Explicabilidad (de la inteligencia artificial)*

Explainable ArtificiaI Intelligence

Lucía Ortiz de Zárate Alcarazo
Universidad Autónoma de Madrid
ORCID ID 0000-0001-7775-4378
lucia.ortizdezarate@uam.es

Cita recomendada:

Ortiz de Zárate Alcarazo, L. (2022). Explicabilidad (de la inteligencia artificial). *Eunomía. Revista en Cultura de la Legalidad*, 22, 328-344.

DOI: https://doi.org/10.20318/eunomia.2022.6819

Recibido / received: 22/12/2021
Aceptado / accepted: 03/02/2022

Resumen
La Inteligencia Artificial (IA) se ha convertido en uno de los ejes centrales de las estrategias de digitalización en la Unión Europea (UE). A pesar de sus grandes beneficios este grupo de tecnologías emergentes también presenta importantes riesgos y desafíos éticos que deben abordarse. Por ello, la UE ha propuesto cuatro principios éticos fundamentales para la IA: la justicia, el respeto por la autonomía humana, la previsión del daño y la explicabilidad. El concepto central de este trabajo es la explicabilidad de la IA (XAI). De este modo, además de poner en contexto la importancia de este principio ético, se abordarán los aspectos centrales en torno al mismo: qué es la explicabilidad; por qué son importantes las explicaciones (en el ámbito de la IA); cuándo es necesario dar explicaciones y de qué tipo de explicaciones disponemos para hacer comprender a la ciudadanía las decisiones tomadas por sistema de IA.

Palabras clave
Inteligencia Artificial, explicabilidad, interpretabilidad, rendición de cuentas, gobernanza algorítmica, ética, ética de la Inteligencia Artificial.

Abstract
Artificial Intelligence (AI) has become one of the central axes of digitization strategies in the European Union (EU). Despite its great benefits, this group of emerging technologies also presents significant ethical risks and challenges that must be addressed. For this reason, the EU has proposed four fundamental ethical principles for AI: fairness, respect for human autonomy, prevention of harm, and explicability. The central concept of the paper is the explicability of AI (XAI). Besides contextualizing the relevance of XAI, this work aims to tackle some of the central aspects around it. Therefore, it will address what explicability is, why

* Este trabajo se ha realizado con el apoyo del Programa On TRUST–CM H 2019/HUM-5699 de la Comunidad de Madrid y el Fondo Social Europeo.

explanations are important (in the field of AI), when it is necessary to give explanations, and the different forms of explanations available to make citizens understand the decisions made by AI systems.

Keywords
Artificial Intelligence, explainability, interpretability, accountability, algorithmic governance, ethics, ethics of Artificial Intelligence.

SUMARIO. 1. Introducción. 2. La ética de la Inteligencia Artificial. 2.1. Problemas éticos de la IA. 2.2. Principios éticos de la IA. 3. La explicabilidad de la IA (XAI). 3.1. El concepto de explicabilidad 4. ¿Por qué son importantes las explicaciones? 4.1. ¿Cuándo hay que dar explicaciones? 5. Tipos de explicaciones. 6. Conclusiones.

1. Introducción

La tecnología se ha convertido en un eje fundamental en la vida de las personas. Sería imposible poder dar sentido a las sociedades actuales sin tener en cuenta el papel de la tecnología y su interacción con la ciudadanía a todos los niveles. Aunque en el ámbito de la filosofía existen diversas discusiones, aún vigentes, sobre qué es exactamente la tecnología (Dusek, 2006) –algunos se remontan tan atrás como hasta la invención de la rueda, mientras que todos señalan la revolución industrial como el momento fundacional– normalmente se entiende que la tecnología son todos aquellos sistemas compuestos por *hardware* y *software* (Orlikowski, 1992) y será así como la entenderemos durante el presente texto.

La presencia de la tecnología en distintas parcelas de nuestra vida ha dado lugar a cambios importantes en nuestras costumbres, dinámicas y formas de socializar habituales, a veces, de forma casi imperceptible. En el sector público estos cambios han sido más notorios ya que la introducción de la tecnología y el tránsito hacia modelos basados en este tipo sistemas son el resultado de proyectos políticos conocidos desde hace décadas (Criado, 2009). Durante los últimos años, la aparición de las llamadas tecnologías disruptivas (IA, *Blockchain*, *Internet of Things* (IoT), *Big Data*, etc.) ha provocado una nueva ola tecnológica en el sector público. La posibilidad real de incorporar estas nuevas tecnologías en las administraciones públicas, los órganos de gobierno y la elaboración de políticas públicas ha acelerado un cambio de paradigma en el sector: ‹‹la gobernanza algorítmica›› (Just y Latzer, 2017; Gasser y Almeida, 2017). Este nuevo modelo de gobernanza se caracterizaría por la relevancia que los algoritmos (inteligentes) adquirirían en la mediación de las relaciones entre la ciudadanía y las administraciones públicas, la realización de trámites burocráticos, la asistencia en la toma de decisiones, tanto de gobierno como dirigidas a la ciudadanía, la tramitación de demandas, la ejecución de servicios públicos (de atención al ciudadano, sanidad, justicia, empleo, etc.), etc. (Margetts y Dorobantu, 2019).

Todos estos cambios procedentes de la algoritmización del sector público (Meijer y Grimmelikhuijsen, 2020), se prevé que traigan importantes mejoras tanto para el funcionamiento interno del mismo, como de cara a los ciudadanos y a las ciudadanas. Normalmente, al hablar de las ventajas de la introducción de tecnologías disruptivas en el sector público, se destacan los aumentos en la eficacia, la eficiencia, la disponibilidad y la automatización de muchos servicios, la personalización de los mismos, la capacidad de predicción y detección de problemas, etc. Por ello, la mayor

parte de los países de la OCDE, entre ellos España, ya cuenta con planes y estrategias para promover la adopción de algunas de estas tecnologías emergentes y, más concretamente, la IA (Comisión Europea, 2019; Ministerio de Asuntos Económicos y Transformación Digital, 2021).

Sin embargo, también existen muchos retos y riesgos asociados a la adopción y uso de la IA en el sector público. Por ello, la UE lleva desde hace años trabajando en el desarrollo de un marco ético y jurídico común para todo el territorio que garantice que todos aquellos sistemas de IA que se usen dentro de la UE cumplan ciertos estándares éticos y jurídicos, de tal modo que los derechos y libertades de los ciudadanos de la Unión sean respetados y adecuadamente protegidos (Comisión Europea, 2019).

2. La ética de la Inteligencia Artificial

2.1. Problemas éticos de la IA

Los avances en IA y sus, aún limitadas, aplicaciones en distintos ámbitos como la sanidad, la seguridad o la justicia, han puesto de manifiesto los riesgos que un mal uso de este grupo de tecnologías puede suponer para la ciudadanía. Entre estos, los más conocidos son la presencia de sesgos, la opacidad algorítmica, la falta de transparencia en los datos y los algoritmos, el determinismo tecnológico, la violación de la privacidad, la falta de rendición de cuentas, etc. (Cath et al., 2018; Coeckelbergh, 2020a). Los problemas de sesgos asociados a la IA surgen, principalmente, por dos motivos: o bien por una falta de representatividad en los datos de los que se nutren los algoritmos (sesgos en los datos); o por sesgos presentes en el diseño de los algoritmos.

