puede dar lugar a decisiones problemáticas sobre el diseño de la capa exterior, así como a resultados de evaluación engañosos. Discutimos dos objetivos principales de evaluación para la capa más interna:

(1) Evaluación de la confiabilidad del modelo.
(2) Evaluación de la fidelidad del explicador ad hoc.

El primer objetivo de la evaluación tiene como objetivo utilizar técnicas de interpretabilidad como herramienta de depuración para analizar la confiabilidad del modelo en el aprendizaje de conceptos más allá de las medidas generales de desempeño [99]. Ejemplos de validación de la confiabilidad del modelo incluyen la evaluación de la confiabilidad del modelo en la evaluación de riesgos financieros [60], la equidad del modelo en aplicaciones de influencia social [220] y la seguridad del modelo para su funcionalidad prevista [22]. Los investigadores también han propuesto varias técnicas de regularización para mejorar el aprendizaje de características confiables en modelos de aprendizaje automático [76, 178]. A continuación, el segundo objetivo de la evaluación apunta a la fidelidad de las técnicas explicativas ad hoc al modelo de caja negra. Investigación

muestra que diferentes técnicas de interpretabilidad ad-hoc tienen inconsistencias y pueden ser engañosas [3]. Evaluar la confiabilidad de la explicación puede verificar la fidelidad del explicador en términos de qué tan bien representa el modelo de caja negra (ver Sección 7.5).

Aplicación en el estudio de caso: en nuestro estudio de caso, prestamos especial atención a la revisión cualitativa de las explicaciones del modelo después de cada iteración del diseño. Nuestra calidad inicial La revisión de las explicaciones del modelo condujo a la limpieza del conjunto de datos a través de una búsqueda heurística dirigida a la eliminación de ejemplos mal etiquetados y artículos de noticias no relacionados. una mejora a El rendimiento del modelo se logró después de la limpieza del conjunto de datos. Luego, después de la primera ronda de En la evaluación humano-sujeto de la interfaz explicable (ver Directriz 5), el equipo identificó efectos negativos de las explicaciones de palabras clave con bajas puntuaciones de atención por parte de los usuarios finales. El equipo decidió utilizar un umbral más bajo para visualizar mapas de atención para reducir desorden y “explicaciones ruidosas” para los usuarios finales. Finalmente, después de una ronda del resultado XAI En la evaluación (ver Directriz 3), el análisis de los modelos mentales de los usuarios reveló que un desequilibrio en el conjunto de datos entre las "noticias falsas" y las "noticias verdaderas" estaba causando un sesgo en el modelo en que el modelo generalmente tenía más confianza en predecir noticias falsas que noticias verdaderas.

9 DISCUSIÓN

En nuestra revisión, discutimos múltiples objetivos de diseño de XAI y medidas de evaluación apropiadas para varios tipos de usuarios específicos. La Tabla 2 presenta la categorización de los métodos de diseño y evaluación existentes seleccionados que organiza la literatura según tres perspectivas: objetivos de diseño, métodos de evaluación, y los usuarios objetivo del sistema XAI. Nuestra categorización reveló la necesidad de un esfuerzo interdisciplinario para diseñar y evaluar sistemas XAI. Para abordar estos problemas, propusimos una Marco de diseño y evaluación que conecta los objetivos de diseño y los métodos de evaluación para el diseño de sistemas XAI de extremo a extremo, como se presenta a través de un modelo (Figura 4) y pautas. En esta sección, nosotros discutir consideraciones adicionales para que los diseñadores de XAI se beneficien del conjunto de conocimientos del diseño y evaluación del sistema XAI. Las siguientes recomendaciones apoyan y promueven diferentes capas. del marco de diseño y evaluación propuesto.

9.1 Emparejamiento de objetivos de diseño con métodos de evaluación

Es esencial utilizar medidas adecuadas para evaluar la eficacia de los elementos de diseño. A El error común al elegir medidas de evaluación en sistemas XAI es que a veces se utiliza la misma medida de evaluación para múltiples objetivos de diseño. Una solución sencilla para abordar este problema es distinguir entre mediciones mediante el uso de múltiples escalas para capturar diferentes atributos en cada una objetivo de evaluación. Por ejemplo, el concepto de confianza del usuario consta de múltiples construcciones [28] que podría medirse con escalas separadas en cuestionarios y entrevistas (ver Sección 7.3). Usuario Las mediciones de satisfacción también podrían diseñarse para diversos atributos, como la comprensibilidad de las explicaciones, la utilidad de las explicaciones y la suficiencia de los detalles [85] para apuntar a objetivos específicos. cualidades explicativas (ver Sección 7.2).

Una manera eficiente de combinar objetivos de diseño con medidas de evaluación apropiadas es equilibrar diferentes métodos de diseño y tipos de evaluación en ciclos iterativos de diseño. Gestionar las compensaciones entre métodos cualitativos y cuantitativos en el proceso de diseño puede permitir a los diseñadores tomar aprovechar diferentes enfoques, según sea necesario. Por ejemplo, mientras que los grupos focales y las entrevistas proporcionan comentarios más detallados y profundos sobre el modelo mental de los usuarios [109], las mediciones remotas son muy valiosos debido a la escalabilidad de los datos recopilados, aunque proporcionan menos detalles para sacar conclusiones [112]. Por lo tanto, un enfoque exitoso podría ser comenzar con múltiples

prototipos a pequeña escala y estudios formativos que recopilan medidas cualitativas en las primeras etapas del diseño (por ejemplo, para la capa de objetivos del sistema XAI en el marco) y continúan con estudios a mayor escala y medidas cuantitativas en las etapas posteriores (por ejemplo, para modelos interpretables y evaluaciones de interfaz en el marco).

9.2 Superposición entre los objetivos de diseño

En nuestra categorización de los sistemas XAI, elegimos dos dimensiones principales para organizar los sistemas XAI según sus objetivos de diseño y medidas de evaluación en la Sección 5. Los objetivos de diseño de XAI (G1–G8) se basaron en los objetivos extraídos de los objetivos encuestados. artículos, y dado que los objetivos de diseño de XAI se derivan principalmente de sus grupos de usuarios objetivo, observamos que existen superposiciones entre los objetivos en todas las disciplinas. Por ejemplo, existe una superposición de los objetivos de G1: Transparencia algorítmica para usuarios novatos en investigación de HCI, G5: Visualización de modelos para expertos de datos en análisis visual y G7: Técnicas de interpretabilidad para expertos en IA en investigación de aprendizaje automático. Si bien se superponen, estos objetivos similares se estudian con diferentes objetivos en las tres disciplinas de investigación, lo que lleva a diversos conjuntos de requisitos de diseño y rutas de implementación. Por ejemplo, diseñar sistemas XAI para principiantes en IA requiere procesos y pasos para construir interfaces explicables centradas en el ser humano para comunicar explicaciones del modelo a los usuarios finales, mientras que diseñar nuevas técnicas de interpretabilidad para expertos en IA tiene un conjunto diferente de requisitos computacionales. Otro ejemplo de superposición en los objetivos de XAI es entre el objetivo de G6: Visualización e inspección de modelos para expertos en datos y G8: Depuración de modelos para expertos en IA, en el que se utilizan diferentes conjuntos de herramientas y requisitos para abordar diferentes objetivos de investigación.

Para abordar la superposición entre los objetivos de XAI entre las disciplinas de investigación, utilizamos los Grupos de Usuarios de XAI como una dimensión auxiliar para organizar los objetivos de XAI en este tema interdisciplinario (Sección 6) y enfatizar la diversidad de diversos objetivos de investigación. Los tres grupos de usuarios fueron elegidos para organizar los objetivos y esfuerzos de investigación en los campos de investigación de HCI (para principiantes en IA), análisis visual (para expertos en datos) y aprendizaje automático (para expertos en IA). Además, como se describe en el marco, los tres grupos de usuarios priorizan los objetivos de diseño en el proceso de diseño del sistema XAI en lugar de una separación absoluta de los objetivos de diseño. Por ejemplo, los objetivos y prioridades en el diseño del sistema XAI para la mitigación de sesgos algorítmicos para expertos en el dominio en una firma de abogados son ciertamente diferentes de los de las herramientas de capacitación y ajuste de modelos para expertos en IA. Sin embargo, al seguir el marco de diseño multidisciplinario, un equipo de diseño puede traducir los objetivos del sistema XAI en objetivos de diseño para una interfaz explicable y técnicas de aprendizaje automático para mejorar el proceso de diseño en diferentes capas. Por lo tanto, en el ejemplo anterior, el equipo de diseño puede centrarse en diversos objetivos de técnicas de interpretación y diseño de interfaz para lograr el objetivo principal de XAI de mitigar el sesgo para los expertos en el dominio. Tenga en cuenta que las características específicas de cualquier sistema en particular determinarán las prioridades de los diferentes objetivos.

9.3 Evaluación del sistema a lo largo del tiempo

Un aspecto importante en la evaluación de sistemas complejos de IA y XAI es tener en cuenta el aprendizaje del usuario. La capacidad de aprendizaje es aún más crítica cuando se miden los modelos mentales y la confianza del usuario en el sistema. Un usuario aprende y se familiariza más con el sistema con el tiempo gracias a la interacción continua con el sistema. Esto plantea la importancia de la captura repetida de datos temporales (en contraste con las mediciones estáticas) para las evaluaciones XAI. Holliday et al. [89] presentan un ejemplo de múltiples evaluaciones de confianza durante el estudio de usuarios. Midieron la confianza del usuario a intervalos regulares durante el estudio para capturar los cambios en la confianza del usuario a medida que el usuario interactúa más con el sistema. Sus resultados indican que un sistema XAI superó a uno que no era XAI a la hora de mantener la confianza del usuario a lo largo del tiem
Las mediciones basadas en el tiempo, también conocidas como mediciones dinámicas, permiten a los diseñadores monitorear la usabilidad y efectividad de las explicaciones en diversos contextos y situaciones [50, 185]. Por ejemplo,

Zhang et al. [222] exploran el efecto de las explicaciones del sistema inteligente en la calibración de la confianza del usuario.
En sus experimentos, observaron un efecto significativo en la calibración de la confianza cuando se mostró a los participantes la puntuación
de confianza de predicción del modelo. En otro ejemplo, un estudio de Nourani et al.
[159] controlaron si las experiencias tempranas de los usuarios con un sistema de reconocimiento de actividad explicable
tuvieron mejores o peores resultados del modelo, y las primeras impresiones afectaron significativamente tanto el desempeño de la tarea
como la confianza del usuario en comprender cómo funciona el sistema. En un estudio con reseña de noticias
tarea, Mohseni et al. [150] identificaron diferentes perfiles de usuario para detectar cambios en la confianza a lo largo del tiempo (dinámica de
confianza) mientras trabajaban con la ayuda de un detector de noticias falsas explicable. Su análisis de
Los resultados revelaron un efecto significativo de las explicaciones del aprendizaje automático en la dinámica de confianza del usuario.

