Transparencia y explicabilidad de la inteligencia artificial y "compañía" (comunicación, interpretabilidad, inteligibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad…). Para qué, para quién y cuánta

Lorenzo Cotino Hueso[1]
Catedrático de Derecho Constitucional
Universidad de Valencia

Transparencia y explicabilidad de la inteligencia artificial y "compañía" (comunicación, interpretabilidad, inteligibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad…). Para qué, para quién y cuánta

[1] Realizado en el marco de los proyectos MICINN Retos "Derechos y garantías frente a las decisiones automatizadas..." (RTI2018-097172-B-C21), proyecto "Derecho, Cambio Climático y Big Data", Universidad Católica de Colombia, "Algoritmic law" (Prometeo/2021/009, 2021-24) Generalitat Valenciana" y estancia de personal investigador en empresa Generalitat Valenciana (AEST/2021/012).

La transparencia, un concepto polisémico, evocador y esencialmente instrumental, ahora aplicado al ámbito de la inteligencia artificial

Hace veinte años, afirmé las notas del emergente concepto de transparencia[2]. Y aunque ahora se trate de un contexto muy diferente, en modo alguno está de más recordar las mismas.

1. Así, en primer lugar, la "transparencia" está caracterizada por una "marcada polisemia, bajo la medida que le recubre un conjunto de significaciones complejas. En buena medida este estudio está destinado a recoger todas estas acepciones.

2. En segundo lugar, cabe apuntar que la "transparencia" viene cargada de connotaciones, en principio positivas, un simbolismo asociado a lo que puede ser conocido y comprendido, por contraposición a lo cerrado, misterioso, inaccesible o inexplicable. En ocasiones se considera la transparencia un mito, siempre con expectativas de crecer, quedando por ello por encima de ser un mero principio político, jurídico u organizativo[3]. Más allá de sus múltiples concreciones, por amplias que sean, Schram subraya que hay que tener en cuenta que la "transparencia tiene también una mentalidad de apertura"[4]. En el caso de la IA sin duda se da este supuesto tanto en las declaraciones normativas y principiales de la transparencia, como su asunción y demanda por la sociedad civil. De hecho, en ocasiones parece que estas reclamaciones de transparencia parecen olvidar las finalidades concretas y utilidad efectiva de la transparencia algorítmica que, en muchos casos, si se me permite, es bastante prosaica y tecnológica en el contexto de relaciones privadas entre proveedores o desarrolladores y las empresas que contratan sus productos de IA.

3. A la hora de fijar el significado de la transparencia, no contribuye en demasía que sea una noción empleada en muchas disciplinas y en muy diferentes contextos[5]. Decía entonces que en general puede hablarse de su uso tanto en la Ciencia política y de la Administración, la Economía y el Derecho, como en ámbitos de la Teoría de la Información, Teorías de los juegos, Teoría del desarrollo, etc.[6] Y ahora esencialmente la transparencia despliega todos sus efectos en el contexto de una tecnología disruptiva. Ya señalaba que dentro de estos ámbitos había que discernir entre muchas disciplinas, y en este sentido, además de a todas las ciencias vinculadas a la inteligencia artificial, la transparencia interesa al Derecho (constitucional, privado, administrativo,

---

[2] Cotino Hueso, L., *Teoría y realidad de la transparencia pública en Europa*, Premio Alcubilla, INAP, 2003, págs. 14 y ss. Acceso en https://www.researchgate.net/publication/349493901_Teoria_y_realidad_de_la_transparencia_publica_en_la_Union_Europea

[3] Inicialmente, J. Chevalier, "Le mythe de la transparence administrative", cit. Recientemente, J. Rideau, *La transparence dans l'Union européenne : mythe ou principe juridique*, L.G.D.J, Paris, 1998.

[4] F. Schram, "Debating Transparency: A Challenge for the Belgian Presidency", en Deckmyn, Veerle (ed.), *Increasing Transparency in the European Union?*, Instituto Europeo de Administración Pública, Maastricht, Países Bajos, 2002, págs. 33-90, pág. 35, en cursiva en el original.

[5] Aspecto que, entre otros, lo recuerda J. Waincymer, "Transparency of Dispute Settlement within the World Trade Organization", Centennial Symposium: An Australian Retrospective Article International Economic Law en *Melbourne University Law Review*, Diciembre de 2000, págs. 797-838, págs. 804 y ss.

[6] En este sentido, W. B. Mock, "An Interdisciplinary Introduction to Legal Transparency: A Tool for Rational Development", en *Dickinson Journal of International Law*, 2000, págs. 293-304, en pág. 294 y a lo largo de todo su estudio con relación a dichas disciplinas.

internacional, etc.) o a otras como la Contabilidad, la Economía política,, las Teorías de la Organización y la Ciencia de la Administración, y un largo etcétera de ramas en donde la "transparencia" adquiere variados matices. Y recordaba que la transparencia se utiliza en contextos diferentes relativos a la corrupción, desarme y control de armamento, desarrollo económico, protección ambiental, mercados financieros, gobernanza, organizaciones internacionales, contabilidad, Economía política, regulación de los mercados, comercio y comercio internacional[7], etc. Y, ahora en nuestro caso, respecto de la inteligencia artificial se exige transparencia respecto de contextos muy diferenciados, como al gobierno y al sector público, organizaciones internacionales, grandes plataformas y pequeñas empresas privadas, burocracias privadas, etc.[8].