En el caso de los sesgos, estos pueden ser de distinto tipo. Hasta la fecha se han registrado discriminaciones provocadas por sesgos de raza, género, orientación sexual, etc. (Guevara et al., 2021). Sesgos de raza han sido visibles, por ejemplo, en el caso Loomis (Berchi, 2020). En este caso se condenó a un hombre negro basándose en las estimaciones de una IA (COMPAS) y años después se descubrió que la raza era una de las variables que el algoritmo tenía en cuenta para hacer sus cálculos y, por tanto, tomar decisiones. Esto se debía a que en las bases de datos que servían de aprendizaje al *software* COMPAS había más hombres negros condenados y con sentencias superiores al de los hombres blancos. De este modo, COMPAS ‹‹aprendió›› que el hecho de pertenecer a una raza concreta debía estar relacionado con las posibilidades para delinquir.

Otro tipo de sesgos también muy conocidos son los sesgos de género (De Zárate-Alcarazo y Guevara-Gómez, 2021). Por ejemplo, en el año 2014 la empresa Amazon puso en marcha un sistema de IA para acelerar y optimizar su proceso de contratación de personal. Al cabo del tiempo se descubrió que este sistema discriminaba de forma sistemática a las mujeres dando preferencia a los currículos de los hombres (D.J.O., 2018). Esto se debía a que los datos con los que había sido entrenada la IA para llevar a cabo esta tarea reflejaban que anteriormente la mayor parte de las personas contratadas para ese tipo de puestos eran hombres que reunían unas características muy concretas. De este modo, el sistema tomó el género como una variable que restaba a la hora de desempeñar ciertas tareas. Por ello, todos aquellos currículos en los que aparecían referencias a mujeres eran automáticamente calificados con menor puntuación que los de los hombres.

Además de los sesgos, otros problemas éticos como el determinismo tecnológico, la falta de privacidad, la opacidad o la rendición de cuentas, mencionados

anteriormente, son también riesgos importantes de la IA. El determinismo tecnológico se refiere a la posibilidad de que en un escenario asentado de gobernanza algorítmica las recomendaciones y, en algunos casos, decisiones, de los algoritmos inteligentes condicionen excesivamente la vida de la ciudadanía, de tal modo que su libertad de elección se viese reducida significativamente sin ser ellos conscientes de este suceso (Taddeo y Floridi, 2018). Otro de los grandes problemas de la IA es la falta de privacidad (Véliz, 2020). El correcto funcionamiento de la IA conlleva, necesariamente, el uso de grandes cantidades de datos. Estos datos son necesarios para entrenar a los sistemas de IA y optimizar sus procesos de aprendizaje y toma de decisiones. De este modo, una de las principales premisas de la IA sería «cuantos más datos mejor» o «a más datos, mejor funcionamiento». Aunque esto no es del todo cierto, pues en el correcto funcionamiento de la IA no solo importa la cantidad de datos, sino también la calidad de los mismos, la realidad es que los datos personales se han convertido en uno de los bienes más deseados (Khatri y Brown, 2010). Sin embargo, la alta demanda de datos entra en conflicto con el derecho de las personas a mantener su propia privacidad y a reservarse cierta información para ellos mismos. La recopilación de cierta información sensible, como es toda aquella que hace referencia a las prácticas sexuales de las personas, las relaciones familiares, preferencias vitales, etc., es una práctica ilegal, y, sin embargo, cada vez son más frecuentes los casos en los que distintas entidades demandan este tipo de información. A veces de forma explícita y otras, sin nuestro consentimiento (Zuboff, 2019).

Finalmente, otro de los problemas fundamentales que se dan con el uso de la IA son las dificultades implícitas en los procesos que permiten garantizar la rendición de cuentas (Doshi-Velez et al., 2017b). La rendición de cuentas es uno de los pilares básicos sobre los que se asientan las democracias liberales modernas (Linz y Miley, 2014). Estos procedimientos son esenciales en cualquier sistema garantista, es decir, en todos aquellos sistemas en los que, cuando se produce una deficiencia, malfuncionamiento y/o error, los responsables del mismo pueden ser sometidos a un proceso (judicial, social y/o político) a través del cual deben dar cuenta de sus hechos y decisiones y, en la medida que les corresponda, pagar por ellos mediante el cumplimiento de una sentencia (judicial, social y/o política).

En el caso de la IA, garantizar la rendición de cuentas es una cuestión especialmente compleja por distintos motivos. Pensemos en el caso de los vehículos autónomos, que es uno de los casos más socorridos para explicitar esta problemática. Si un vehículo de estas características está circulando por una vía habilitada para ello, a una velocidad correcta, pero un peatón cruzase por un lugar donde no se puede de tal forma que el vehículo le atropellase y la persona falleciese, ¿quién sería el responsable? En este tipo de casos, cuando el peatón es el infractor, la atribución de responsabilidades parece más sencilla que en otras circunstancias. El responsable del atropello es el infractor. Pero ¿qué sucedería en el caso de que el peatón cruzase por un paso de cebra con el semáforo en verde y el vehículo le atropellase? ¿quién sería entonces el responsable? ¿la persona que iba dentro, los ingenieros que diseñaron el vehículo, la empresa que lo pone en venta, el gobierno por permitir ese tipo de vehículos, el propio vehículo? En este caso, la atribución de responsabilidades y, por tanto, la rendición de cuentas se complica considerablemente (Quarta y Trezza, 2021; Trezza, 2021).

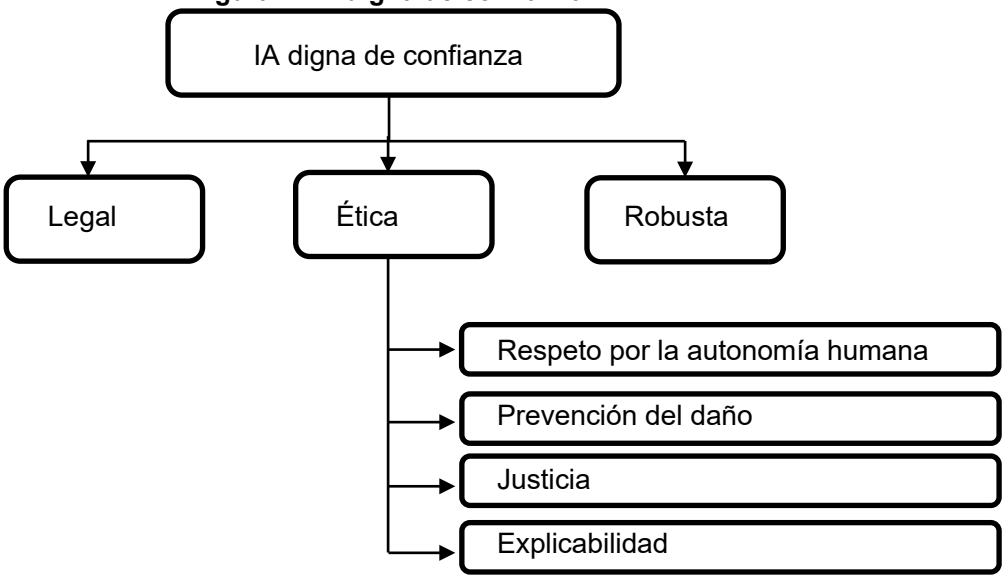
Sin embargo, no solo en el caso de los vehículos autónomos se manifiestan los problemas de rendición de cuentas de la IA. Cuando se produce un sesgo del tipo de los mencionados anteriormente (género, raza, etc.), de tal modo que un sistema de IA recomienda una decisión discriminatoria que perjudica directamente a una