La evaluación a largo plazo de los sistemas XAI también puede permitir a los diseñadores estimar factores valiosos de la experiencia del
usuario, como el exceso y la falta de confianza en el sistema. La precisión del sistema percibida por el usuario [110] y la transparencia [171]
son ejemplos de medidas a largo plazo para explicar la usabilidad.
que dependen de generar confianza en el usuario en la interpretabilidad del sistema. A medida que se proporciona más información
Mediante explicaciones a lo largo del tiempo, el razonamiento y las estrategias mentales pueden cambiar a medida que los usuarios crean
nuevas hipótesis sobre la funcionalidad del sistema. Por tanto, es fundamental considerar también los modelos mentales de los usuarios.
y confianza en estudios extensos para evaluar todos los aspectos del sistema XAI.

Otro caso de uso de mediciones a largo plazo es evaluar los efectos de los sistemas inteligentes.
Comportamientos no uniformes en escenarios del mundo real. Esto significa que, aunque en un entorno de estudio controlado,
un conjunto equilibrado de ejemplos de entrada presentará el sistema al usuario; en escenarios del mundo real, los usuarios
pueden enfrentar alteraciones en el desempeño del sistema en la interacción a largo plazo con el sistema. A largo plazo
Las mediciones identificarán la confianza injusta del usuario en el sistema debido a un conjunto limitado o sesgado de interacciones con el
sistema. Por ejemplo, en el contexto de los vehículos autónomos, Kraus et al. [104]
presentó un modelo de calibración de confianza y presentó estudios sobre la dinámica de la confianza en las primeras fases
de interacción del usuario con el sistema. Sus resultados indican los efectos de la automatización sin errores en
aumento constante de la confianza del usuario, así como los efectos de la información a priori del usuario en la eliminación de la
Disminución de la confianza en caso de mal funcionamiento del sistema.

9.4 Verdad fundamental de la evaluación

La investigación sobre sistemas XAI estudia varios objetivos con diferentes medidas en múltiples dominios.
La amplitud de la investigación de XAI dificulta la interpretación y transferencia de hallazgos de una sola tarea.
y dominio a otro. Conocer los factores clave para interpretar las implicaciones de los resultados de la evaluación es
esencial para agregar hallazgos en todos los dominios y disciplinas. Un factor importante para comprender los resultados de la evaluación
XAI y comparar los resultados entre múltiples estudios es la elección del terreno.
verdad. A continuación, revisamos las opciones comunes de verdad fundamental tanto para el sujeto humano como para el
métodos de evaluación computacional.

Los experimentos con sujetos humanos a menudo toman la forma de estudios controlados para examinar los efectos
de explicaciones de aprendizaje automático en un grupo de control en comparación con un grupo de referencia. En estos
configuraciones, la elección de la línea de base podría afectar las implicaciones y la importancia de los resultados. Nuestra reseña
de artículos en el espacio de evaluación XAI muestra que la mayoría de los diseños de estudio utilizan una condición de no explicación como
condición de referencia para medir la efectividad de las explicaciones del modelo en
un grupo explicativo. Los ejemplos de la línea de base incluyen enfoques que eliminan los componentes y características relacionados con
las explicaciones del modelo de la interfaz en la condición de la línea de base [103, 160]. En
otro trabajo, Poursabzi-Sangdeh et al. [168] también incluyeron una línea de base sin IA para medir la capacidad de los participantes.
rendimiento sin la ayuda de predicciones del modelo. Otra forma es comparar los efectos de
tipo de explicación o complejidad entre las condiciones del estudio sin la línea de base sin explicación.
Por ejemplo, Lage et al. [112] presentan un estudio para evaluar los efectos de la complejidad de la explicación en la comprensión y el
desempeño de los participantes. Utilizaron regresión lineal y logística para

estimar los efectos de la complejidad de la explicación en el tiempo de respuesta normalizado de los participantes, la precisión de la respuesta y la calificación subjetiva de la dificultad de la tarea.

Aunque los estudios mencionados anteriormente son experimentos controlados, es posible que todavía haya implicaciones no contabilizadas en el comportamiento humano debido a diferencias en el complejo proceso de explicación que merecen consideración. Langer et al. [114] presentan un experimento sobre explicaciones "placebicas" que muestra el comportamiento irresponsable de las personas cuando se enfrentan a explicaciones de acciones. En una configuración simple, su estudio demostró que al realizar una solicitud, la inclusión de explicaciones y justificaciones aumentaba la disposición del usuario a cumplir, incluso si las explicaciones no transmiten información significativa. Recientemente, Eiband et al. [55] propusieron utilizar explicaciones placebicas en lugar de una condición sin explicación como base para los estudios XAI en sujetos humanos. Por lo tanto, el uso de explicaciones no informativas o incluso generadas aleatoriamente como condición inicial podría potencialmente contrarrestar la tendencia positiva de un participante hacia las explicaciones y mejorar los resultados del estudio.

Considerando otros enfoques, una técnica computacional comúnmente aceptada para evaluar cuantitativamente explicaciones de instancias es crear una verdad fundamental basada en las características de entrada que contribuyen semánticamente a la clase objetivo. Por ejemplo, los mapas de segmentación de imágenes (anotaciones de objetos en imágenes) se utilizan para evaluar mapas de prominencia generados por modelos en tareas de localización de objetos débilmente supervisadas [121]. Mohseni y Ragan [149] propusieron una línea base de atención humana de múltiples capas para la evaluación a nivel de características de las explicaciones del aprendizaje automático. Su línea base de Atención Humana proporciona un mapa de atribución de características basado en humanos con un mayor nivel de granularidad en comparación con los mapas de segmentación de objetos. De manera similar, las anotaciones a nivel de características se han utilizado como explicación de la verdad fundamental en el dominio de clasificación de textos [53]. Se han utilizado otros medios menos precisos de atribución de características, como el cuadro delimitador en conjuntos de datos de imágenes, para la evaluación cuantitativa de mapas de prominencia. Por ejemplo, Du et al. [52] evaluaron mapas de prominencia generados a partir de un modelo CNN calculando IOU (intersección sobre unión) por píxeles de cuadros delimitadores de explicación del modelo y cuadros delimitadores de verdad fundamental.

9.5 Papel de las interacciones del usuario en XAI

Otra consideración importante al diseñar sistemas XAI es cómo aprovechar las interacciones del usuario para respaldar mejor la comprensión del sistema. Los beneficios del diseño de sistemas interactivos se han explorado previamente en el tema del aprendizaje automático interactivo [6, 7] para usuarios finales novatos. Los expertos en inteligencia artificial y datos también se benefician de las herramientas visuales interactivas para mejorar el rendimiento del modelo y de las tareas [57].
En esta sección, analizamos múltiples ejemplos de diseño de interacción que respaldan la comprensión del usuario del modelo de caja negra subyacente.

Centrándose en el diseño interactivo de sistemas basados en IA para principiantes en IA, Amershi et al. [6] revisó múltiples estudios de casos que demuestran la efectividad de la interactividad con un estrecho acoplamiento entre el algoritmo y el usuario. Enfatizan cómo los procesos interactivos de aprendizaje automático permiten a los usuarios examinar instantáneamente el impacto de sus acciones y adaptar sus próximas consultas para mejorar los resultados. Estas interacciones permiten a los usuarios probar varias entradas y aprender sobre el modelo mediante la creación de explicaciones hipotéticas [204]. En particular, los ciclos de prueba y error dirigidos por el usuario ayudan a los principiantes a comprender cómo funciona el modelo de aprendizaje automático y cómo dirigir el modelo para mejorar los resultados. En el contexto de XAI, Cai et al. [29] presentan un estudio en el que los usuarios dibujan imágenes para ver si un algoritmo de reconocimiento de imágenes puede reconocer correctamente el boceto deseado. Su sistema y estudio permiten realizar pruebas y errores interactivos para explorar cómo funciona el algoritmo. Además, su sistema proporciona explicaciones basadas en ejemplos en los casos en que el algoritmo no logra clasificar correctamente los dibujos. Otro enfoque es permitir a los usuarios controlar o ajustar los parámetros algorítmicos para lograr mejores resultados. Por ejemplo, Kocielnik et al. [103] presentan un estudio en el que los usuarios pudieron controlar libremente la sensibilidad de detección en un asistente de IA. Sus resultados mostraron un efecto significativo en la percepción de control y aceptación del usuario.

Las herramientas de análisis visual también respaldan la comprensión del modelo para usuarios expertos a través de la interacción. con algoritmos. Los ejemplos incluyen permitir a los científicos de datos y expertos en modelos explorar interactivamente representaciones de modelos [86], analizar procesos de entrenamiento de modelos [129] y detectar sesgos de aprendizaje [27]. Además, las técnicas de interacción integradas pueden respaldar la exploración de espacios profundos de gran tamaño. redes de aprendizaje. Por ejemplo, Hohman et al. [86] presenta múltiples funciones interactivas para seleccionar y filtrar neuronas y ampliar y desplazar representaciones de características para ayudar a los expertos en inteligencia artificial a interpretar y revisar modelos entrenados.