4. En cuarto lugar, y quizá como carácter más importante, cabe tener en cuenta que la transparencia es un concepto metafórico e instrumental y aglutinador de toda una serie de contenidos que se ordenan según las finalidades perseguidas en cada contexto determinado. Como afirmaba Arena respecto de la transparencia administrativa, la "transparencia" es una gran metáfora, pues la transparencia en sí misma es solamente una propiedad física de los cuerpos. Cuando utilizamos la expresión "transparencia administrativa" estamos en realidad expresando una cierta idea de cómo *debe ser* . Y claro, la determinación del contenido de ese "deber ser" se hace depender de las finalidades. La transparencia es, pues, una cualidad instrumental en beneficio de unos valores, normas o finalidades. Y dicho carácter instrumental ha de ser ya subrayado desde un inicio. En consecuencia, el contenido de dicha "transparencia" se hace depender de las finalidades por las que la estructura se quiere que sea transparente. En este sentido no puede obviarse que el concepto es, sobre todo, un aglutinador, una noción que "acumula acepciones"[9], "un conjunto de institutos y de normas"[10] que quedan a su amparo.

Es de ahí que haya de subrayarse el carácter objetivo de este concepto, frente a una concepción subjetiva, más propia de los derechos que en el marco de la transparencia puedan reconocerse[11]. Ello no excluye, obviamente, que la transparencia se configure como un derecho subjetivo. Lo mismo sucede en el ámbito de la IA, pues la transparencia tiene un contenido objetivo variable, además de que el ordenamiento jurídico la reconoce como una obligación para diversos sujetos y, al mismo tiempo, forma parte de diversos derechos subjetivos que reconocen diversos contenidos de la transparencia (derecho de acceso a la información pública, debido proceso, derecho frente actuaciones automatizadas, derecho de acceso de protección de datos, derechos del consumidor, garantías de quienes usan sistemas de IA frente a proveedores o desarrolladores). Asimismo, la transparencia se configura como atribuciones, facultades o

---

[7] Así, W. B. Mock, "On the centrality of information law: a rational choice discussion of information law and transparency", en *John Marshall Journal of Computer and Information Law*, Verano de 1999, págs. 1069-1100, en concreto, págs. 1078-1079.

[8] *Ibidem*, pág. 1079.

[9] P. Dyrberg, "El acceso público a los documentos y las autoridades comunitarias", en *Revista de Derecho Comunitario Europeo*, nº 2, vol I, julio/diciembre de 1997, págs. 377-411, pág. 377.

[10] G. Arena, "Transparencia administrativa y democracia", en *Revista Vasca de Administración Pública* nº 37, 1993, págs. 9-20, pág. 9.

[11] En este sentido, aproximadamente, A Cerrillo Martínez, *La transparencia administrativa: Unión Europea y Medio Ambiente. El derecho de acceso a la documentación administrativa)*, Tirant lo Blanch, Valencia, 1998, págs. 15-18.

potestades de sujetos muy diversos que verifican, controlan, inspeccionan o supervisan sistemas de IA.

En razón de este último carácter instrumental, que es el más importante, afirmaba hace 20 años y cabe reiterar para la transparencia algorítmica, se deriva el hecho de que no sea posible considerar a la transparencia como un concepto cerrado, esto es, todos los atributos, derechos, principios que se incluyan en la transparencia obedecen al logro de una finalidad, y según sea dicha finalidad, el contexto donde deba desarrollarse y alcanzarse, se determinará su contenido. Así que espacial, temporal, contextualmente el contenido de este concepto puede variar, como lo hará, indefectiblemente si se matiza la finalidad que se quiere lograr a través del recurso a la "transparencia".

Transparencia y explicabilidad: el principio de referencia de la inteligencia artificial confiable

Mantelero, siguiendo a Jobin recuerda que el principio de transparencia algorítmica está presente en más de 84 documentos políticos sobre la IA. Así, en tales documentos se identificaron diez valores éticos clave y la transparencia era la más mencionada en 73 de los 84[12]. Considera que es uno de los principios de "fuerte implantación legal". Otro estudio esencial sobre los principios éticos de la IA señala que el 94% de los documentos exhaustivamente analizados incluyen la transparencia como uno de los principios consensuados en los documentos básicos[13]. Y resulta de especial interés porque se describen los diferentes contenidos a los que se suele aludir bajo el principio de "transparencia y explicabilidad", que se refiere de modo conjunto. Así, se señala que a su amparo un 72% de casos hacen referencia a la transparencia, un 78% a la explicabilidad. Ya bastante por debajo se incluiría que haya apertura de datos y algoritmos de código abierto (28%), obligaciones de comunicar o "notificar" que se interactúa con IA (25%) o de cuándo la IA toma una decisión sobre un individuo (19%), un 17% de documentos lo identifica con la obligación de difusión pública de Informes periódicos con información relevante, un 11% con un "derecho a la información", especialmente de los afectados y finalmente sólo un 3% con obligaciones de contratación pública abierta cuando se use por el sector público.

---

[12] Se señala que la transparencia aparece en 73/84; no maleficencia 60/84; responsabilidad 60/84; privacidad 47/84; beneficencia 41/84; libertad y autonomía 34/84; confianza 28/84; sostenibilidad 14/84; dignidad 13/84, y solidaridad 6/84. Jobin, A, Ienca, M, Vayena, E., "The Global Landscape of AI Ethics Guidelines", en *Nature Machine Intelligence* 1, 2019, págs. 389–399, sigo por A. Mantelero, *Beyond Data. Human Rights, Ethical and Social Impact Assessment in AI*, Information Technology and Law Series, IT&LAW 36, 2022, pág. 98.