persona hasta el punto de vulnerar sus derechos fundamentales, ¿cómo dirimimos responsabilidades? ¿y cuando se viola la privacidad y la intimidad de las personas? ¿y cuando un médico hace un diagnóstico o prescribe un tratamiento erróneo porque el sistema de IA que le asistía en la toma de decisiones falló? Los problemas de rendición de cuentas se encuentran presentes en todos aquellos procesos de toma de decisiones en los que se vea involucrado un sistema de IA, ya que se trata de una cuestión fundamental para poder garantizar que estas nuevas herramientas se someten y cumplen los mismos estándares y leyes que el resto (Wachter et al., 2017c).

2.2. Principios éticos de la IA

Debido a todos estos (y otros) problemas éticos y riesgos asociados a la IA, la Comisión Europea lleva desde 2018 trabajando en la elaboración de un marco ético y legislativo que permita acomodar el uso de estas nuevas tecnologías con los elementos propios de la cultura de legalidad vigente, es decir, con las normas, deberes y derechos fundamentales que rigen la vida de todos los ciudadanos y las ciudadanas de la UE (De Zárate-Alcarazo y Guevara-Gómez, 2021). En este sentido, el marco ético de la UE busca garantizar que todo sistema de IA usado en territorio europeo sea «digno de confianza»[1], lo que implica que las tecnologías de IA sean legales, éticas y robustas (Comisión Europea, 2019). A su vez, la legalidad de la IA depende del cumplimiento de la regulación en materia de IA propuesta por UE (Comisión Europea, 2021a). La robustez de la IA depende de cuestiones principalmente técnicas, es decir, de la capacidad de los sistemas inteligentes para funcionar correctamente, así como de resistir ciberataques o amenazas de otro tipo. Finalmente, la IA ética en Europa se basa en el cumplimiento de cuatro principios fundamentales que deben guiar el desarrollo, la implementación y el uso de la IA dentro del territorio de la UE. Esos principios éticos son: el respeto por la autonomía humana, la prevención del daño, la justicia y la explicabilidad (Comisión Europea, 2019). Estos cuatro principios constituyen los fundamentos éticos necesarios para garantizar una IA que sea digna de confianza (Figura 1).

Figura 1: IA digna de confianza



[1] Digna de confianza se ha traducido de la expresión en inglés «*trustworthy*» que es la que usa la UE para referirse a este tipo de IA.

Fuente: elaboración propia

Según la normativa y marco ético europeo (Comisión Europea 2018, 2019, 2020), el cumplimiento de todos estos principios y características debería garantizar la compatibilidad entre el uso de la IA y el respecto y salvaguarda de los derechos fundamentales de los ciudadanos de la UE. El principio de respeto por la autonomía humana hace referencia a la necesidad (amparada por la nueva regulación europea) de prohibir el uso de todas aquellas tecnologías de IA cuya finalidad sea manipular y/o coaccionar a las personas para que tomen ciertas acciones o adquieran determinados comportamientos en contra de su voluntad. En el caso de la IA el principal peligro es que esta amenaza tenga lugar de forma subliminal, es decir, sin que los usuarios de la tecnología sean conscientes de que se les está manipulando y, por tanto, violando su autonomía. La prevención del daño pretende asegurar que ninguna aplicación de la IA pueda causar un daño físico o moral a las personas. En virtud de este principio se prohibiría el uso de armas autónomas, se blindaría la protección de los datos de los ciudadanos, etc. El principio de justicia es, probablemente, el principio más amplio y esencial de todos, pues todos los demás se engloban en este. Por un lado, la justicia sustancial implica una distribución justa e igualitaria de los beneficios y los costes de la IA, respetar el principio de proporcionalidad, así como el compromiso de que estas tecnologías no comentan sesgos ni provoquen discriminaciones ni estigmas de ningún tipo prestando especial atención a todos aquellos grupos vulnerables debido a discriminaciones ya existentes. Por otro lado, la justicia procedimental es de carácter retroactivo y busca garantizar que, en el caso de que una decisión tomada, propuesta o asesorada por un sistema inteligente salga mal, será posible la revisión del proceso en su totalidad para que se puedan rendir cuentas (Comisión Europea, 2019).

Finalmente, la UE ha incluido el principio de explicabilidad como un elemento crucial para poder generar y mantener la confianza en la IA. Aunque la propia UE no proporciona una definición precisa de qué es la explicabilidad algorítmica o explicabilidad de la IA (XAI), sí hace referencia a la necesidad de que los procesos sean transparentes y que las decisiones sean bien comunicadas, así como la imposibilidad de que ciertas acciones sean contestadas si no se proporcionan las explicaciones necesarias (Comisión Europea, 2019, p. 13).

3. La explicabilidad de la IA (XAI)

La XAI es uno de los cuatro principios éticos propuestos por la UE para garantizar una IA que sea digna de confianza. Sin embargo, de entre los cuatro principios propuestos, probablemente se trate del menos conocido y explorado de todos ellos a pesar de que la necesidad de proporcionar explicaciones no es algo novedoso en el ámbito de la ética y la aplicación de la justicia (Díez et al., 2011). Las explicaciones tienen un papel muy destacado a la hora de justificar, o no, moralmente ciertas acciones y/o decisiones y son varias las doctrinas éticas que hacen referencia a este hecho, llegando a situarlo en el centro de la racionalidad moral (Copp, 1990). También en muchos casos involucrados en la aplicación de la justicia proporcionar explicaciones es un requisito indispensable para poder atribuir responsabilidad, atenuar condenas, etc. (Doshi-Velez et al., 2017b). Incluso en el ámbito tecnológico la explicabilidad no es una característica desconocida, en el caso de varias tecnologías anteriores a la IA ya eran importantes las explicaciones (Coeckelbergh, 2020b).