9.6 Generalización y Ampliación del Marco

Nuestro marco es ampliable y compatible con la interfaz e interacción existentes basadas en IA. Guía de diseño. Por ejemplo, Amershi et al. [7] proponen 18 pautas de diseño para el diseño de interacción humano-IA. Sus directrices se basan en una revisión de una gran cantidad de diseños relacionados con la IA. Fuentes de recomendación. Validaron sistemáticamente las directrices a través de múltiples rondas de evaluaciones con 49 profesionales del diseño en 20 productos con IA. Sus pautas de diseño proporcionan más detalles dentro de la capa de diseño de la interfaz de usuario de nuestro marco (Sección 8.2) para guiar el desarrollo de interacciones apropiadas del usuario con los resultados y las interacciones del modelo. En otra trabajo, Dudley y Kristensson [54] presentan una revisión y caracterización del diseño de interfaz de usuario Principios para sistemas interactivos de aprendizaje automático. Proponen un desglose estructural de los sistemas interactivos de aprendizaje automático y presentan seis principios para respaldar el diseño del sistema. Este trabajo también beneficia nuestro marco al contribuir con prácticas de diseño de aprendizaje automático interactivo para la capa de objetivos del sistema XAI (Sección 8.1) y la capa de diseño de interfaz de usuario (Sección 8.2). Desde el Desde el punto de vista de los métodos de evaluación, Mueller y Klein [153] analizan cómo las pruebas de usabilidad comunes No podemos abordar herramientas inteligentes donde el software replica la inteligencia humana. Sugieren nuevos Se necesitan soluciones que permitan a los usuarios experimentar las fortalezas y debilidades de una herramienta basada en IA. Del mismo modo, nuestro marco anidado señala el potencial de propagación de errores desde las capas internas (por ejemplo, diseño de algoritmos interpretables) a las capas externas (por ejemplo, resultados del sistema) en el XAI. polo de evaluación del sistema. El vaivén iterativo entre capas en el modelo anidado fomenta la revisión por expertos de los resultados del sistema, la evaluación centrada en el usuario de la interfaz explicable y Evaluación computacional de algoritmos de aprendizaje automático.

9.7 Limitaciones del marco

Nuestro marco proporciona una base para el diseño del sistema XAI en el trabajo en equipo interdisciplinario y tenemos Describimos nuestro ejemplo de estudio de caso para validar y mejorar el marco. El estudio de caso presentado sirve como un ejemplo práctico del uso de nuestro marco en un XAI colaborativo multidisciplinario esfuerzo de diseño y desarrollo. Nuestro caso de uso es el resultado de una investigación de un año (y en curso) realizado por un equipo de ocho investigadores universitarios con diversos orígenes. Las lecciones aprendidas y los obstáculos en nuestro estudio de caso de implementación de extremo a extremo se incorporan en el diseño presentado pautas. Sin embargo, ningún marco es perfecto o completamente integral. Reconocemos que La validez y utilidad de un marco deben demostrarse en la práctica con más estudios de casos. En nuestro trabajo futuro, planeamos ejecutar múltiples estudios de casos de validación para examinar la practicidad y utilidad de este marco.

Además, este marco tiene una limitación común a muchos marcos de diseño multidisciplinarios: ser liviano en detalles específicos en cada paso. En lugar de aportar directrices detalladas Para cada capa del marco, el marco está destinado a allanar el camino para una colaboración eficiente. entre y dentro de diferentes equipos, lo cual es esencial para el diseño del sistema XAI dada la naturaleza inherentemente interdisciplinaria de este campo. Este mayor nivel de libertad permite la extensibilidad con otros directrices de diseño (ver la discusión en la Sección 9.6) para integrar con enfoques más personalizados para dominios específicos. Además, la diversidad de objetivos de diseño y métodos de evaluación en cada capa

Puede ayudar a mantener el equilibrio de atención del equipo de diseño hacia diferentes aspectos de un XAI.
sistema.

10 CONCLUSIÓN

Revisamos la investigación relacionada con XAI para organizar múltiples objetivos de diseño y medidas de evaluación de XAI.
La Tabla 2 presenta nuestra categorización de métodos de diseño y evaluación existentes seleccionados que organiza la
literatura según tres perspectivas: objetivos de diseño, métodos de evaluación y usuarios objetivo.
del sistema XAI. Proporcionamos tablas resumidas listas para usar de métodos de evaluación y recomendaciones para
diferentes objetivos en la investigación XAI. Nuestra categorización reveló la necesidad de una
Esfuerzo interdisciplinario para diseñar y evaluar sistemas XAI. Queremos llamar la atención
Recursos relacionados en las ciencias sociales que pueden facilitar el alcance de los aspectos sociales y cognitivos.
de explicaciones. Para abordar estos problemas, propusimos un marco de diseño y evaluación que
conecta objetivos de diseño y métodos de evaluación para el diseño de sistemas XAI de extremo a extremo, como se presenta
a través de un modelo y una serie de pautas. Esperamos que nuestro marco impulse una mayor discusión sobre
la interacción entre el diseño y la evaluación de sistemas de inteligencia artificial explicables. A pesar de
El marco presentado está organizado para proporcionar una guía de alto nivel para un esfuerzo multidisciplinario para construir
sistemas XAI, no pretende ofrecer todos los aspectos del diseño y la interacción de interfaz e interacción.
Desarrollo de técnicas interpretables de aprendizaje automático. Por último, analizamos brevemente otros
Consideraciones para que los diseñadores de XAI se beneficien del conjunto de conocimientos sobre el diseño y el diseño del sistema XAI.
evaluación.

EXPRESIONES DE GRATITUD

Los autores desean agradecer a los revisores anónimos por sus útiles comentarios sobre versiones anteriores de este
manuscrito. Los puntos de vista y conclusiones de este artículo son responsabilidad de los autores y
no debe interpretarse como que representa a ninguna agencia de financiación.

REFERENCIAS

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim y Mohan Kankanhalli. 2018. Tendencias y trayectorias de
Sistemas explicables, responsables e inteligibles: una agenda de investigación de HCI. En actas de la Conferencia CHI 2018
sobre factores humanos en sistemas informáticos. ACM, 582.

[2] Amina Adadi y Mohammed Berrada. 2018. Echando un vistazo al interior de la caja negra: una encuesta sobre inteligencia artificial explicable (XAI).
Acceso IEEE 6 (2018), 52138–52160.

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt y Been Kim. 2018. Controles de cordura para
mapas de prominencia. En Avances en sistemas de procesamiento de información neuronal. 9505–9515.

[4] Yongsu Ahn y Yu-Ru Lin. 2019. FairSight: Análisis visual para la equidad en la toma de decisiones. Transacciones IEEE en
Visualización y gráficos por computadora 26, 1 (2019), 1086–1095.

[5] Eric Alexander y Michael Gleicher. 2015. Comparación de modelos temáticos basada en tareas. Transacciones IEEE sobre visualización y gráficos por
computadora 22, 1 (2015), 320–329.

[6] Saleema Amershi, Maya Cakmak, William Bradley Knox y Todd Kulesza. 2014. Poder para el pueblo: el papel de
humanos en el aprendizaje automático interactivo. Revista AI 35, 4 (2014), 105–120.

[7] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi
Iqbal, Paul N. Bennett, Kori Inkpen y otros. 2019. Directrices para la interacción entre humanos y IA. En actas del CHI 2019
Jornada sobre Factores Humanos en Sistemas Informáticos. ACM, 3.

[8] Darío Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman y Dan Mané. 2016. Problemas concretos
Problemas en la seguridad de la IA. arXiv:1606.06565. http://arxiv.org/abs/1606.06565.

[9] Mike Ananny y Kate Crawford. 2018. Ver sin saber: limitaciones del ideal de transparencia y su aplicación.
Aplicación a la responsabilidad algorítmica. Nuevos medios y sociedad 20, 3 (2018), 973–989.

[10] Stavros Antifakos, Nicky Kern, Bernt Schiele y Adrian Schwaninger. 2005. Hacia mejorar la confianza en sistemas sensibles al contexto mostrando la
confianza en el sistema. En Actas de la Séptima Conferencia Internacional sobre Interacción Humano-Computadora con Dispositivos y Servicios
Móviles. ACM, 9-14.

[11] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado,
Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Inteligencia artificial explicable

(XAI): Conceptos, taxonomías, oportunidades y desafíos hacia una IA responsable. Fusión de información 58 (2020), 82–115.

[12] Jimmy Ba, Volodymyr Mnih y Koray Kavukcuoglu. 2014. Reconocimiento de múltiples objetos con atención visual. arXiv:1412.7755.

[13] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller y Wojciech Samek. 2015. Sobre explicaciones en píxeles para decisiones de clasificadores no lineales mediante propagación de relevancia en capas. PloS One 10, 7 (2015), e0130140.

[14] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen y Klaus-Robert Müller. 2010. Cómo explicar las decisiones de clasificación individuales. Revista de investigación de aprendizaje automático 11, (junio de 2010), 1803–1831.

[15] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld y Eric Horvitz. 2019. Más allá de la precisión: el papel de los modelos mentales en el desempeño del equipo Human-AI. En Actas de la Conferencia AAAI sobre Computación Humana y Crowdsourcing, vol. 7. 2–11.

[16] Victoria Bellotti y Keith Edwards. 2001. Inteligibilidad y responsabilidad: consideraciones humanas en contexto sistemas. Interacción persona-computadora 16, 2-4 (2001), 193–212.

[17] Shlomo Berkovsky, Ronnie Taib y Dan Conway. 2017. ¿Cómo recomendar?: Factores de confianza del usuario en los sistemas de recomendación de películas. En actas de la 22ª Conferencia Internacional sobre Interfaces de Usuario Inteligentes (IUI'17). ACM, Nueva York, Nueva York, 287–300. https://doi.org/10.1145/3025171.3025209

[18] Daniel M. Best, Alex Endert y Daniel Kidwell. 2014. 7 desafíos clave para la visualización en la defensa de redes cibernéticas. En actas del undécimo taller sobre visualización para seguridad cibernética. ACC, 33–40.

[19] Mustafa Bilgic y Raymond J. Mooney. 2005. Explicación de recomendaciones: satisfacción versus promoción. En el más allá Taller de personalización, IUI, vol. 5. 153.

[20] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao y Nigel Shadbolt. 2018. “Es reducir al ser humano a un porcentaje”: Percepciones de justicia en las decisiones algorítmicas. En actas de la Conferencia CHI de 2018 sobre factores humanos en sistemas informáticos. ACM, 377.

[21] Philip Bobko, Alex J. Barelka y Leanne M. Hirshfield. 2014. El constructo de la sospecha a nivel estatal: un modelo y una agenda de investigación para contextos automatizados y de tecnología de la información (TI). Factores humanos 56, 3 (2014), 489–508.

[22] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry J. Ackel, Urs Muller, Phil Yeres y Karol Zieba. 2018. Visualbackprop: visualización eficiente de CNN para conducción autónoma. En 2018 Conferencia Internacional IEEE sobre Robótica y Automatización (ICRA'18). IEEE, 1–8.

[23] Engin Bozdag y Jeroen van den Hoven. 2015. Rompiendo la burbuja del filtro: Democracia y diseño. Ética e Información Tecnología de la información 17, 4 (2015), 249–265.