Remite también a Hagendorff, T. "The Ethics of AI Ethics: An Evaluation of Guidelines", nº. 30 *Minds and Machines* 99., pág. 102, en donde la transparencia también era uno de los principios mencionados en más del 50% de documentos.

[13] El estudio de campo y síntesis de referencia sobre la ética de la IA, J. Fjeld, et. al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI", Berkman Klein Center for Internet & Society Research at Harvard University. January, 2020, pág. 4, https://dash.harvard.edu/handle/1/42160420

Como en otro lugar he analizado[14], la Declaración del Parlamento UE[15] sobre robótica es buena expresión de lo que constituyen los principios éticos esenciales de la IA. Así, además del "principio de transparencia" (nº 12)[16] indica que "este marco de orientaciones éticas debe basarse en los principios de beneficencia, no maleficencia, autonomía y justicia. Así pues, ya condensó en buena medida todos los principios en estos cinco. En 2018, el proyecto *AI4People* contabilizó 47 principios éticos proclamados internacionalmente. Y considera que hay cinco principios que sintetizan o captan el significado de 47 (Floridi *et al*, 2018: 696): beneficencia ("hacer el bien"), no maleficencia ("no hacer daño"), autonomía o acción humana ("human agency") ("respeto por la autodeterminación y elección de los individuos") y justicia ("Trato justo y equitativo para todos"). Estos cuatro principios básicos traen causa del ámbito de la biomedicina desde los años 2001 (Beauchamp y Childress, 2001). El principio de rendición de cuentas, explicabilidad y transparencia[17] se añadiría a los otros cuatro principios básicos.

Este principio de transparencia y explicabilidad es la "pieza crucial que falta en el rompecabezas" pues "complementa los otros cuatro principios"[18]. Así, sin transparencia no se entiende el bien o el daño de la IA, sin transparencia no se sabe cómo actuaría la IA en lugar de nosotros (autonomía). Tampoco se puede determinar la justicia y la responsabilidad.

Transparencia y explicabilidad ¿para qué? Los muchos "caminos" los que llevan a la "Roma" de la transparencia

Son muchos *caminos* los que llevan a la *Roma* de la transparencia. La transparencia pasa a configurarse como la clave de bóveda y premisa de los principios éticos y garantías jurídicas del uso de la IA. Esencialmente la transparencia es una herramienta básica. Y de una parte es un instrumento vicarial para la garantía de toda una serie de principios democráticos y derechos

---

[14] L. Cotino Hueso, "Ética en el diseño para el desarrollo de una inteligencia artificial, robótica y big data confiables y su utilidad desde el derecho" en *Revista Catalana de Derecho Público* nº 58 (junio 2019). http://revistes.eapc.gencat.cat/index.php/rcdp/issue/view/n58 http://dx.doi.org/10.2436/rcdp.i58.2019.3303

[15] Parlamento Europeo (2017 a). *Normas de Derecho civil sobre robótica. Resolución del Parlamento Europeo*, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2103(INL)). Acceso en http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0051+0+DOC+XML+V0//ES

[16] "12. Pone de relieve el principio de transparencia, que consiste en que siempre ha de ser posible justificar cualquier decisión que se haya adoptado con ayuda de la inteligencia artificial y que pueda tener un impacto significativo sobre la vida de una o varias personas; considera que siempre debe ser posible reducir los cálculos del sistema de inteligencia artificial a una forma comprensible para los humanos; estima que los robots avanzados deberían estar equipados con una «caja negra» que registre los datos de todas las operaciones efectuadas por la máquina, incluidos, en su caso, los pasos lógicos que han conducido a la formulación de sus decisiones;".

[17] HLEG (Comisión Europea - Grupo Independiente de Expertos de Alto nivel sobre Inteligencia Artificial), *Directrices éticas para una IA fiable*, 2019, pág. 10, https://op.europa.eu/es/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

L. Floridi, *et al.,* "*AI4People* —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", *Minds and Machines* 28(4), 2018, págs. 689-707, págs. 699-700, https://doi.org/10.1007/s11023-018-9482-5

[18] HLEG, *Directrices éticas*… cit. pág. 10.

fundamentales. Así sucede especialmente respecto de la que podemos llamar "transparencia externa", esto es, de cara al público y a la ciudadanía en general, o a los afectados o interesados por el uso de la IA. Así, la transparencia algorítmica es instrumento o herramienta de[19]:

- la *accountability* y el principio democrático obligan a que se pueda controlar la actuación de programadores, funcionarios y analistas de bajo nivel que "traducen" la ley en un código, muchas veces sin capacidad técnica ni jurídica para hacerlo[20];

-respecto de los usos públicos de la IA, se incluyen tanto obligaciones de publicidad activa como el derecho de acceso a la información pública, como los son los programas, algoritmos y sistemas de IA y a los registros que deja la actuación automatizada.

-Igualmente, el régimen de protección de datos incluye fuertes obligaciones de transparencia y de derecho de acceso y suplementariamente, el nuevo "derecho" respecto de las decisiones automatizadas (art. 22 RGPD-UE) también incluye especiales obligaciones de información que han sido detalladas por las autoridades de protección de datos[21] y la doctrina[22].