A pesar de esto, debido al auge de las nuevas tecnologías de IA el principio de explicabilidad ha ganado relevancia y, actualmente, se trata de uno de los

principales ámbitos emergentes de estudio científico, tanto en áreas científicas y técnicas, como en el derecho, la ciencia política y la filosofía (Miller, 2017). Las previsiones de que la IA se aplique a la mayor parte de áreas que conforman el sector público, así como otros espacios más propios de la vida privada de las personas, pero igualmente importantes, ha convertido en una necesidad repensar y expandir los horizontes de la explicabilidad algorítmica. En el paradigma de la gobernanza algorítmica hacia el que caminan Europa, EE. UU. y China, gran parte de las decisiones más importantes que afectan a la vida de la ciudadanía estarán asesoradas y, en ciertos casos, tomadas por algoritmos inteligentes. En este escenario, como se explicará más adelante, la explicabilidad de los algoritmos se convertirá en uno de los pilares esenciales para garantizar el respeto por la cultura de la legalidad, la democracia y la justicia.

3.1. El concepto de explicabilidad

En gran parte, debido a la situación emergente que vive el área de estudio relacionado con la explicabilidad de la IA aún no existe una definición consensuada de qué significa exactamente explicabilidad. El principal problema en torno a la definición exacta del concepto de explicabilidad tiene que ver con la proximidad del mismo con otros términos similares como «interpretabilidad», «comprensibilidad», «inteligibilidad», «legibilidad» o, incluso, «transparencia». En este sentido, es frecuente encontrar en la literatura académica especializada un uso indistinto entre los términos «explicabilidad» e «interpretabilidad». Así, la interpretabilidad, o la explicabilidad, harían referencia a la «habilidad para explicar o presentar sistemas de IA en términos comprensibles para los humanos» (Doshi-Velez y Kim, 2017). También hay quienes (Cabitza et al., 2019; Lou et al., 2013) entienden que no solo «explicabilidad» e «interpretabilidad» son términos equivalentes y, por tanto, intercambiables, sino que también lo son «comprensibilidad», «inteligibilidad» y «legibilidad».

Sin embargo, otras personas consideran que estos conceptos son distintos y, por tanto, que «interpretabilidad» no es un sinónimo de «explicabilidad» aunque sí se encuentran próximos entre sí. En este caso la interpretabilidad tendría que ver con la capacidad de un sistema de IA de ser comprensible para los humanos, mientras que la explicabilidad iría más allá de la interpretabilidad al englobar también la propiedad de ser fidedigno con la realidad. Es decir, que un sistema de IA explicable sería aquel que pudiese ser comprensible para los seres humanos al mismo tiempo que su contenido (la explicación) se corresponde con la realidad que trata de explicar (Markus et al., 2021).

De una forma u otra, lo que late de fondo en toda la discusión en torno al concepto de explicabilidad es la necesidad de hacer comprensible, entendible o inteligible (tomando todos estos términos como sinónimos) los sistemas de IA a las personas, es decir, de dar a entender la racionalidad o criterios detrás de una decisión (Lipton, 2018; Miller, 2017). Para realizar esta tarea, es decir, para poder llevar a cabo un ejercicio de interpretación entre el lenguaje computacional y el lenguaje humano, es necesario que los datos y algoritmos que componen el sistema, así como todos los pasos que han tenido lugar hasta alcanzar el resultado final sean visibles, es decir, transparentes. En este sentido, la transparencia sería la habilidad para hacer visible las componentes de un sistema de IA y sería condición necesaria, pero no suficiente para se cumpliera con el principio de explicabilidad. Para esto es necesario, como hemos mencionado, un ejercicio de interpretación posterior que haga comprensible para los seres humanos todo lo que tiene lugar dentro del sistema.

Dada la falta de consenso, en este trabajo usaremos la definición de explicabilidad propuesta por la UNESCO (2021), habiendo sido esta abalada por 193 estados en todo el mundo. Así, la explicabilidad «hace referencia al hacer inteligible los resultados de los sistemas de IA. La XAI también hace referencia a la comprensibilidad de los datos, procesos y comportamientos de los distintos bloques algorítmicos y como cada uno de ellos contribuye al resultado del sistema. Así, la explicabilidad está estrechamente relacionada con la transparencia, ya que los procesos y sub-procesos que conducen a los resultados deberían ser comprensibles y trazables, apropiados para el contexto»[2].

4. ¿Por qué son importantes las explicaciones?

Otra cuestión que es importante abordar cuando hablamos de XAI es la de por qué son importantes las explicaciones. Aunque esta pregunta puede parecer, a priori, innecesaria dada la familiaridad que todos sentimos al hablar de explicaciones, se trata de una cuestión que es importante abordar pues su adecuada justificación es la única forma de garantizar su inclusión en los planes y las estrategias de IA y digitalización. En el contexto de la IA la relevancia de las explicaciones suele depender del contexto y el tipo de aplicación de IA que se utilice (Markus et al., 2021). Sin embargo, existen distintas razones comunes que justifican la importancia de la XAI.

Una de las principales motivaciones detrás de la XAI es la explicación de mecanismos causales. La capacidad de ser explicable garantiza que sea posible comprender la cadena de causalidades que tiene lugar dentro de la IA hasta llegar a sus resultados o decisiones finales. La explicación de causalidades en sí misma no es especialmente importante, sin embargo, sí que está estrechamente relacionada con dos aspectos de la IA que son esenciales. Por un lado, la posibilidad de explicar los mecanismos causales de la IA posibilita la resolución de problemas a nivel técnico (Athey y Imbens, 2015). En este caso la necesidad de dar explicaciones se justificaría a través del hecho de que cuando algo sale mal, la única posibilidad de poder corregir aquella cuestión técnica que ha fallado (datos o algoritmo) es a través del conocimiento tanto de los datos de los que se ha nutrido el sistema de IA, como del propio algoritmo, así como todos los pasos que se han sucedido hasta llegar a los resultados. La capacidad para rastrear todo el proceso y componentes de la IA se conoce con el nombre de trazabilidad (Hamon et al., 2020). En este sentido la trazabilidad –estrechamente ligada a la transparencia– sería uno de los requisitos para poder garantizar la explicabilidad de los algoritmos y, por tanto, poder corregir los errores que hayan tenido lugar, pero desde un punto de vista técnico.