[24] Nicholas Bryan y Gautham Mysore. 2013. Un modelo de variable latente regularizado posterior eficiente para interacción Separación de fuentes de sonido. En Conferencia Internacional sobre Aprendizaje Automático. 208–216.

[25] Andrea Bunt, Matthew Lount y Catherine Lauzon. 2012. ¿Son siempre importantes las explicaciones?: Un estudio de sistemas interactivos inteligentes de bajo costo implementados. En actas de la Conferencia internacional ACM de 2012 sobre interfaces de usuario inteligentes. ACM, 169-178.

[26] Adrian Bussone, Simone Stumpf y Dympna O'Sullivan. 2015. El papel de las explicaciones sobre la confianza en los sistemas de apoyo a las decisiones clínicas. En Conferencia Internacional sobre Informática Sanitaria (ICHI'15). IEEE, 160–169.

[27] Ángel Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern y Duen Horng Chau. 2019. FairVis: análisis visual para descubrir sesgos interseccionales en el aprendizaje automático. Conferencia IEEE sobre ciencia y tecnología de análisis visual (VAST'19).

[28] Béatrice Cahour y Jean-François Forzy. 2009. ¿La proyección sobre el uso mejora la confianza y la exploración? Un ejemplo con un sistema de control de crucero. Ciencia de la seguridad 47, 9 (2009), 1260–1270.

[29] Carrie J. Cai, Jonas Jongejan y Jess Holbrook. 2019. Los efectos de las explicaciones basadas en ejemplos en una interfaz de aprendizaje automático. En actas de la 24ª Conferencia Internacional sobre Interfaces de Usuario Inteligentes. 258–262.

[30] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C. Stumpe, et al. 2019. Herramientas centradas en el ser humano para hacer frente a algoritmos imperfectos durante la toma de decisiones médicas. En actas de la Conferencia CHI de 2019 sobre factores humanos en sistemas informáticos. 1–14.

[31] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm y Noemie Elhadad. 2015. Modelos inteligibles para la atención sanitaria: predicción del riesgo de neumonía y reingreso hospitalario a los 30 días. En actas de la 21ª Conferencia internacional ACM SIGKDD sobre descubrimiento de conocimientos y minería de datos. ACM, 1721-1730.

[32] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha y Madeleine Udell. 2019. Equidad bajo el desconocimiento: evaluación de la disparidad cuando no se observa la clase protegida. En Actas de la Conferencia sobre Equidad, Responsabilidad y Transparencia. ACM, 339–348.

[33] Jaegul Choo, Hanseung Lee, Jaeyeon Kihm y Haesun Park. 2010. iVisClassifier: un sistema de análisis visual interactivo para clasificación basado en reducción de dimensiones supervisada. En 2010, Simposio IEEE sobre ciencia y tecnología de análisis visual (VAST'10). IEEE, 27–34.

[34] Jaegul Choo y Shixia Liu. 2018. Análisis visual para un aprendizaje profundo explicable. Aplicaciones y gráficos por computadora IEEE 38, 4 (2018), 84–92.

[35] Alexandra Chouldechova. 2017. Predicción justa con impacto dispar: un estudio del sesgo en la predicción de la reincidencia en instrumentos. Grandes datos 5, 2 (2017), 153–163.

[36] Michael Chromik, Malin Eiband, Sarah Theres Völkel y Daniel Buschek. 2019. Patrones oscuros de explicabilidad, transparencia y control del usuario para sistemas inteligentes. En Talleres IUI.

[37] Lingyang Chu, Xia Hu, Juhua Hu, Lanjun Wang y Jian Pei. 2018. Interpretación exacta y consistente para redes neuronales lineales por partes: una solución de forma cerrada. En actas de la 24ª Conferencia internacional ACM SIGKDD sobre descubrimiento de conocimientos y minería de datos. 1244-1253.

[38] Miruna-Adriana Clinciu y Helen Hastie. 2019. Un estudio sobre terminología de IA explicable. En actas del primer taller sobre tecnología de lenguaje natural interactivo para inteligencia artificial explicable (NL4XAI'19). 8–13.

[39] Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna Van der Lek-Ciudin, Tom Vanallemeersch y Vincent Vandeghinste. 2018. Intellingo: un entorno de traducción inteligible. En actas de la Conferencia CHI de 2018 sobre factores humanos en sistemas informáticos. ACM, 524.

[40] Enrico Costanza, Joel E. Fischer, James A. Colley, Tom Rodden, Sarvapali D. Ramchurn y Nicholas R. Jennings. 2014. Lavar la ropa con agentes: una prueba de campo de un futuro sistema de energía inteligente en el hogar. En actas de la Conferencia SIGCHI sobre factores humanos en sistemas informáticos. ACM, 813–822.

[41] William Curran, Travis Moore, Todd Kulesza, Weng-Keen Wong, Sinisa Todorovic, Simone Stumpf, Rachel White y Margaret Burnett. 2012. Hacia el reconocimiento de lo interesante: ¿Pueden los usuarios finales ayudar a la visión por computadora a reconocer los atributos subjetivos de los objetos en las imágenes? En actas de la Conferencia internacional ACM de 2012 sobre interfaces de usuario inteligentes. ACM, 285–288.

[42] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh y Dhruv Batra. 2016. Atención humana en la respuesta visual a preguntas: ¿los humanos y las redes profundas miran las mismas regiones? En Conferencia sobre Métodos Empíricos en Procesamiento del Lenguaje Natural (EMNLP'16). https://computing.ece.vt.edu/~abhshkdz/vqa-hat/

[43] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh y Dhruv Batra. 2017. Atención humana en la respuesta visual a preguntas: ¿los humanos y las redes profundas miran las mismas regiones? Visión por computadora y comprensión de imágenes 163 (2017), 90–100.

[44] Amit Datta, Michael Carl Tschantz y Anupam Datta. 2015. Experimentos automatizados sobre la configuración de privacidad de los anuncios. Pro-Actas sobre tecnologías de mejora de la privacidad 2015, 1 (2015), 92–112.

[45] Nicolás Diakopoulos. 2014. Responsabilidad algorítmica: la investigación de cajas negras. Centro de remolque para digital Periodismo (2014).

[46] Nicolás Diakopoulos. 2017. Habilitación de la responsabilidad de los medios algorítmicos: la transparencia como elemento constructivo y crítico. lente física. En Minería de datos transparente para datos grandes y pequeños. Saltador, 25–43.

[47] Jonathan Dodge, Sean Penney, Andrew Anderson y Margaret M. Burnett. 2018. ¿Qué debería estar en una explicación XAI? ¿nación? Lo que revela el IFT. En Talleres IUI.

[48] Final Doshi-Velez y Been Kim. 2017. Hacia una ciencia rigurosa del aprendizaje automático interpretable. arXiv:1702.08608. http://arxiv.org/abs/1702.08608.

[49] Final Doshi-Velez, Mason Kortz, Ryan Budish, Christopher Bavitz, Samuel J. Gershman, David O'Brien, Stuart Shieber, Jim Waldo, David Weinberger y Alexandra Wood. 2017. Responsabilidad de la IA según la ley: el papel de la explicación. Publicación de investigación del Centro Berkman de próxima publicación (2017), 18–07.

[50] James K. Doyle, Michael J. Radzicki y W. Scott Trees. 2008. Medición del cambio en modelos mentales de complejos. sistemas dinámicos. En la toma de decisiones complejas. Saltador, 269–294.

[51] Fan Du, Catherine Plaisant, Neil Spring, Kenyon Crowley y Ben Shneiderman. 2019. EventAction: un enfoque de análisis visual para recomendaciones explicables para secuencias de eventos. Transacciones ACM sobre sistemas inteligentes interactivos (TiiS) 9, 4 (2019), 1–31.

[52] Mengnan Du, Ninghao Liu, Qingquan Song y Xia Hu. 2018. Hacia la explicación de la predicción basada en DNN con inversión de características guiada. En actas de la 24ª Conferencia internacional ACM SIGKDD sobre descubrimiento de conocimientos y minería de datos. 1358-1367.

[53] M. Du, N. Liu, F. Yang y X. Hu. 2019. Aprendizaje de redes neuronales profundas creíbles con regularización racional. En Conferencia Internacional IEEE 2019 sobre Minería de Datos (ICDM'19). 150–159.

[54] John J. Dudley y Per Ola Kristensson. 2018. Una revisión del diseño de la interfaz de usuario para el aprendizaje automático interactivo. Transacciones ACM sobre sistemas inteligentes interactivos (TiiS) 8, 2 (2018), 8.

[55] Malin Eiband, Daniel Buschek, Alexander Kremer y Heinrich Hussmann. 2019. El impacto de las explicaciones placebicas en la confianza en los sistemas inteligentes. En resúmenes ampliados de la Conferencia CHI de 2019 sobre factores humanos en sistemas informáticos. ACM, LBW0243.

[56] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug y Heinrich Hussmann. 2018. Llevar el diseño de transparencia a la práctica. En la 23ª Conferencia Internacional sobre Interfaces de Usuario Inteligentes (IUI'18). ACM, Nueva York, Nueva York, 211–223. https://doi.org/10.1145/3172944.3172961

[57] A. Endert, W. Ribarsky, C. Turkay, BL Wong, Ian Nabney, I. Díaz Blanco y F. Rossi. 2017. El estado del arte en la integración del aprendizaje automático en el análisis visual. En Foro de gráficos por computadora, vol. 36. Biblioteca en línea Wiley, 458–486.

[58] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton y Christian Sandvig. 2015. Siempre supuse que no era tan cercano a [ella]: Razonamiento sobre algoritmos invisibles en las fuentes de noticias. En actas de la 33ª Conferencia Anual de ACM sobre factores humanos en sistemas informáticos. ACM, 153-162.

[59] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios y Kevin Hamilton. 2017. "Ten cuidado; las cosas pueden ser peores de lo que parecen": comprender los algoritmos sesgados y el comportamiento de los usuarios en torno a ellos en las plataformas de calificación. En la 11ª Conferencia Internacional AAAI sobre Web y Redes Sociales.

[60] Raquel Florez López y Juan Manuel Ramón Jerónimo. 2015. Mejora de la precisión y la interpretabilidad de las estrategias conjuntas en la evaluación del riesgo crediticio. Una propuesta de bosque de decisión correlacionada y ajustada. Sistemas expertos con aplicaciones 42, 13 (2015), 5737–5753.