- La transparencia también es esencial para poder controlar errores, discriminación y sesgo algorítmicos públicos y privados[23].

-Además, respecto del uso de la IA en el ámbito sancionador, policial y judicial[24], además, será especialmente importante la transparencia para garantizar la individualización del uso de la IA respecto del afectado y que se

---

[19] Lo sigo de mi trabajo "Derechos y garantías ante el uso público y privado de inteligencia artificial, robótica y big data", en Bauzá, Marcelo (dir.), *El Derecho de las TIC en Iberoamérica*, Obra Colectiva de FIADI (Federación Iberoamericana de Asociaciones de Derecho e Informática), La Ley- Thompson-Reuters, Montevideo, 2019, págs. 917-952, http://links.uv.es/BmO8AU7.

[20] En esta perspectiva, entre otros, A. Boix Palop, "Los algoritmos son reglamentos: la necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones", en *Revista de Derecho Público: Teoría y Método*, Vol. 1, 2020. https://doi.org/10.37417/RPD/vol_1_2020_33

[21] En la concreción de la información y garantías específicas en razón del artículo 22 RGPD, Grupo del Artículo 29, *Directrices sobre decisiones automatizadas de 6 de febrero de 2018*, https://www.aepd.es/sites/default/files/2019-12/wp251rev01-es.pdf . Destaca la concreción de la AEPD, *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción, de febrero* 2020, pág 22. https://www.aepd.es/media/guias/adecuacion-rgpd-ia.pdf De igual modo, cabe seguir especialmente a este respecto las Directrices para el sector público, Rijksoverheid, *Richtlijnen voor het toepassen van algoritmen…* cit.

[22] En particular por cuanto a la transparencia en razón de este derecho, cabe remitir al exhaustivo trabajo M. Medina Guerrero, "El derecho a conocer los algoritmos utilizados en la toma de decisiones. Aproximación desde la perspectiva del derecho fundamental a la protección de datos personales", en *Teoría y realidad constitucional*, nº 49, 2022, págs. 141-171, https://dialnet.unirioja.es/descarga/articulo/8450038.pdf

Sobre este "derecho", me remito especialmente a mi trabajo "Derechos y garantías ante el uso público y privado… cit., así como a los trabajos de A. Palma Ortigosa, su tesis doctoral ahora *Decisiones automatizadas y protección de datos personales. Especial atención a los sistemas de inteligencia artificial*, Dykinson, 2022.

[23] Al respecto, cabe seguir especialmente la tesis doctoral y los trabajos de A. Soriano Arnanz, entre otros, "Decisiones automatizadas y discriminación: aproximación y propuestas generales", *Revista General de Derecho Administrativo* (Iustel, enero 2021) nº 56, acceso, https://laadministracionaldia.inap.es/noticia.asp?id=1511706

[24] La doctrina ya es muy abundante, destacan los trabajos de S. Barona, Algoritmización del derecho y de la justicia: De la inteligencia artificial a la Smart Justice, Tirant lo Blanch, 2021 y de P. Simó, La prisión algorítmica, Tirant lo Blanch, 2022. O, con V. Magro, *Justicia cautelar e inteligencia artificial: la alternativa a los atávicos heurísticos judiciales*, J.M. Bosch Editor, 2021.

tiene en cuenta la "totalidad" de circunstancias que se requieren especialmente por cuanto a las garantías de la actuación administrativa y el debido proceso.

-El acceso a la información algorítmica es elemento esencial para poder impugnar y alegar frente a las decisiones basadas en IA, ya sean del sector público o privado.

Así pues, desde el punto de vista de la Recomendación sobre la ética de la IA UNESCO[25] (nº 37): "transparencia y la explicabilidad de los sistemas de IA suelen ser condiciones previas fundamentales para garantizar el respeto, la protección y la promoción de los derechos humanos, las libertades fundamentales y los principios éticos." Esencialmente permiten "conocer los motivos por los que se ha tomado una decisión […] y tener la posibilidad de presentar alegaciones. […] La falta de transparencia también podría mermar la posibilidad de impugnar eficazmente las decisiones.

Recuerda también Mantelero que pese a los muchos obstáculos de la transparencia de los algoritmos, la transparencia y la inteligibilidad de la IA son elementos esenciales para salvaguardar la autodeterminación individual y colectiva[26]. Y ello especialmente debe ser garantizado en el sector público. Sin duda alguna para el sector público debe hacerse un análisis específico. En el presente estudio se siguen los materiales de referencia de diversas autoridades, así como las declaraciones y manifiestos que hemos hecho desde la Red DAIA[27] o los mejores trabajos en nuestro país como los del monográfico que tuve ocasión de coordinar con Boix[28], destacando muy especialmente respecto de la transparencia los trabajos de Gutiérrez[29], Vestri[30] y Cerrillo[31], por supuesto, los que acompañan ahora a este estudio.

Pero lo cierto es que la transparencia es instrumento esencial no sólo de principios, intereses públicos y derechos fundamentales, sino que es pieza esencial para el conocimiento y comprobación del buen funcionamiento del sistema de IA por los sujetos de la cadena de valor (usuarios del sistema, importadores, distribuidores, etc.), así como todos aquellos que tienen que verificar, comprobar el mismo (autoridades, evaluadores, etc.). Como luego se concreta, ésta sería la "transparencia interna".