Por otro lado, la explicación de causalidades es esencial para poder garantizar la rendición de cuentas (Doshi-Velez et al., 2017; Raji et al., 2020). La rendición de cuentas es uno de los principios básicos de los Estados de Derecho y las democracias liberales modernas (Linz y Miley, 2014) ya que es el principal mecanismo que existe para contestar las decisiones tomadas (Schedler, 1999). La rendición de cuentas es «la antítesis del poder monolítico» e involucra «el derecho (de la ciudadanía) a recibir información y la obligación (de los gobernantes y poderes políticos) de divulgar todos los datos necesarios» (Schedler, 2004), se trata, por tanto, de la justificación del

[2] Traducción del texto original en inglés «*Explainability refers to making intelligible and providing insight into the outcome of AI systems. The explainability of AI systems also refers to the understandability of the input, output and behaviour of each algorithmic building block and how it contributes to the outcome of the systems. Thus, explainability is closely related to transparency, as outcomes and sub-processes leading to outcomes should be understandable and traceable, appropriate to the use context*».

ejercicio del poder. En el caso de los sistemas de IA, al tratarse de máquinas, los procesos de rendición de cuentas pueden y suelen complicarse. En este caso, la XAI es esencial para poder determinar qué parte del proceso ha fallado y, de ese modo, poder dirimir responsabilidades a nivel social, político y/o jurídico. Volvamos al caso del vehículo autónomo. Si un sistema de estas características causase algún tipo de daño, o incluso la muerte, de un peatón que cumple todas las normas de circulación ¿quién sería el responsable de tal accidente? En estos casos los procesos de rendición de cuentas solo son viables si se cumple el principio de explicabilidad que garantiza que se pueda hallar el fallo que ha conducido a la toma de una decisión fatal. Casos de este tipo ponen de relieve que, para ciertos supuestos, la XAI debe ser un requisito legal que tienen que cumplir ciertas aplicaciones de IA, tal y como veremos más adelante.

Otro de los principales motivos que se suelen utilizar para señalar la relevancia de la XAI es el papel que juega este elemento en la generación de confianza en los sistemas de IA (Kim, 2015; Ribeiro et al., 2016; Doshi-Velez et al., 2017a; Lipton, 2018). Aunque todavía queda mucho trabajo por hacer para determinar cuál es la relación entre la XAI y la confianza en la IA, es razonable pensar que aquellos sistemas que sean comprensibles generarán mayores niveles de confianza en la ciudadanía. Disponer de un nivel alto de confianza en los sistemas de IA es importante pues de ello dependerá, en gran parte, el adecuado funcionamiento de los mismos. Al contrario, bajos niveles de confianza pueden conllevar trabas e incluso rechazo al uso de estas nuevas tecnologías y, por tanto, dificultar la gobernanza. Si bien, también es cierto que la confianza es un concepto complejo de medir y estudiar y que con frecuencia depende de cuestiones menos racionales y/o evidentes de lo que se podría pensar en un primer momento. La generación de confianza suele estar ligada a procesos heurísticos y prejuicios existentes en la mente humana (Güemes, 2016) como podrían ser en el caso de la IA: los juicios prexistentes acerca de la institución que adopta el sistema, el área de aplicación, etc. De hecho, en esta misma línea, se ha demostrado que hay cosas en los que la generación de confianza en la IA no crece con las explicaciones más comprensibles, sino justamente, al contrario. Ehsan y Riedl (2021) publicaron un estudio en el que las mayores cotas de confianza eran obtenidas a través de una explicación analítica del sistema de IA. Una explicación analítica es aquella que se formula usando lenguaje matemático y computacional, por tanto, a priori, se trataba de la explicación más difícil de entender. Estudios de este tipo señalan que la relación entre XAI y confianza probablemente sea más compleja de lo que podríamos pensar y que podría estar mediada por otro tipo de asociaciones como el grado de competencia percibida en el sistema de IA y las instituciones que lo adoptan.

Otras razones que se suelen alegar para poner en valor la XAI son la transferibilidad, el derecho a la información y la mejora en la toma de decisiones éticas. La XAI es importante para la transferibilidad en la medida que solo las tecnologías de IA cuyo comportamiento se conoce y comprende pueden ser reproducidas con éxito en entornos diferentes al original (Lipton, 2018). También es importante la XAI para garantizar el cumplimiento del derecho de la ciudadanía a estar informada (Kim, 2015; Lipton, 2018). Siguiendo esta misma lógica, dentro del campo de la IA hay quienes abogan por el derecho a la explicación (Selbst y Powles, 2017; Bryce Goodman y Flaxman, 2017; Kaminski, 2019) para convertir la XAI en una obligación en términos legales. Sin embargo, también hay quienes consideran que tal derecho no existe (Wachter et al., 2017a). Finalmente, la XAI juega un papel decisivo en el establecimiento de principios éticos en las tecnologías de IA. Solo si estas tecnologías son explicables y, por tanto, conocemos su funcionamiento, podremos

evaluar sus resultados y decisiones, y, en el caso de que sea necesario, establecer otros criterios y principios éticos que guíen su uso.

4.1. ¿Cuándo hay que dar explicaciones?

Las explicaciones están presentes en todos los ámbitos de nuestra vida y suelen demandarse cuando tiene lugar algún tipo de suceso o acto que no logramos comprender parcial o totalmente. Esto no significa que las personas demandemos explicaciones para todas aquellas cosas que no comprendemos o que estas siempre sean la mejor solución, principalmente porque dar explicaciones es un proceso costoso (Doshi-Velez et al., 2017b). Es decir, procurar explicaciones es un proceso que lleva tiempo y esfuerzo, algo de lo que no siempre disponemos. Por este motivo, no todas las tecnologías de IA transparentes tienen por qué ser al mismo tiempo explicables. Dado los costos que supone proporcionar explicaciones la obligatoriedad de las mismas se ha visto limitada a aquellos usos de la IA que la Comisión Europea ha calificado como de riesgo alto (Comisión Europea, 2021a).

En este sentido, según la propuesta de regulación realizada por la UE, la XAI será obligatoria por ley cuando se usen tecnologías de IA en los siguientes supuestos (Comisión Europea, 2021b):

1. Identificación biométrica y categorización de personas físicas.
2. Gestión y funcionamiento de infraestructuras esenciales.
3. Educación y formación profesional.
4. Empleo, gestión de los trabajadores y acceso al autoempleo.
5. Acceso y disfrute de servicios públicos y privados esenciales y sus beneficios.
6. Asuntos relacionados con la aplicación de la ley.
7. Gestión de la migración, el asilo y el control fronterizo.
8. Administración de justicia y procesos democráticos.

Por ley, las ocho situaciones expuestas anteriormente serán los únicos casos que, a priori, necesitarán garantizar el uso de sistemas de IA que sean explicables. En estos supuestos, contemplados por la regulación europea, la explicabilidad es obligatoria por el carácter excepcional de las mismas. Estas particularidades conllevan que los sistemas de IA que se apliquen cumplan con los más altos estándares de calidad para que, en el caso de que algo saliese mal, se pudiese garantizar la rendición de cuentas, solucionar problemas, respetar los derechos humanos, etc.

Sin embargo, esto no significa que solo en estas situaciones sea importante cumplir el principio de explicabilidad. Tal y como señalábamos al tratar la importancia de las explicaciones, podría suceder que, aún sin ser una obligación en términos legales, hubiese situaciones en las que resultase beneficioso proporcionar explicaciones acerca del funcionamiento de las tecnologías de IA, por ejemplo, para generar confianza. Esto sería interesante, principalmente, en las primeras etapas del proceso de implementación de la IA en el sector público ya que coincidiría con los momentos de mayor incertidumbre y desconfianza por parte de la ciudadanía respecto al uso de estas tecnologías.