[61] Ruth C. Fong y Andrea Vedaldi. 2017. Explicaciones interpretables de cajas negras mediante perturbación significativa. En Actas de la Conferencia Internacional IEEE sobre Visión por Computadora. 3429–3437.

[62] Fatih Gedikli, Dietmar Jannach y Mouzhi Ge. 2014. ¿Cómo debo explicar? Una comparación de diferentes tipos de explicaciones para sistemas de recomendación. Revista internacional de estudios humanos-computadores 72, 4 (2014), 367–382.

[63] Amirata Ghorbani, James Wexler, James Y. Zou y Been Kim. 2019. Hacia explicaciones automáticas basadas en conceptos. En Avances en sistemas de procesamiento de información neuronal. 9273–9282.

[64] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter y Lalana Kagal. 2018. Explicaciones explicativas: una descripción general de la interpretabilidad del aprendizaje automático. En 2018, quinta conferencia internacional del IEEE sobre ciencia de datos y análisis avanzado (DSAA'18). IEEE, 80–89.

[65] Alyssa Glass, Deborah L. McGuinness y Michael Wolverton. 2008. Hacia el establecimiento de confianza en agentes adaptativos. En actas de la 13ª Conferencia Internacional sobre Interfaces de Usuario Inteligentes. ACM, 227–236.

[66] John Goodall, Eric D. Ragan, Chad A. Steed, Joel W. Reed, G. David Richardson, Kelly MT Huffer, Robert A. Bridges y Jason A. Laska. 2018. Situ: Identificar y explicar comportamientos sospechosos en redes. Transacciones IEEE sobre visualización y gráficos por computadora 25, 1 (2018), 204–214.

[67] Bryce Goodman y Seth Flaxman. 2017. Regulaciones de la Unión Europea sobre la toma de decisiones algorítmicas y el "derecho a la explicación". Revista AI 38, 3 (2017), 50–57.

[68] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt y Austin L. Toombs. 2018. El lado oscuro (patrones) del diseño UX. En actas de la Conferencia CHI de 2018 sobre factores humanos en sistemas informáticos. ACM, 534.

[69] Shirley Gregor e Izak Benbasat. 1999. Explicaciones de los sistemas inteligentes: fundamentos teóricos e implicaciones. cationes para la práctica. Sistemas de información de gestión trimestral 23, 4 (1999), 2.

[70] Alex Groce, Todd Kulesza, Chaoqiang Zhang, Shalini Shamasunder, Margaret Burnett, Weng-Keen Wong, Simone Stumpf, Shubhomoy Das, Amber Shinsel, Forrest Bice, et al. 2014. Eres el único oráculo posible: selección de pruebas efectiva para usuarios finales de sistemas interactivos de aprendizaje automático. Transacciones IEEE sobre ingeniería de software 40, 3 (2014), 307–323.

[71] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti y Dino Pedreschi. 2018. Un estudio de métodos para explicar los modelos de caja negra. Encuestas de Computación ACM (CSUR) 51, 5 (2018), 93.

[72] David Gunning. 2017. Inteligencia artificial explicable (XAI). Agencia de Proyectos de Investigación Avanzada de Defensa (DARPA) (2017).

[73] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove y Christo Wilson. 2013. Medición de la personalización de la búsqueda web. En actas de la 22ª Conferencia Internacional sobre la World Wide Web. ACM, 527–538.

[74] Steven R. Haynes, Mark A. Cohen y Frank E. Ritter. 2009. Diseños para explicar agentes inteligentes. Revista internacional de estudios humanos-computadores 67, 1 (2009), 90–110.

[75] Jeffrey Heer. 2019. Agencia más automatización: diseño de inteligencia artificial en sistemas interactivos. Procedimientos de la Academia Nacional de Ciencias USA 116, 6 (2019), 1844–1850.

[76] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell y Anna Rohrbach. 2018. Las mujeres también practican snowboard: superar los prejuicios en los subtítulos de modelos. En Conferencia Europea sobre Visión por Computador. Saltador, 793–811.

[77] Jonathan L. Herlocker, Joseph A. Konstan y John Riedl. 2000. Explicación de las recomendaciones de filtrado colaborativo. En actas de la Conferencia ACM de 2000 sobre trabajo cooperativo asistido por computadora. ACM, 241–250.

[78] Bernease Herman. 2017. La promesa y el peligro de la evaluación humana para la interpretabilidad del modelo. arXiv:1711.07414. https://arxiv.org/abs/1711.07414.

[79] Robert Hoffman, Tim Miller, Shane T. Mueller, Gary Klein y William J. Clancey. 2018. Explicando explicación, parte 4: Una inmersión profunda en redes profundas. Sistemas inteligentes IEEE 33, 3 (2018), 87–95.

[80] Robert R. Hoffman. 2017. Teoría conceptos medidas pero políticas métricas. En Métricas y Escenarios de Macrocognición. Prensa CRC, 35–42.

[81] Robert R. Hoffman, John K. Hawley y Jeffrey M. Bradshaw. 2014. Mitos de la automatización, parte 2: Algunos muy humanos consecuencias. Sistemas inteligentes IEEE 29, 2 (2014), 82–85.

[82] Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw y Al Underbrink. 2013. Confianza en la automatización. IEEE Sistemas inteligentes 28, 1 (2013), 84–88.

[83] Robert R. Hoffman y Gary Klein. 2017. Explicación explicativa, parte 1: Fundamentos teóricos. IEEE inteligente Sistemas 32, 3 (2017), 68–73.

[84] Robert R. Hoffman, Shane T. Mueller y Gary Klein. 2017. Explicación explicativa, parte 2: Fundamentos empíricos. Sistemas inteligentes IEEE 32, 4 (2017), 78–86.

[85] Robert R. Hoffman, Shane T. Mueller, Gary Klein y Jordan Litman. 2018. Métricas para una IA explicable: desafíos y perspectivas. arXiv:1812.04608. https://arxiv.org/abs/1812.04608.

[86] Fred Hohman, Haekyu Park, Caleb Robinson y Duen Horng Polo Chau. 2019. Cumbre: Ampliación de la interpretabilidad del aprendizaje profundo mediante la visualización de resúmenes de activación y atribución. Transacciones IEEE sobre visualización y gráficos por computadora 26, 1 (2019), 1096–1106.

[87] Fred Hohman, Arjun Srinivasan y Steven M. Drucker. 2019. TeleGam: combinación de visualización y verbalización para un aprendizaje automático interpretable. Conferencia de visualización IEEE (VIS'19).

[88] Fred Matthew Hohman, Minsuk Kahng, Robert Pienta y Duen Horng Chau. 2018. Análisis visual en el aprendizaje profundo: una encuesta interrogativa para las próximas fronteras. Transacciones IEEE sobre visualización y gráficos por computadora 25, 8 (agosto de 2018), 2674–2693.

[89] Daniel Holliday, Stephanie Wilson y Simone Stumpf. 2016. Confianza del usuario en sistemas inteligentes: un viaje en el tiempo. En actas de la XXI Conferencia Internacional sobre Interfaces de Usuario Inteligentes. ACM, 164-168.

[90] Kristina Höök. 2000. Pasos a seguir antes de que las interfaces de usuario inteligentes se vuelvan reales. Interactuando con Computadoras 12, 4 (2000), 409–426.

[91] Philip N. Howard y Bence Kollanyi. 2016. Bots, #StrongerIn y #Brexit: propaganda computacional durante el Referéndum entre el Reino Unido y la UE.

[92] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff y Alison Smith. 2014. Modelado de temas interactivos. Máquina Aprendizaje 95, 3 (2014), 423–469.

[93] Shagun Jhaver, Yoni Karpfen y Judd Antin. 2018. Ansiedad algorítmica y estrategias de afrontamiento de los anfitriones de Airbnb. En Actas de la Conferencia CHI de 2018 sobre factores humanos en sistemas informáticos. ACM, 421.

[94] Jiun-Yin Jian, Ann M. Bisantz y Colin G. Drury. 2000. Fundamentos para una escala de confianza determinada empíricamente en sistemas automatizados. Revista internacional de ergonomía cognitiva 4, 1 (2000), 53–71.

[95] Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro y Duen Horng Polo Chau. 2018. ActiVis: exploración visual de modelos de redes neuronales profundas a escala industrial. Transacciones IEEE sobre visualización y gráficos por computadora 24, 1 (2018), 88–97.

[96] Matthew Kay, Tara Kola, Jessica R. Hullman y Sean A. Munson. 2016. ¿Cuándo (más o menos) es mi autobús?: visualizaciones de incertidumbre centradas en el usuario en sistemas predictivos móviles cotidianos. En actas de la Conferencia CHI de 2016 sobre factores humanos en sistemas informáticos. ACM, 5092–5103.

[97] Frank C. Keil. 2006. Explicación y comprensión. Revisión anual de psicología 57 (2006), 227–254.

[98] Sido Kim, Rajiv Khanna y Oluwasanmi O. Koyejo. 2016. Los ejemplos no bastan, ¡aprende a criticar! crítica por interpretabilidad. En Avances en sistemas de procesamiento de información neuronal. 2280–2288.

[99] Sido Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretabilidad más allá de la atribución de características: pruebas cuantitativas con vectores de activación de conceptos (TCAV). En Conferencia Internacional sobre Aprendizaje Automático. 2673–2682.

[100] Jaedeok Kim y Jingoo Seo. 2017. Extracción de explicaciones comprensibles para los humanos para modelos de clasificación de caja negra. basado en la factorización matricial. arXiv:1709.06201. https://arxiv.org/abs/1709.06201.

[101] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan y Been Kim. 2019. La (in)confiabilidad de los métodos de prominencia. En IA explicable: interpretación, explicación y visualización del aprendizaje profundo. Saltador, 267–280.

[102] Gary Klein. 2018. Explicación explicativa, parte 3: El panorama causal. Sistemas inteligentes IEEE 33, 2 (2018), 83–88.

[103] Rafal Kocielnik, Saleema Amershi y Paul N. Bennett. 2019. ¿Aceptarás una IA imperfecta? Explorar diseños para ajustar las expectativas del usuario final de los sistemas de IA. En actas de la Conferencia CHI de 2019 sobre factores humanos en sistemas informáticos. 1–14.

[104] Johannes Kraus, David Scholz, Dina Stiegemeier y Martin Baumann. 2019. Cuanto más sepa: confíe en la dinámica y la calibración en la conducción altamente automatizada y los efectos de las tomas de control, el mal funcionamiento del sistema y la transparencia del sistema. Factores humanos (2019), 0018720819853686.