Y es que como recuerda Ortiz de Zárate, en buena medida, la transparencia y la explicabilidad permiten comprender resolver problemas técnicos del

---

[25] Conferencia General 41ª reunión - París, 41 C/73, 22 de noviembre de 2021. Anexo. https://unesdoc.unesco.org/ark:/48223/pf0000379920_spa

[26] A. Mantelero, *Beyond Data… cit.* pág. 31.

[27] Desde la Red Derecho Administrativo de la IA (DAIA) desde 2019 hemos subrayado elementos básicos jurídicos respecto del uso de la IA en el sector público, entre ellos las cuestiones de transparencia, que sin duda constituyen un referente. Así en las Conclusiones de Toledo de 1 de abril de 2019 y la Declaración final de Valencia de 24 de octubre de 2019.

[28] A. Boix Palop, y L. Cotino Hueso, (coords.) *Monográfico Derecho Público, derechos y transparencia ante el uso de algoritmos, inteligencia artificial y big data RGDA Iustel*, nº 50, febrero 2019. Acceso en https://www.iustel.com/v2/revistas/detalle_revista.asp?id=1

[29] Posiblemente el trabajo más riguroso, aunando además los elementos técnicos esenciales de la cuestión, Gutiérrez David M.E., Gutiérrez David, "Administraciones inteligentes y acceso al código fuente y los algoritmos públicos. Conjurando riesgos de cajas negras decisionales", en *Derecom*, nº 31, págs. 19-105, ver págs. 55-56, 2021, http://www.derecom.com/derecom/

[30] G. Vestri, "La inteligencia artificial ante el desafío de la transparencia algorítmica: Una aproximación desde la perspectiva jurídico-administrativa", *Revista Aragonesa de Administración Pública*, nº 56, Zaragoza, 2021, págs. 368-398.

[31] A. Cerrillo i Martínez, "La transparencia de los algoritmos que utilizan las administraciones públicas", *Anuario de Transparencia Local*, nº. 3, 2020, págs. 41-78.

funcionamiento del sistema, especialmente para comprender la cadena de causalidades[32]. En ocasiones, la única posibilidad de poder corregir un problema técnico del sistema o de los datos que lo alimentan 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. Se requiere la posibilidad también de rastrear todo el proceso y componentes de la IA.

De ahí que ya se trate de intereses más públicos y vinculados a valores y derechos constitucionales, como a otros intereses de consumidores o usuarios que contratan y emplean sistemas de IA, o el de los sujetos que deben evaluar, verificar, inspeccionar, auditar los sistemas de IA, "la transparencia es el elemento es clave para fomentar la confianza (Recomendación UNESCO 2021, nº 39), para la generación de confianza en los sistemas de IA[33]. Ello se encuentra en la base de las políticas de IA de la UE: nuestro objetivo es crear una cultura de "IA confiable en Europa" [34]. Esta idea se refuerza en el Plan coordinado de la UE sobre la IA: la IA "Made in Europe" y es su pieza angular y el propio título del Plan[35]. Al fin ya al cabo, se busca una "IA confiable" basada en la "ética en el diseño".[36]

Pues bien, en tanto en cuanto la transparencia es un instrumento o herramienta al servicio de los concretos intereses perseguidos, el contenido o alcance de la transparencia se hace depender en muy buena medida de tales finalidades. Y para ello habrá que estar al contexto jurídico concreto respecto del que se trate.

Un esbozo de las exigencias normativas de transparencia algorítmica

De modo muy sintético en este estudio, cabe mencionar algunas regulaciones que imponen o impulsan la transparencia algorítmica en España.

Las ya mencionadas intensas obligaciones de información y explicabilidad respecto de las sobre decisiones de IA que supongan tratamiento de datos personales (art. 22 RGPD, 9.1 Convenio 108 del Consejo de Europa, artículo 20 Lei nº 13.709, de 14 de agosto de 2018 de Brasil, artículo 20 Ley orgánica de protección de datos personales de Ecuador de 2021, entre otras).

En el contexto de la transparencia pública no puede obviarse que se consideran "información pública" los "contenidos o documentos, cualquiera que sea su formato o soporte, que obren en poder de alguno de los sujetos [públicos…] y que hayan sido elaborados o adquiridos en el ejercicio de sus

---

[32] L. Ortiz de Zárate Alcarazo, "Explicabilidad (de la inteligencia artificial)". En *Eunomía. Revista en Cultura de la Legalidad*, nº 22, 2022, págs. 328-344, pág. 334 https://doi.org/10.20318/eunomia.2022.6819

[33] *Ibídem*, pág 335 remite en esta dirección a autores como Kim, Ribeiro, Doshi-Velez, Lipton, etc.

[34] HLEG, *Directrices éticas*… cit. pág. 29.