5. Tipos de explicaciones

Como hemos dicho anteriormente, normalmente, los seres humanos demandamos explicaciones cuando sucede algo que no comprendemos (Leake, 1992; Lipton, 2018) y esto sucede en muchos ámbitos distintos de la vida, desde el privado hasta el científico. Por tanto, es importante destacar qué es exactamente una explicación y cuáles son los requisitos que esta debe cumplir para realizar su función de forma satisfactoria según el contexto en el que nos situemos. No es lo mismo tener que explicarle a tu pareja por qué no sacaste la basura tal y como te pidió, que explicar cuáles son los mecanismos internos de una estrella y su proceso evolutivo (Díez et al., 2011; Haslanger, 2016). Las explicaciones que se circunscriben al ámbito familiar o sentimental, es decir, a nuestra vida cotidiana, no suelen ser muy restrictivas en cuanto a su forma y contenido. De hecho, éstas suelen ser altamente contingentes y dependientes de las particularidades emocionales y cognitivas de nuestro interlocutor. Además, estas explicaciones no tienen pretensiones de universalidad, es decir, de ser válidas en todo tipo de contextos, sino que más bien responden a la necesidad del momento.

Sin embargo, las explicaciones científicas, o aquellas que aspiran a ser consideradas como tal, sí que deben cumplir una serie de normas estrictas. Estas explicaciones tienen que ser lo más objetivas posibles, es decir, no deben ser dependientes del contexto en el que se formulan (De Zárate-Alcarazo, 2020). Por tanto, las explicaciones científicas, al contrario que otro tipo de explicaciones, sí que tienen pretensión de universalidad pues deben de servir de herramienta aclaratoria para todas las personas que componen una misma comunidad epistémica y, al mismo tiempo, posibilitadoras de la creación de nuevo conocimiento (Kuhn, 1989). Esto implica que las explicaciones científicas también deben generar un consenso más o menos sólido. Pensemos en las explicaciones que se dan a los alumnos de física a la hora de comprender las leyes de la termodinámica. Aunque las explicaciones pueden incorporar ciertos matices y añadiduras propias del «explicante» (persona que explica), la mayor parte de su contenido es compartido por toda la comunidad científica y se puede encontrar incluso en los manuales sobre la materia sin que ello suponga grandes diferencias. Por otro lado, las explicaciones científicas, tienen que ser lo más precisas posibles, es decir, no admiten vaguedad en los conceptos que la componen (Friedman, 1974).

Para saber qué tipo de explicaciones es necesario proporcionar en el campo de la IA el primer paso es saber quién es el destinatario del mensaje (explicación). En este sentido, podríamos diferenciar dos tipos de destinatarios: los expertos y la ciudadanía[3]. Los expertos serían todas aquellas personas que conocen en profundidad el campo de la IA (principalmente pertenecientes al ámbito de la tecnología, pero también pueden proceder de áreas de conocimiento mixtas) y que su acercamiento a la IA tiene lugar debido a conocimiento de la materia. En este caso las explicaciones, pueden ser el propio código que se ha usado para programar el algoritmo sin necesidad de que se realice ningún tipo de filtro o ejercicio de interpretación. Los expertos pueden interactuar con la IA de forma directa porque conocen su funcionamiento y el lenguaje de programación en el que trabaja. Además, esta forma de presentar los algoritmos es la necesaria para poder solucionar, a nivel

[3] Es importante tener en cuenta que los primeros, en función del contexto, pueden enmarcarse dentro del segundo grupo, es decir, una persona experta en IA es también una ciudadana y, por tanto, no todos sus encuentros con este grupo de tecnologías los hará en calidad de experta, sino más bien como ciudadana.

técnico, los problemas que pueda presentar un sistema concreto de IA o llevar a cabo los cambios que sean oportunos (Markus et al., 2021).

En cambio, la ciudadanía (o usuarios no expertos) no puede interactuar de forma directa con los sistemas de IA en la medida que entrar en contacto con el código y los datos en bruto no supondría para ellos ningún tipo de evento explicatorio. En estos casos, que son la mayoría, es necesario realizar un ejercicio de interpretación entre el lenguaje computacional y el lenguaje natural (que usamos las personas humanas) para poder proporcionar explicaciones que permitan comprender como funciona un sistema de IA y por qué toma las decisiones que toma sin necesidad de ser un experto en la materia. Mientras que el primer tipo de explicaciones se circunscribe al ámbito de las explicaciones científicas, este segundo tipo de explicaciones se encuentran a medio camino entre el lenguaje coloquial y el científico, pues deben ser fácilmente comunicables y entendibles para la mayor parte de las personas, pero, al mismo tiempo, tienen que ser rigurosas y cumplir una serie de requisitos para asegurar que la función explicativa se cumple.

A su vez, las explicaciones que se dan a la ciudadanía pueden presentarse de distinto modo. Las explicaciones se componen de dos elementos: el primero hace referencia al nivel o grado de la explicación; el segundo a la forma o al objetivo de la explicación. Por un lado, el grado o nivel de explicación depende, a su vez, de dos elementos: la precisión y la completitud (Díez et al., 2011; Kulesza et al., 2013). La precisión hace referencia al grado de detalle con el que se describe el sistema o, lo que es lo mismo, el nivel de correspondencia entre la realidad de funcionamiento del sistema y la explicación. La completitud tiene que ver con qué porcentaje del sistema se explica. En este sentido, las explicaciones pueden ser locales o completas. Puede haber casos en los que la mejor opción sea explicar una parte del sistema que sea representativa de forma muy precisa, mientras que en otros casos quizás resulte más conveniente dar explicaciones de la totalidad del sistema de forma algo más vaga. Por otro lado, la forma de la explicación depende de qué se explique. En este caso encontramos dos tipos de explicaciones, las explicaciones descriptivas, que son aquellas que tratan de explicar el cómo (cómo funciona la tecnología de IA), y las explicaciones justificativas, que explican los criterios que se utilizan para alcanzar una explicación, es decir, el porqué. De este modo, parece evidente que, en el caso de la XAI no existe una única explicación que sirva para explicar todos los usos de la IA (Arya et al., 2021).

En la literatura especializada las explicaciones justificativas han acaparado más atención al considerar que éstas suelen ser el tipo de explicación que más satisface a la ciudadanía. Sin embargo, debido a la reciente emergencia de este campo de estudio aún no existe consenso sobre esta cuestión. Si bien es cierto que tradicionalmente las explicaciones justificativas han dado mejores resultados en otras áreas (Wachter et al., 2017b), aún es pronto para decir que esta misma situación se repite en el campo de la IA. Lo que es importante tener en cuenta es que el tipo de explicación necesaria, probablemente, sea altamente dependiente del contexto en el que se usó. Lo más probable es que la ciudadanía no demande el mismo tipo de explicaciones por parte de una tecnología de IA que le ayuda a hacer la declaración de la renta que de una que asiste a su médico en el diagnóstico de una enfermedad.