[105] Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs y Enrico Bertini. 2017. Un flujo de trabajo para el diagnóstico visual de clasificadores binarios utilizando explicaciones a nivel de instancia. En 2017, Conferencia IEEE sobre ciencia y tecnología de análisis visual (VAST'17). IEEE, 162–172.

[106] Josua Krause, Adam Perer y Enrico Bertini. 2014. INFUSE: Selección de funciones interactivas para el modelado predictivo de datos de alta dimensión. Transacciones IEEE sobre visualización y gráficos por computadora 20, 12 (2014), 1614–1623.

[107] Josua Krause, Adam Perer y Kenney Ng. 2016. Interactuar con predicciones: inspección visual de modelos de aprendizaje automático de caja negra. En actas de la Conferencia CHI de 2016 sobre factores humanos en sistemas informáticos. ACM, 5686–5697.

[108] Todd Kulesza, Margaret Burnett, Weng-Keen Wong y Simone Stumpf. 2015. Principios de depuración explicativa para personalizar el aprendizaje automático interactivo. En actas de la vigésima conferencia internacional sobre interfaces de usuario inteligentes. ACM, 126-137.

[109] Todd Kulesza, Simone Stumpf, Margaret Burnett e Irwin Kwan. 2012. ¿Dime más?: Los efectos de la solidez del modelo mental en la personalización de un agente inteligente. En actas de la Conferencia SIGCHI sobre factores humanos en sistemas informáticos (CHI'12). ACM, Nueva York, Nueva York, 1–10.

[110] Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel y Kevin McIntosh. 2010. Depuración explicativa: soporte a la depuración por parte del usuario final de programas aprendidos por máquina. En 2010, Simposio IEEE sobre lenguajes visuales y computación centrada en el ser humano (VL/HCC'10). IEEE, 41–48.

[111] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan y Weng-Keen Wong. 2013. ¿Demasiado, muy poco o justo? Las formas en que las explicaciones impactan los modelos mentales de los usuarios finales. En 2013, Simposio IEEE sobre lenguajes visuales y computación centrada en las personas (VL/HCC'13). IEEE, 3–10.

[112] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman y Finale Doshi-Velez. 2019. Evaluación humana de modelos construidos para la interpretabilidad. En Actas de la Conferencia AAAI sobre Computación Humana y Crowdsourcing, vol. 7. 59–67.

[113] Himabindu Lakkaraju, Stephen H. Bach y Jure Leskovec. 2016. Conjuntos de decisiones interpretables: un marco conjunto para la descripción y la predicción. En actas de la 22ª Conferencia internacional ACM SIGKDD sobre descubrimiento de conocimientos y minería de datos. ACM, 1675–1684.

[114] Ellen J. Langer, Arthur Blank y Benzion Chanowitz. 1978. La insensatez de la acción ostensiblemente reflexiva: el papel de la información "placebica" en la interacción interpersonal. Revista de Personalidad y Psicología Social 36, 6 (1978), 635.

[115] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha y Daniel Kusbit. 2019. Justicia procesal en equidad algorítmica: aprovechar la transparencia y el control de resultados para una mediación algorítmica justa. Actas de la ACM sobre interacción persona-computadora 3, CSCW (2019), 182.

[116] Min Kyung Lee, Daniel Kusbit, Evan Metsky y Laura Dabbish. 2015. Trabajar con máquinas: el impacto de la gestión algorítmica y basada en datos en los trabajadores humanos. En actas de la 33ª Conferencia Anual de ACM sobre factores humanos en sistemas informáticos. MCA, 1603–1612.

[117] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland y Patrick Vinck. 2017. Justo, transparente y ac- procesos algorítmicos contables de toma de decisiones. Filosofía y tecnología (2017), 1–17.

[118] Piyawat Lertvittayakumjorn y Francesca Toni. 2019. Evaluaciones basadas en humanos de métodos de explicación para la clasificación de textos. En actas de la Conferencia de 2019 sobre métodos empíricos en el procesamiento del lenguaje natural y la novena Conferencia Internacional Conjunta sobre Procesamiento del Lenguaje Natural (EMNLP-IJCNLP'19). 5198–5208.

[119] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, David Madigan y otros. 2015. Clasificadores interpretables que utilizan reglas y análisis bayesiano: construcción de un mejor modelo de predicción de accidentes cerebrovasculares. Los anales de la estadística aplicada 9, 3 (2015), 1350-1371.

[120] Alexander Lex, Marc Streit, H.-J. Schulz, Christian Partl, Dieter Schmalstieg, Peter J. Park y Nils Gehlenborg. 2012. StratomeX: Análisis visual de datos genómicos heterogéneos a gran escala para la caracterización de subtipos de cáncer. En Foro de gráficos por computadora, vol. 31. Biblioteca en línea Wiley, 1175–1184.

[121] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst y Yun Fu. 2018. Dime dónde buscar: Red de inferencia de atención guiada. En actas de la Conferencia IEEE sobre visión por computadora y reconocimiento de patrones. 9215–9223.

[122] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li y Yun Fu. 2019. Red puente de atención para la transferencia de conocimiento. En actas de la Conferencia Internacional IEEE sobre Visión por Computadora. 5198–5207.

[123] Brian Lim. 2011. Mejorar la comprensión, la confianza y el control con inteligibilidad en aplicaciones sensibles al contexto. Universidad de Carnegie mellon.

[124] Brian Y. Lim y Anind K. Dey. 2009. Evaluación de la demanda de inteligibilidad en aplicaciones sensibles al contexto. En curso- ings de la 11ª Conferencia Internacional sobre Computación Ubicua. ACM, 195–204.

[125] Brian Y. Lim, Anind K. Dey y Daniel Avrahami. 2009. Las explicaciones de por qué y por qué no mejoran la inteligibilidad de los sistemas inteligentes conscientes del contexto. En actas de la Conferencia SIGCHI sobre factores humanos en sistemas informáticos. JCA, 2119–2128.

[126] Brian Y. Lim, Qian Yang, Ashraf M. Abdul y Danding Wang. 2019. ¿Por qué estas explicaciones? Selección de tipos de inteligibilidad para objetivos de explicación. En Talleres IUI.

[127] Zachary C. Lipton. 2016. El mito de la interpretabilidad del modelo. arXiv:1606.03490. https://arxiv.org/abs/1606.03490.

[128] Mengchen Liu, Shixia Liu, Xizhou Zhu, Qinying Liao, Furu Wei y Shimei Pan. 2016. Un enfoque consciente de la incertidumbre para la recuperación exploratoria de microblogs. Transacciones IEEE sobre visualización y gráficos por computadora 22, 1 (2016), 250–259.

[129] Mengchen Liu, Jiaxin Shi, Kelei Cao, Jun Zhu y Shixia Liu. 2018. Analizando los procesos formativos de generación profunda. modelos activos. Transacciones IEEE sobre visualización y gráficos por computadora 24, 1 (2018), 77–87.

[130] Mengchen Liu, Jiaxin Shi, Zhen Li, Chongxuan Li, Jun Zhu y Shixia Liu. 2017. Hacia un mejor análisis de redes neuronales convolucionales profundas. Transacciones IEEE sobre visualización y gráficos por computadora 23, 1 (2017), 91–100.

[131] Shixia Liu, Xiting Wang, Jianfei Chen, Jim Zhu y Baining Guo. 2014. TopicPanorama: una imagen completa de temas relevantes. En 2014, Conferencia IEEE sobre ciencia y tecnología de análisis visual (VAST'14). IEEE, 183–192.

[132]Tania Lombrozo. 2006. La estructura y función de las explicaciones. Tendencias en ciencias cognitivas 10, 10 (2006), 464–470.

[133]Tania Lombrozo. 2009. Explicación y categorización: ¿Cómo “¿por qué?” informa "¿qué?". Cognición 110, 2 (2009), 248–253.

[134] Scott M. Lundberg y Su-In Lee. 2017. Un enfoque unificado para interpretar las predicciones de los modelos. En avances en neurología Sistemas de procesamiento de información. 4765–4774.

[135] Laurens van der Maaten y Geoffrey Hinton. 2008. Visualización de datos utilizando t-SNE. Revista de aprendizaje automático Investigación 9, (noviembre de 2008), 2579–2605.

[136] María Madsen y Shirley Gregor. 2000. Medición de la confianza entre humanos y computadoras. En la 11ª Conferencia de Australasia sobre Infor- Sistemas de información, vol. 53. Citador, 6–8.

[137] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman y Aram Galstyan. 2019. Una encuesta sobre el sesgo y la equidad en el aprendizaje automático. arXiv:1908.09635. https://arxiv.org/abs/1908.09635.

[138] Sarah Mennicken, Jo Vermeulen y Elaine M. Huang. 2014. De las casas aumentadas de hoy a las casas inteligentes del mañana: nuevas direcciones para la investigación en automatización del hogar. En actas de la Conferencia conjunta internacional ACM de 2014 sobre informática omnipresente y ubicua. ACM, 105-115.

[139] Stephanie M. Merritt, Heather Heimbaugh, Jennifer LaChapell y Deborah Lee. 2013. Confío en ello, pero no sé por qué: Efectos de las actitudes implícitas hacia la automatización en la confianza en un sistema automatizado. Factores humanos 55, 3 (2013), 520–534.

[140] Miriah Meyer, Michael Sedlmair, P. Samuel Quinan y Tamara Munzner. 2015. El modelo de pautas y bloques anidados. Visualización de información 14, 3 (2015), 234–249.

[141] Debra Meyerson, Karl E. Weick y Roderick M. Kramer. 1996. Confianza rápida y grupos temporales. Confianza en la organización zaciones: Fronteras de la teoría y la investigación 166 (1996), 195.

[142] Tim Miller. 2019. Explicación de la inteligencia artificial: conocimientos de las ciencias sociales. Inteligencia artificial 267 (2019), 1–38.

[143] Yao Ming, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song y Huamin Qu. 2017. Comprensión de los recuerdos ocultos de las redes neuronales recurrentes. En 2017, Conferencia IEEE sobre ciencia y tecnología de análisis visual (VAST'17). IEEE, 13-24.

[144] Yao Ming, Huamin Qu y Enrico Bertini. 2018. Rulematrix: visualización y comprensión de clasificadores con reglas. Transacciones IEEE sobre visualización y gráficos por computadora 25, 1 (2018), 342–352.