[35] Comisión Europea, *Comunicación. Plan coordinado sobre la Inteligencia artificial* COM(2018) 795 final. Bruselas, 7 de diciembre de 2018. Acceso en http://www.ipex.eu/IPEXL-WEB/dossier/files/download/082dbcc5679fb7b40167a1b3581a006c.do

Y, en particular, su Anexo, acceso en http://data.consilium.europa.eu/doc/document/ST-15641-2018-ADD-1/es/pdf

[36] Puede seguirse un análisis de la construcción de las políticas de IA en la UE y la ética de la IA en L. Cotino Hueso, "Ética en el diseño… cit. También, un análisis de la transparencia en diversos documentos internacionales en W. Arellano Toledo, "El derecho a la transparencia algorítmica en big data e inteligencia artificial", en *RGDA Iustel*, nº 50, febrero 2019. Acceso en https://www.iustel.com/v2/revistas/detalle_revista.asp?id=1

funciones" (art. 13 Ley 19/2013 de transparencia). Así, potencialmente los algoritmos usados por el poder público son susceptibles de ser solicitados por el público, así como toda información sobre sistemas de IA, su contratación, etc.[37]

También en el ámbito de la transparencia pública, especialmente novedosa es la Ley 1/2022, de 13 de abril, de la Generalitat, de Transparencia y Buen Gobierno de la Comunitat Valenciana. Su artículo 16.1 l) impone la publicidad activa "de sistemas algorítmicos o de inteligencia artificial que tengan impacto en los procedimientos administrativos o la prestación de los servicios públicos"[38]. También cabe tener en cuenta la imposición de la publicidad activa del inventario de actividades de tratamiento en aplicación del artículo 31 de la Ley orgánica 3/2018, de 5 de diciembre, de protección de datos personales y garantía de los derechos digitales. Considero que como regla general el uso de IA que suponga un tratamiento de datos personales debe quedar reflejado en dicho inventario.

Asimismo, en el ámbito administrativo, en razón del artículo 13 Decreto 203/2021 para la AGE, es precisa una "resolución" de autorización de las decisiones íntegramente automatizadas que "establecerá medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos de las personas interesadas." Dicha resolución, expresando los recursos procedentes, se publicará en la sede de la AGE (11. 1º i).[39]. La regulación, ciertamente escasa, cabe completarla con la relativa al ámbito de la contratación pública al que luego se hace alguna referencia. Cabe remitir a otros estudios en esta obra.

Aunque no es norma jurídica, no está de más recordar que la Carta de Derechos Digitales de 2021 por cuanto su aparatado XVIII 6 c) afirma el derecho a "obtener una motivación comprensible en lenguaje natural de las decisiones que se adopten en el entorno digital, con justificación de las normas jurídicas relevantes, tecnología empleada, así como de los criterios de aplicación de las mismas al caso. El interesado tendrá derecho a que se motive o se explique la decisión administrativa cuando esta se separe del criterio propuesto por un sistema automatizado o inteligente."[40]

Igualmente cabe mencionar el muy reciente artículo 23 Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación. Respecto del uso público de "algoritmos", el mismo hace referencia a la "transparencia y rendición de cuentas, siempre que sea factible técnicamente." Y el mandado de priorizar "la transparencia en el diseño y la implementación y la capacidad de interpretación de las decisiones adoptadas por los mismos."

En el ámbito laboral, el artículo 64.4.d) Estatuto de los Trabajadores dispone que el comité de empresa, con la periodicidad que proceda en cada

---

[37] La primera resolución al respecto fue de la Comisión de Garantía del Derecho de Acceso a la Información Pública (GAIP), Resolución GAIP 200/2017, de 21 de junio. En esta obra el tema es objeto de riguroso análisis al cual remitir.

[38] l) La relación de sistemas algorítmicos o de inteligencia artificial que tengan impacto en los procedimientos administrativos o la prestación de los servicios públicos con la descripción de manera comprensible de su diseño y funcionamiento, el nivel de riesgo que implican y el punto de contacto al que poder dirigirse en cada caso, de acuerdo con los principios de transparencia y explicabilidad.

[39] habrá de incluir una "relación actualizada de las actuaciones administrativas automatizadas vinculadas a los servicios, procedimientos […] Cada una se acompañará de la descripción de su diseño y funcionamiento, los mecanismos de rendición de cuentas y transparencia, así como los datos utilizados en su configuración y aprendizaje."

[40] Me permito remitir a mi estudio de dicho apartado "Derechos ante la Administración digital y la inteligencia artificial" en L. Cotino Hueso, (editor), *La Carta de Derechos Digitales*, Tirant Lo Blanch, Valencia, 2022.

caso, tendrá derecho a conocer los algoritmos que afectan a la toma de decisiones laborales[41], objeto también de otro estudio de interés en esta monografía.

A las anteriores normas, obviamente hay que añadir el futuro Reglamento de IA de la UE que someramente se ha expuesto. Asimismo, cabría añadir la incidencia en la transparencia de algoritmos que tienen diversas normas europeas, como el más reciente Reglamento de servicios digitales que se aprueba en 2022 que incluye respecto de las más grandes plataformas exigencias de transparencia de los sistemas de información o recomendación algorítmicos que determinan lo que ven los usuarios y en especial los utilizados en la lucha contra la desinformación.

¿Cuánta transparencia?

La transparencia en sentido amplio se acaba concretando en buena medida -aunque no sólo- en el grado de acceso a toda una serie de datos, información y conocimiento. Y el alcance y la disposición de esta información que exige la transparencia y en su caso la explicabilidad depende de diversos factores.

No es sencillo determinar el alcance o contenido de la transparencia algorítmica. A este respecto, se señala desde ISO que un sistema no transparente genera problemas de equidad, seguridad y responsabilidad[42]. Y además se dificulta el control de calidad. Por el contrario, mucha transparencia puede resultar excesiva e incluso contraproducente para conocer el sistema de IA. Y, sobre todo, mucha transparencia puede generar problemas de privacidad, seguridad, confidencialidad y propiedad intelectual. Encontrar el balance adecuado para las finalidades que se persigan es la clave.