Finalmente, es importante tener en cuenta que dentro de las explicaciones justificativas también existen distintas formulaciones. Por ejemplo, las explicaciones contrafactuales suelen tener la siguiente forma: «Te han denegado el préstamo porque tus ingresos anuales son de 30.000€, si hubieran sido iguales o superiores a 45.000€/año se te habría concedido el préstamo» (Wachter et al., 2017b). Se trata de

explicaciones en las que el elemento explicatorio es el contrafáctico, es decir, aquello que es alternativo a lo que sucede –una opción posible en otro escenario–. Pero también existen otro tipo de explicaciones más allá de las contrafactuales, como las explicaciones basadas en datos demográficos, casos concretos, etc. Además, vale la pena señalar que las explicaciones no siempre tienen por que ser escritas. También se pueden dar explicaciones de forma visual, oral, analítica, etc. (Markus et al., 2021), aunque en estos casos probablemente entrarían en juego otras variables que no hemos analizado en este trabajo como la entonación, la calidad de las imágenes, etc.

6. Conclusiones

Tras lo expuesto en este trabajo, podemos concluir que la XAI es uno de los principios fundamentales en el campo de la ética y la gobernanza de la IA. El principio de explicabilidad es uno de los principios éticos esenciales sobre el que se sostienen el resto de los principios. Así pues, sería difícil pensar en una IA justa, que no dañe a las personas o que respete la autonomía humana, si no se garantiza previamente la explicabilidad de estos sistemas. Al mismo tiempo, la XAI no solo es importante en el plano ético y moral, sino también en el ámbito de la gobernanza en la medida que se trata de un elemento indispensable para poder asegurar la rendición de cuentas de las nuevas tecnologías y, previsiblemente, facilitar la buena gobernanza a través de la generación de confianza por parte de la ciudadanía en los algoritmos inteligentes.

Sin embargo, a pesar de la relevancia de la XAI resulta evidente que todavía se trata de un área de estudio emergente en el que no solo existen cuestiones por resolver, sino también consensos por establecer. Primero, es necesario hallar una definición de consenso para el concepto de XAI. Como se ha intentado mostrar en este trabajo, aún no existe una definición unívoca y que genere consenso dentro de la comunidad científica en torno a qué es exactamente la XAI. Esta cuestión, que podría parecer menor, es fundamental si queremos consolidar este campo de estudio. En el ámbito científico, en el que se sitúa la XAI, solo será posible generar conocimiento en la medida que la comunidad científica reconozca y comparta ciertas asunciones y premisas en torno a conceptos básicos (Kuhn, 1989; 2017). En este caso, las bases o fundamentos del edificio de la ciencia serían, entre otros, los conceptos de explicabilidad, IA, etc. Por tanto, alcanzar consensos en torno a definiciones básicas como estas es esencial si queremos que este ámbito de estudio emergente, pero potencialmente estratégico, pueda aportar enseñanzas y conocimientos útiles, que se puedan usar para mejorar la vida de las personas y que nos permitan avanzar en el estudio de la IA y el ser humano.

Otros desafíos presentes en el campo de la XAI son la falta de evidencia empírica que demuestre y explique las relaciones entre la explicabilidad y otros elementos esenciales como la rendición de cuentas, la generación de confianza, etc. Esta problemática, que se debe principalmente al hecho de que este es un campo científico aún emergente, se ve acrecentada por las particularidades del objeto de estudio. Estudiar cómo influyen las explicaciones en la rendición de cuentas o la confianza no es tan simple como probar qué explicaciones prefieren los ciudadanos, pues como se ha comentado a lo largo del trabajo, estos procesos no siempre son completamente racionales. Para poder llevar a cabo una investigación efectiva de estas relaciones es necesario tener en cuenta potenciales elementos mediadores como los prejuicios existentes en torno a la IA, la reputación de las instituciones públicas, la percepción de la competencia de los sistemas, las necesidades cognitivas de las personas, etc. Además, existen otros problemas como el quién evalúa la idoneidad de las explicaciones, cómo podemos llegar a determinar si estás cumplen o no la función para la cual se establecieron, etc., es decir, una serie de problemáticas

relacionadas con el proceso de interlocución entre instituciones públicas, desarrolladores tecnológicos y ciudadanía.

Por tanto, aunque la XAI es un área de estudio apasionante, con vocación de futuro y estrechamente ligado a la cultura de la legalidad, aún presenta muchos puntos oscuros relacionados, por un lado, con la falta de madurez de esta disciplina y, por otro lado, con las problemáticas propias del objeto de estudio. De este modo, algunos retos de cara a asentar este campo y futuras líneas de investigación para ahondar en el mismo serían la puesta en común de una definición en torno al concepto de explicabilidad; determinar qué tipos de explicaciones son más eficaces para la rendición de cuentas y la generación de confianza; determinar cómo influyen elementos heurísticos y cognitivos en estos procesos; analizar cuándo son mayores las demandas de explicaciones por parte de la ciudadanía, etc.

Bibliografía

Arya, V., Bellamy, R. K. E., Chen, Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., y Zhang, Y. (3 de marzo de 2021). One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques.

Athey, S. y Imbens, G. W. (2015) Machine-learning methods. https://arxiv.org/abs/1504.01132v1.

Berchi, Mauro. (3 de marzo de 2020). La inteligencia artificial se asoma a la justicia, pero despierta dudas éticas. *El País*. https://elpais.com/retina/2020/03/03/innovacion/1583236735_793682.html

Bryce Goodman, y Flaxman, S. (2017). European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". *AI Magazine*, *38*(3).

Cabitza, F., Campagner, y A., Ciucci, D. (23 de agosto de 2019). *New frontiers in explainable AI: Understanding the GI to interpret the GO* [Comunicación en Congreso]. International Cross-Domain Conference for Machine Learning Knowledge Extract. 27-47, https://doi.org/10.1007/978-3-030-29726-8_3.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., y Floridi, L. (2018). Artificial Intelligence and the "Good Society": the US, EU, and UK approach. *Science and Engineering Ethics*, *24*(2), 505–528.

Coeckelbergh, M. (2020a). *AI Ethics*. MIT Press.

Coeckelbergh, M. (2020b). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, *26*(4), 2051-2068.