[145] Brent Mittelstadt. 2016. Automatización, algoritmos y política: auditoría de transparencia en la personalización de contenidos sistemas. Revista Internacional de Comunicación 10 (2016), 12.

[146] Sina Mohseni, Akshay Jagadeesh y Zhangyang Wang. 2019. Predicción de fallos del modelo mediante mapas de prominencia en sistemas de conducción autónomos. Taller ICML sobre incertidumbre y robustez en el aprendizaje profundo.

[147] Sina Mohseni, Mandar Pitale, Vasu Singh y Zhangyang Wang. 2020. Soluciones prácticas para la seguridad del aprendizaje automático en vehículos autónomos. En el Taller de la AAAI sobre seguridad de la inteligencia artificial (Safe AI'20).

[148] Sina Mohseni, Eric Ragan y Xia Hu. 2019. Cuestiones abiertas en la lucha contra las noticias falsas: la interpretabilidad como oportunidad. arXiv:1904.03016. https://arxiv.org/abs/1904.03016.

[149] Sina Mohseni y Eric D. Ragan. 2018. Un punto de referencia de evaluación basado en humanos para explicaciones locales de máquinas. aprendiendo. arXiv:1801.05075. https://arxiv.org/abs/1801.05075.

[150] Sina Mohseni, Fan Yang, Shiva Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji y Eric Ragan. 2020. Confíe en la evolución a lo largo del tiempo en una IA explicable para la detección de noticias falsas. Taller de IA justa y responsable en CHI 2020.

[151] Christoph Molnar. 2019. Aprendizaje automático interpretable. Lulu.com.

[152] Grégoire Montavon, Wojciech Samek y Klaus-Robert Müller. 2017. Métodos para interpretar y comprender. Redes neuronales profundas. Procesamiento de señales digitales 73 (febrero de 2018), 1–15.

[153] Shane T. Mueller y Gary Klein. 2011. Mejora de los modelos mentales de los usuarios de herramientas de software inteligentes. IEEE inteligente Sistemas 26, 2 (2011), 77–83.

[154] Bonnie M. Muir. 1987. Confianza entre humanos y máquinas y diseño de ayudas para la toma de decisiones. Revista internacional de estudios hombre-máquina 27, 5-6 (1987), 527–539.

[155] Tamara Münzner. 2009. Un modelo de proceso anidado para el diseño y validación de visualización. Transacciones IEEE sobre visualización y gráficos por computadora 6 (2009), 921–928.

[156] Brad A. Myers, David A. Weitzman, Andrew J. Ko y Duen H. Chau. 2006. Respondiendo a las preguntas de por qué y por qué no en las interfaces de usuario. En actas de la Conferencia SIGCHI sobre factores humanos en sistemas informáticos. ACM, 397–406.

[157] Andrew P. Norton y Yanjun Qi. 2017. Adversarial-playground: un conjunto de visualización que muestra cómo los ejemplos contradictorios engañan al aprendizaje profundo. En 2017 Simposio IEEE sobre visualización para seguridad cibernética (VizSec'17). IEEE, 1–4.

[158] Florian Nothdurft, Felix Richter y Wolfgang Minker. 2014. Manejo probabilístico de la confianza entre humanos y computadoras. En Actas de la XV Reunión Anual del Grupo de Interés Especial sobre Discurso y Diálogo (SIGDIAL'14). 51–59.

[159] Mahsan Nourani, Dondald Honeycutt, Jeremy Block, Chiradeep Roy, Tahrima Rahman, Eric D. Ragan y Vibhav Gogate. 2020. Investigando la importancia de las primeras impresiones y la IA explicable con análisis de vídeo interactivo. En resúmenes ampliados de la Conferencia CHI de 2019 sobre factores humanos en sistemas informáticos. ACM.

[160] Mahsan Nourani, Samia Kabir, Sina Mohseni y Eric D. Ragan. 2019. Los efectos de las explicaciones significativas y sin sentido sobre la confianza y la precisión percibida del sistema en sistemas inteligentes. En Actas de la Conferencia AAAI sobre Computación Humana y Crowdsourcing, vol. 7. 97–105.

[161] Besmira Nushi, Ece Kamar y Eric Horvitz. 2018. Hacia una IA responsable: análisis híbridos hombre-máquina para caracterizar fallas del sistema. En la VI Conferencia AAAI sobre Computación Humana y Crowdsourcing.

[162] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye y Alexander Mordvintsev. 2018. Los pilares de la interpretabilidad. Destilar (2018). https://doi.org/10.23915/distill.00010 https://distill. pub/2018/bloques de construcción.

[163] Cathy O'Neil. 2016. Armas de destrucción matemática: cómo los macrodatos aumentan la desigualdad y amenazan la democracia. Amplio-camino Libros.

[164] Sean Penney, Jonathan Dodge, Claudia Hilderbrand, Andrew Anderson, Logan Simpson y Margaret Burnett. 2018. Hacia la búsqueda de comprensión sobre los agentes de StarCraft: un estudio empírico. En la 23ª Conferencia Internacional sobre Interfaces de Usuario Inteligentes (IUI'18). ACM, Nueva York, Nueva York, 225–237. https://doi.org/10.1145/3172944.3172946

[165] Nicola Pezzotti, Thomas Höllt, Jan Van Gemert, Boudewijn PF Lelieveldt, Elmar Eisemann y Anna Vilanova. 2018. DeepEyes: análisis visual progresivo para el diseño de redes neuronales profundas. Transacciones IEEE sobre visualización y gráficos por computadora 24, 1 (2018), 98–108.

[166] Nina Poerner, Hinrich Schütze y Benjamin Roth. 2018. Evaluación de métodos de explicación de redes neuronales utilizando documentos híbridos y predicción morfológica. En la 56ª Reunión Anual de la Asociación de Lingüística Computacional (ACL'18).

[167] Brett Poulin, Roman Eisner, Duane Szafron, Paul Lu, Russell Greiner, David S. Wishart, Alona Fyshe, Brandon Pearcy, Cam MacDonell y John Anvik. 2006. Explicación visual de la evidencia con clasificadores aditivos. En Actas de la 18.ª Conferencia sobre aplicaciones innovadoras de la inteligencia artificial, vol. 2. Prensa AAAI, 1822–1829.

[168] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan y Hanna Wallach. 2018. Manipulación y medición de la interpretabilidad del modelo. arXiv:1802.07810. https://arxiv.org/abs/1802.07810.

[169] Perla Pu y Li Chen. 2006. Fomento de la confianza con interfaces explicativas. En actas de la 11ª Conferencia Internacional sobre Interfaces de Usuario Inteligentes. MCA, 93-100.

[170] Emilee Rader, Kelley Cotter y Janghee Cho. 2018. Explicaciones como mecanismos para apoyar la transparencia algorítmica. En actas de la Conferencia CHI de 2018 sobre factores humanos en sistemas informáticos. ACM, 103.

[171] Emilee Rader y Rebecca Gray. 2015. Comprender las creencias de los usuarios sobre la curación algorítmica en el servicio de noticias de Facebook. En actas de la 33ª Conferencia Anual de ACM sobre factores humanos en sistemas informáticos. ACM, 173–182.

[172] Marco Tulio Ribeiro, Sameer Singh y Carlos Guestrin. 2016. "¿Por qué debería confiar en tí?" Explicar las predicciones de cualquier clasificador. En actas de la 22ª Conferencia internacional ACM SIGKDD sobre descubrimiento de conocimientos y minería de datos. MCA, 1135-1144.

[173] Marco Tulio Ribeiro, Sameer Singh y Carlos Guestrin. 2018. Anclas: explicaciones independientes del modelo de alta precisión. En Conferencia AAAI sobre Inteligencia Artificial.

[174] Caleb Robinson, Fred Hohman y Bistra Dilkina. 2017. Un enfoque de aprendizaje profundo para la estimación de población a partir de imágenes satelitales. En actas del 1er Taller ACM SIGSPATIAL sobre Humanidades Geoespaciales. ACM, 47–54.

[175] Marko Robnik-Šikonja y Marko Bohanec. 2018. Explicaciones de modelos de predicción basadas en perturbaciones. en humano y aprendizaje automático. Saltador, 159-175.

[176] Stephanie Rosenthal, Sai P. Selvaraj y Manuela Veloso. 2016. Verbalización: Narración de la experiencia de un robot autónomo. En actas de la 25ª Conferencia Internacional Conjunta sobre Inteligencia Artificial. 862–868.

[177] Andrew Slavin Ross y Finale Doshi-Velez. 2018. Mejorar la solidez adversarial y la interpretabilidad de los profundos redes neuronales regularizando sus gradientes de entrada. En la 32ª Conferencia AAAI sobre Inteligencia Artificial.

[178] Andrew Slavin Ross, Michael C. Hughes y Finale Doshi-Velez. 2017. Correcto por las razones correctas: entrenar modelos diferenciables restringiendo sus explicaciones. En actas de la 26ª Conferencia Internacional Conjunta sobre Inteligencia Artificial (IJCAI'17). 2662–2670. https://doi.org/10.24963/ijcai.2017/371 [179] Stephen Rudolph, Anya Savikhin y David
S. Ebert. 2009. FinVis: Análisis visual aplicado a la planificación financiera personal. En Simposio IEEE sobre ciencia y tecnología de análisis visual, 2009. Citeseer, 195–202.
[180] Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Aldo Lee, Daniel Weiskopf, Stephen North y Daniel Keim. 2016. Aprendizaje automático centrado en el ser humano a través de visualización interactiva. En el 24º Simposio Europeo sobre Redes Neuronales Artificiales, Inteligencia Computacional y Aprendizaje Automático. 641–646.
[181] Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis y Daniel A. Keim. 2016. El papel de la incertidumbre, la conciencia y la confianza en el análisis visual. Transacciones IEEE sobre visualización y gráficos por computadora 22, 1 (2016), 240–249.
[182] Bahador Saket, Arjun Srinivasan, Eric D. Ragan y Alex Endert. 2017. Evaluación de codificaciones gráficas interactivas para visualización de datos. Transacciones IEEE sobre visualización y gráficos por computadora 24, 3 (2017), 1316–1330.
[183] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin y Klaus-Robert Müller. 2017. Evaluación de la visualización de lo que ha aprendido una red neuronal profunda. Transacciones IEEE sobre redes neuronales y sistemas de aprendizaje 28, 11 (2017), 2660–2673.
[184] Christian Sandvig, Kevin Hamilton, Karrie Karahalios y Cedric Langbort. 2014. Algoritmos de auditoría: métodos de investigación para detectar discriminación en plataformas de Internet. Datos y discriminación: convertir preocupaciones críticas en investigaciones productivas (2014), 1–23.
[185] Martin Schaffernicht y Stefan N. Groesser. 2011. Un método integral para comparar modelos mentales de dy- sistemas námicos. Revista europea de investigación operativa 210, 1 (2011), 57–67.
[186] Ute Schmid, Christina Zeller, Tarek Besold, Alireza Tamaddoni-Nezhad y Stephen Muggleton. 2016. ¿Cómo afecta la invención de predicados a la comprensibilidad humana? En Congreso Internacional sobre Programación Lógica Inductiva. Saltador, 52–67.
[187] Philipp Schmidt y Félix Biessmann. 2019. Cuantificación de la interpretabilidad y la confianza en los sistemas de aprendizaje automático. arXiv:1901.08558. https://arxiv.org/abs/1901.08558.
[188] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh y Dhruv Batra. 2017. Grad-cam: explicaciones visuales de redes profundas mediante localización basada en gradientes. En actas de la Conferencia Internacional IEEE sobre Visión por Computadora. 618–626.
[189] Avanti Shrikumar, Peyton Greenside y Anshul Kundaje. 2017. Aprender características importantes mediante la propagación de diferencias de activación. En Actas de la 34ª Conferencia Internacional sobre Aprendizaje Automático, Volumen 70. JMLR. org, 3145–3153.