Según las Directrices para el sector público de Países Bajos[43], el grado de exigencia de la explicabilidad y la transparencia depende de (1) el impacto del algoritmo en la decisión, el resultado y el ciudadano; (2) el grado de autonomía en la toma de decisiones (es decir, hasta qué punto se garantiza la participación humana); y (3) el tipo y la complejidad del algoritmo.

Así, se pueden afirmar como principios o planteamientos generales que a más impacto o potencialidad de daño de un sistema de IA en derechos e intereses, es precisa más transparencia. Respecto de derechos se afirma que cuanto mayor sea el impacto de los análisis de datos en los ciudadanos y que puedan conducir a conclusiones o decisiones, más importante será la transparencia. Se ha enunciado específicamente para el sector público “Cuanto mayor es el impacto, más importante es la transparencia”, “Cuanto mayor sea el impacto de los análisis de datos en los ciudadanos, más importante será la

---

[41] “d) Ser informado por la empresa de los parámetros, reglas e instrucciones en los que se basan los algoritmos o sistemas de inteligencia artificial que afectan a la toma de decisiones que pueden incidir en las condiciones de trabajo, el acceso y mantenimiento del empleo, incluida la elaboración de perfiles.” A este respecto, Ministerio de Trabajo y Economía Social, *Información algorítmica en el ámbito laboral. Guía práctica y herramienta sobre la obligación empresarial de información sobre el uso de algoritmos en el ámbito laboral*, Gobierno de España, Mayo 2022, https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/trabajo14/Documents/2022/100622-Guia_algoritmos.pdf

[42] Se sigue aquí el proceso de elaboración de normas ISO sobre IA en 2022, en concreto Inteligencia artificial — Seguridad funcional y sistemas de IA (Artificial intelligence — Functional safety and AI systems), el apartado 8.3 se dedica al grado de transparencia y explicabilidad.

[43] Directrices para el sector público, Rijksoverheid, *Richtlijnen voor het toepassen van algoritmen*… cit.

transparencia. Esto se aplica ciertamente en los casos en que tales análisis (pueden) conducir a conclusiones o decisiones." En todo caso, este principio puede considerarse válido en todos los contextos (públicos, privados, para la transparencia interna o externa, etc.). Una buena muestra es la propia regulación de protección de datos respecto de las decisiones automatizadas (además de las garantías generales de protección de datos, el artículo 22 RGPD consagra garantías añadidas respecto de decisiones solo automatizadas y relevantes).

Se afirma también que es necesaria más transparencia a menor intervención, control o supervisión humana del sistema de IA. Se ha señalado que la transparencia algorítmica puede ser graduada en razón de su correlación con la intervención humana. Esto significa que si un servicio público recurre a la transparencia en términos generales con respecto a sus análisis de datos, pero no divulga ciertos aspectos más detallados con vistas a los intereses de la investigación, por ejemplo, debe al menos compensar esa conducta con una supervisión interna y externa suficiente de esos aspectos[44]. Y de hecho, puede considerarse el elemento contrario, a mayor transparencia, puede relajarse la exigencia de otro tipo de garantías. En este sentido cabe recordar el sistema SYRI -también de aquél país- que fue declarado contrario a diferentes derechos fundamentales esencialmente por no ser transparente. La sentencia de 5 de febrero de 2020 del Tribunal de Distrito de la Haya (C / 09/550982 / HA ZA 18-388)[45] entre otros elementos, consideró la vulneración de derechos como el de igualdad o debido proceso, precisamente porque no se daba suficiente transparencia.[46]

Y también cabe aceptar como principio que a mayor opacidad o complejidad del sistema de IA, mayor transparencia y, como se verá, mayor explicabilidad y otros contenidos del genérico de la transparencia serán requeridas. Me remito a otros estudios en los que intento concretar de manera exhaustiva los concretos datos, información y conocimiento que se derivan de las obligaciones de transparencia algorítmica.

La "transparencia" como concepto general inclusivo y la miríada de nociones que orbitan a su alrededor

Apelamos por lo general a la transparencia algorítmica o de la inteligencia artificial. En ocasiones conjuntamente a la transparencia y explicabilidad. No obstante, hay ciertamente una nube de conceptos a veces indistintos, en ocasiones íntimamente relacionados y, cuanto menos afines o en la órbita de la transparencia que bien merece tener en cuenta.

Así, cabe mencionar entre otros, la trazabilidad, explicabilidad, interpretabilidad, comprensibilidad, inteligibilidad, legibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad, descomponibilidad, verificabilidad,

---

[44] *Ibídem*, pág. 29.

[45] https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878

[46] Al respecto, 'SyRI, ¿a quién sanciono?' Garantías frente al uso de inteligencia artificial y decisiones automatizadas en el sector público y la sentencia holandesa de febrero de 2020, en *La Ley Privacidad , Wolters Kluwer* nº 4, mayo 2020. https://diariolaley.laleynext.es/Content/Documento.aspx?params=H4sIAAAAAAAAEAMtMSbF1CTEAAmMDSwNjM7Wy1KLizPw8WyMDIwMDEyNzkEBmWqVLfnJIZUGqbUlRaSoApoQJizQAAAA=WKE y "Hacia la transparencia 4.0: el uso de la inteligencia artificial y big data para la lucha contra el fraude y la corrupción y las (muchas) exigencias constitucionales", en Carles Ramió (coord.), *Repensando la administración digital y la innovación pública*, Instituto Nacional de Administración Pública (INAP), Madrid, 2021. https://links.uv.es/FUW2pz6

replicabilidad, comunicación, código abierto, así como referencias a la transparencia técnica en sentido estricto, o la importante diferencia que se ha seguido entre transparencia externa e interna.