Comisión Europea. (2018). *Artificial Intelligence for Europe*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN

Comisión Europea. (2019). *Building Trust in Human-Centric Artificial Intelligence*. https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

Comisión Europea. (2021a). *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF

Comisión Europea. (2021b). *Annex to the Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_2&format=PDF

Copp, D. (1990). Explanation and justification in ethics. *Ethics*, *100*(2), 237-258.
Criado, J. I. (2009). *Entre sueños utópicos y visiones pesimistas: Internet y las tecnologías de la información y la comunicación en la modernización de las administraciones públicas*. Instituto Nacional de la Administración Pública.
De Zárate-Alcarazo, L.O. (2 de junio de 2020). ¿Por qué confiamos en la ciencia? *El País.* https://elpais.com/ciencia/2020-06-02/por-que-confiamos-en-la-ciencia.html
De Zárate-Alcarazo, L.O. y Guevara-Gómez, A. (2021). *Inteligencia artificial e igualdad de género. Un análisis comparado entre la UE, Suecia y España*. Fundación Alternativas.
Díez, J., Khalifa, K., y Leuridan, B. (2011). General theories of explanation: buyer beware. *Synthese,* 190, 379–296.
D.J.O. (11 de octubre de 2018). Amazon construyó una inteligencia artificial para contratar empleados que discriminaba a las mujeres. *El Mundo*. https://www.elmundo.es/tecnologia/2018/10/11/5bbe3a12e5fdea0f578b467e.html.
Doshi-Velez, F., y Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
Doshi-Velez, F., Budish, R., y Kortz, M. (2017a). *The Role of Explanation in Algorithmic Trust, 9*, 1-93. [Reporte Técnico] Artificial Intelligence and Interpretability Working Group, Berkman Klein Center for Internet y Society.
Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'brien, D., Scott, K., Shieber, S., Waldo, J., Weinberger, D., Weller, A., y Wood, A. (2017b). *Accountability of AI Under the Law: The Role of Explanation*. http://robotics.sciencemag.org/content/2/6/eaan6080/.
Dusek, V. (2006). *Philosophy of technology: An introduction* (Vol. 90). Blackwell Publishing.
Ehsan, U., y Riedl, M. O. (2021). Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. *arXiv preprint arXiv:2109.12480*.
Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, *71*(1), 5-19.
Gasser, U., y Almeida, V. A. F. (2017). A Layered Model for AI Governance. *IEEE Internet Computing*, *21*(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835.
Guevara-Gómez, A., De Zárate-Alcarazo, L. O., y Criado, J. I. (2021). Feminist perspectives to artificial intelligence: Comparing the policy frames of the European Union and Spain. *Information Polity*, *26*(2), 173-192.
Güemes, C. (2016). Confianza. *Eunomia. Cultura de La Legalidad.*, *10*(0), 132-143. https://e-revistas.uc3m.es/index.php/EUNOM/article/view/3053.
Hamon, R., Junklewitz, H., y Sanchez, I. (2020). *Robustness and Explainability of Artificial Intelligence*. https://publications.jrc.ec.europa.eu/repository/handle/JRC119336.
Haslanger, S. (2016). What is a (social) structural explanation? *Philosophical Studies*, *173*(1), 113-130.
Just, N., y Latzer, M. (2017). Governance by algorithms: reality construction by algorithmic selection on the Internet. *Media, Culture and Society*, *39*(2), 238–258. https://doi.org/10.1177/0163443716643157.
Kaminski, M. E. (2019). The Right to Explanation, Explained. *Berkeley Technology Law Journal, 34*. https://heinonline.org/HOL/Page?handle=hein.journals/berktech34&id=202&div=9&collection=journals
Khatri, V., y Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, *53*(1), 148-152.
Kim, B. (2015). *Interactive and interpretable machine- learning models for human-machine collaboration*. [Tesis doctoral]. Massachusetts Institute of Technology

Kuhn, T. (1989). *¿Qué son las revoluciones científicas?* Paidos.
Kuhn, T. (2017). *La estructura de las revoluciones científicas*. Fondo de Cultura Económica.
Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., y Wong, W. K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. *IEEE Symposium on Visual Languages and Human Centric Computing*, 3-10.
Leake, D. (1992). *Evaluating Explanations: A Content Theory*. Psychology Press.
Lewis D. (1987). Causal explanation. En *Philosophical Papers Volume II*. Oxford Scholarship Online.
Lipton, C. Z. (2018). The Mythos of Model Interpretability. *Queue*, *16*(3), 1-27.
Lou, Y, Caruana, R., Gehrke, J., y Hooker, G. (2013). Accurate intelligible models with pairwise interactions [Comunicación en Congreso] Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 623-631. https://doi.org/10.1145/2487575.2487579.
Linz, J. J., y Miley, T. J. (2014). Algunas reflexiones precautorias y no ortodoxas sobre la democracia hoy. *Revista de estudios políticos*, 166, 19-43.
Margetts, H., y Dorobantu, C. (2019). Rethink government with AI. *Nature*, *568*, (7751), 163–165. https://doi.org/10.1038/d41586-019-01099-5.
Markus, A. F., Kors, J. A., y Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113.
Meijer, A., y Grimmelikhuijsen, S. (2020). Responsible and Accountable Algorithmization: How to Generate Citizen Trust in Governmental Usage of Algorithms. En R. P. y M. Schuilenburg (Ed.). *The Algorithmic Society*. Routledge.
Miller, T. (2017). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence,* 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007.
Ministerio de Asuntos Económicos y Transformación Digital. (2021). *Estrategia Nacional de Inteligencia Artificial*. https://portal.mineco.gob.es/RecursosArticulo/mineco/ministerio/ficheros/201202_ENIA_V1_0.pdf.
Orlikowski, W. J. (1992). The duality of technology: Rethinking the concept of technology in organizations. *Organization science*, *3*(3), 398-427.
Quarta, E., y Trezza, R. (2021). Coche sin conductor o ley sin conductor: ¿qué dirección tomará la ley para evitar los accidentes sistemáticos? *Revista de derecho del transporte: Terrestre, marítimo, aéreo y multimodal*, 28, 221-244.
Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., y Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing [Comunicación en Congreso]. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44. http://arxiv.org/abs/1807.00553.
Ribeiro, M. T., Singh, S., y Guestrin, C. (2016) "Why should I trust you?": Explaining the predictions of any classifier. [Comunicación en Congreso]. Proceedings of the 22nd, SIGKDD International, Conference on Knowledge Discovery and Data Mining, 1135-1144.
Schedler, A. (1999). Conceptualizing accountability. *The self-restraining state: Power and accountability in new democracies*, 14.
Schedler, A. (2004). ¿Qué es la rendición de cuentas? Instituto Federal de Acceso a la Información (IFAI), México.
Selbst, A. D., y Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, *7*(4). https://doi.org/10.1007/s13347-017-0263-5.
Taddeo, M., y Floridi, L. (2018). How AI can be a force for good. *Science*, 361, 751-752. https://doi.org/10.1126/science.aat5991.

Trezza, R. (2021). Artificial intelligence, civil law and ethics in Italy: a possible interaction. *Cultura giuridica e diritto vivente*, 9.

UNESCO. (2021). *First draft of the recommendation on the ethics of artificial intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000373434.

Véliz, C. (2020). *Privacy is Power*. Bantam Press.

Wachter, S., Mittelstadt, B., y Floridi, L. (2017a). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76-99. http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf.

Wachter, S., Mittelstadt, B., y Russell, C. (2017b). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31, 841-887.

Wachter, S., Mittelstadt, B., y Floridi, L. (2017c). Transparent, explainable, and accountable AI for robotics. *Science robotics*, *2*(6).

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Public Affairs.