[190] Karen Simonyan, Andrea Vedaldi y Andrew Zisserman. 2013. En lo profundo de las redes convolucionales: visualización de modelos de clasificación de imágenes y mapas de prominencia. arXiv:1312.6034. https://arxiv.org/abs/1312.6034.
[191] Daniel Smilkov, Shan Carter, D. Sculley, Fernanda B. Viégas y Martin Wattenberg. 2017. Manipulación directa Visualización de redes profundas. arXiv:1708.03788. http://arxiv.org/abs/1708.03788.
[192] Thilo Spinner, Udo Schlegel, Hanna Schäfer y Mennatallah El-Assady. 2019. explAIner: un marco de análisis visual para el aprendizaje automático interactivo y explicable. Transacciones IEEE sobre visualización y gráficos por computadora 26, 1 (2020), 1064–1074.

[193] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister y Alexander M. Rush. 2018. LSTMVis: una herramienta para el análisis visual de la dinámica de estados ocultos en redes neuronales recurrentes. Transacciones IEEE sobre visualización y gráficos por computadora 24, 1 (2018), 667–676.
[194] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan y Jonathan Herlocker. 2009. Interactuar significativamente con sistemas de aprendizaje automático: tres experimentos. Revista internacional de estudios humanos-computadores 67, 8 (2009), 639–662.
[195] Simone Stumpf, Simonas Skrebe, Graeme Aymer y Julie Hobson. 2018. Explicando los sistemas de calefacción inteligentes para dis- coraje jugando con un comportamiento optimizado.
[196] Latanya Sweeney. 2013. Discriminación en la entrega de publicidad online. Comunicaciones de la JCA 56, 5 (2013), 44–54.
[197] Jiliang Tang, Huiji Gao, Huan Liu y Atish Das Sarma. 2012. eTrust: Comprender la evolución de la confianza en un mundo en línea. En actas de la 18.ª Conferencia internacional ACM SIGKDD sobre descubrimiento de conocimientos y minería de datos. ACM, 253–261.
[198] Christina Tikkinen-Piri, Anna Rohunen y Jouni Markkula. 2018. Reglamento general de protección de datos de la UE: cambios e implicaciones para las empresas de recopilación de datos personales. Revisión de seguridad y derecho informático 34, 1 (2018), 134–153.
[199] Nava Tintarev y Judith Masthoff. 2011. Diseño y evaluación de explicaciones para sistemas de recomendación. En el Manual de sistemas recomendadores. Saltador, 479–510.
[200] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece y Supriyo Chakraborty. 2018. ¿Interpretable para quién? Un modelo basado en roles para analizar sistemas de aprendizaje automático interpretables. arXiv:1806.07552. https://arxiv.org/abs/1806.07552 .

[201] Matteo Turilli y Luciano Floridi. 2009. La ética de la transparencia informativa. Ética y tecnología de la información 11, 2 (2009), 105-112.

[202] Jo Vermeulen, Geert Vanderhulst, Kris Luyten y Karin Coninx. 2010. PervasiveCrystal: Preguntar y responder preguntas sobre por qué y por qué no sobre aplicaciones de computación ubicua. En 2010 Sexta Conferencia Internacional sobre Entornos Inteligentes (IE'10). IEEE, 271–276.

[203] Sandra Wachter, Brent Mittelstadt y Chris Russell. 2017. Explicaciones contrafactuales sin abrir la caja negra: Decisiones automatizadas y el RGPD. Revista de Derecho y Tecnología de Harvard 31 (2017), 841.

[204] Danding Wang, Qian Yang, Ashraf Abdul y Brian Y. Lim. 2019. Diseño de IA explicable centrada en el usuario basada en la teoría. En Actas de la Conferencia CHI de 2019 sobre factores humanos en sistemas informáticos (CHI'19). ACM, Nueva York, NY, Artículo 601, 15 páginas.

[205] Fulton Wang y Cynthia Rudin. 2015. Listas de reglas en caída. En Inteligencia Artificial y Estadística. 1013–1022.

[206] Qianwen Wang, Jun Yuan, Shuxin Chen, Hang Su, Huamin Qu y Shixia Liu. 2019. Genealogía visual de redes neuronales profundas. Transacciones IEEE sobre visualización y gráficos por computadora 26, 11 (2020), 3340–3352.

[207] Daniel S. Weld y Gagan Bansal. 2019. El desafío de crear inteligencia inteligible. Comunicaciones de la JCA 62, 6 (mayo de 2019), 70–79.

[208] Adrián Weller. 2017. Retos para la transparencia. arXiv:1708.01870. https://arxiv.org/abs/1708.01870.

[209] Gesa Wiegand, Matthias Schmidmaier, Thomas Weber, Yuanting Liu y Heinrich Hussmann. 2019. Yo conduzco-tú confías: Explicando el comportamiento de conducción de los coches autónomos. En resúmenes ampliados de la Conferencia CHI de 2019 sobre factores humanos en sistemas informáticos. ACM, LBW0163.

[210] James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur y Vern Crow. 1995. Visualizando lo no visual: análisis espacial e interacción con información de documentos de texto. En Actas de visualización de información, 1995. IEEE, 51–58.

[211] Kanit Wongsuphasawat, Daniel Smilkov, James Wexler, Jimbo Wilson, Dandelion Mane, Doug Fritz, Dilip Krishnan, Fernanda B. Viégas y Martin Wattenberg. 2017. Visualización de gráficos de flujo de datos de modelos de aprendizaje profundo en tensorflow. Transacciones IEEE sobre visualización y gráficos por computadora 24, 1 (2017), 1–12.

[212] Samuel C. Woolley. 2016. Automatización del poder: interferencia de los robots sociales en la política global. Primer lunes 21, 4 (2016).

[213] Mike Wu, Michael C. Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth y Finale Doshi-Velez. 2018. Más allá de la escasez: regularización de árboles de modelos profundos para la interpretabilidad. En la 32ª Conferencia AAAI sobre Inteligencia Artificial.

[214] Ming Yin, Jennifer Wortman Vaughan y Hanna Wallach. 2019. Comprender el efecto de la precisión en la confianza en los modelos de aprendizaje automático. En actas de la Conferencia CHI de 2019 sobre factores humanos en sistemas informáticos. 1–12.

[215] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs y Hod Lipson. 2015. Comprensión de las redes neuronales. a través de una visualización profunda. En el Taller de Aprendizaje Profundo ICML 2015.

[216] Rulei Yu y Lei Shi. 2018. Una taxonomía basada en usuarios para la visualización de aprendizaje profundo. Informática Visual 2, 3 (2018), 147–154.

[217] Tom Zahavy, Nir Ben-Zrihem y Shie Mannor. 2016. Atenuar la caja negra: comprender los DQN. En internacional Jornada sobre Aprendizaje Automático. 1899-1908.

[218] Tal Zarsky. 2016. El problema de las decisiones algorítmicas: una hoja de ruta analítica para examinar la eficiencia y la equidad en la toma de decisiones automatizada y opaca. Ciencia, tecnología y valores humanos 41, 1 (2016), 118–132.

[219] Matthew D. Zeiler y Rob Fergus. 2014. Visualización y comprensión de redes convolucionales. En Conferencia Europea sobre Visión por Computador. Saltador, 818–833.

[220] Quanshi Zhang, Wenguan Wang y Song-Chun Zhu. 2018. Examinando las representaciones de CNN con respecto al sesgo del conjunto de datos. En la 32ª Conferencia AAAI sobre Inteligencia Artificial.

[221] Quan-shi Zhang y Song-Chun Zhu. 2018. Interpretabilidad visual para el aprendizaje profundo: una encuesta. Fronteras de la información Tecnología de ción e ingeniería electrónica 19, 1 (2018), 27–39.

[222] Yunfeng Zhang, Q. Vera Liao y Rachel KE Bellamy. 2020. Efecto de la confianza y la explicación sobre la precisión y la calibración de la confianza en la toma de decisiones asistida por IA. En Actas de la Conferencia de 2020 sobre Equidad, Responsabilidad y Transparencia (FAT*'20).

[223] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju y Avishek Anand. 2019. Disonancia entre la comprensión humana y la máquina. Actas de la ACM sobre interacción persona-computadora 3, CSCW (2019), 56.

[224] Wen Zhong, Cong Xie, Yuan Zhong, Yang Wang, Wei Xu, Shenghui Cheng y Klaus Mueller. 2017. Análisis visual evolutivo de redes neuronales profundas. En Taller ICML sobre visualización para aprendizaje profundo.

[225] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra y G. Michael Youngblood. 2018. IA explicable para diseñadores: una perspectiva centrada en el ser humano sobre la cocreación de iniciativas mixtas. En la Conferencia IEEE de 2018 sobre juegos e inteligencia computacional (CIG'18). IEEE, 1–8.

[226] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel y Max Welling. 2017. Visualización de decisiones de redes neuronales profundas: Análisis de diferencias de predicción. arXiv:1702.04595. http://arxiv.org/abs/1702.04595.

Recibido en noviembre de 2019; revisado en julio de 2020; aceptado en julio de 2020