A continuación se profundiza en los conceptos básicos de transparencia y explicabilidad, así como la interpretabilidad. Ello acompañado de las nociones que orbitan sobre estos conceptos. Luego hace referencia a las íntimas conexiones entre estos tres conceptos. De igual modo, se determinan los elementos básicos para lograr la transparencia y explicabilidad así como se detalla el alcance o contenido que han de tener.

Para el NIST la transparencia sería un "principio rector" [47] Mientras que la "Explicabilidad" y la interpretatiblidad serían "características sociotécnicas" [48]. Según el NIST la transparencia busca remediar un desequilibrio informativo común entre los operadores de sistemas de IA y los consumidores del sistema de IA. La transparencia refleja la medida en que la información está disponible para un usuario cuando interactúa con un sistema de IA. Se afirma que a falta de transparencia, los usuarios tienen que adivinar estos factores y puede hacer suposiciones injustificadas y poco fiables sobre la procedencia del modelo[49].

En la estandarización de la IA la transparencia sería la propiedad de un sistema que aporta información sobre los procesos internos de un sistema de IA y se pone a disposición de las partes interesadas pertinentes[50].

El HLEG acude a un concepto inclusivo[51], *Transparencia = "Incluidas la trazabilidad, la explicabilidad y la comunicación".* Su check list o lista de control de "Transparencia" incluye la t*razabilidad, la explicabilidad y la comunicación.*

Como a continuación se expone, la transparencia en sentido más estricto o técnico obliga acudir a nociones de interpretabilidad y comprensibilidad del sistema de IA de ahí la relación de la transparencia con la explicabilidad se hacen más intensas.

La auditabilidad y con ella la testabilidad, comprobabilidad, verificabilidad y replicabilidad son elementos que pueden permitir la transparencia y explicabilidad. Asimismo, la transparencia incluye también la notificación o comunicación de la existencia de interacciones o de decisiones de sistemas IA. De igual modo, bajo el genérico paraguas de la transparencia se incluyen la exigencia de informaciones e informes periódicos, sistemas IA de código abierto, la contratación pública abierta o, entre otros, la difusión activa y transparencia de las evaluaciones de impacto que se realizan.

Transparencia "técnica", transparencia intrínseca, interpretabilidad, comprensibilidad, descomponibilidad y simulabilidad del sistema de inteligencia artificial

Distingue el Gobierno de Países Bajos en su referente sobre el estudio de impacto de algoritmos la transparencia técnica como "la visión del método algorítmico utilizado (árbol de decisión, red neuronal), el código fuente, cómo se entrena el algoritmo, así como los datos utilizados, las variables de entrada, los

---

[47] NIST, *AI Risk Management Framework… cit.* apartado 5.3.3, págs. 12-13.

[48] *Ibídem*, apartados 5.2.2. págs. 10 y ss.

[49] *Ibídem* págs. 12 y ss., en concreto, pág. 13.

[50] Así en el proceso de elaboración de normas ISO sobre IA en 2022, en concreto Artificial intelligence — Functional safety and AI systems, el apartado 8.3, nº 672- 694.

[51] HLEG, *Directrices éticas*… cit. págs. 18 y 22.

parámetros y umbrales utilizados, etc." [52] En su excelente trabajo, Gutiérrez David, recuerda que desde un plano estrictamente técnico, no hay un concepto unívoco de transparencia algorítmica.[53] Diferencia entre la transparencia intrínseca del modelo, la transparencia algorítmica en estricto sentido y, en los casos de modelos opacos, los métodos o técnicas de interpretabilidad en mayor o menor medida hacen más comprensible y transparente al humano el funcionamiento[54]. Sobre esta base, identifica la "transparencia técnica" como una característica *pasiva* del modelo de IA, es decir el grado de comprensibilidad y la interpretabilidad de un modelo específico por sí mismo, esto es, si el funcionamiento global del modelo, de sus componentes individuales y de su algoritmo de aprendizaje resultan inteligibles o comprensibles para un humano.

Así las cosas, podría diferenciarse entre los modelos técnicamente transparentes, que son interpretables por diseño, mientras que habría otros sistemas de IA que no son transparentes, pero pueden llegar a ser explicables mediante distintas técnicas. En consecuencia, la transparencia técnica se vincula inextricablemente con la interpretabilidad y comprensibilidad del modelo de IA.

La transparencia o interpretabilidad dependerá de un adecuado equilibrio entre la *simulabilidad*, la *descomponibilidad* y la *transparencia* algorítmica[55]:

Simulabilidad supone que el conjunto del modelo puede ser reproducido o replicado por un humano en un tiempo razonable a partir de los datos y parámetros del modelo mediante los cálculos necesarios para generar la predicción

Descomponibilidad. Gutiérrez supone que los componentes del modelo (inputs, parámetros y cálculo) de pueden descomponer y permiten una explicación intuitiva[56].

Para lograr la interpretabilidad de los resultados, los métodos pueden ser variados. Los modelos agnósticos se basan en analizar las variables de entrada y los resultados, sin acceder a las operaciones internas del modelo y valen para cualquier modelo de *machine learning*, sea o no opaco. En cambio, los modelos específicos sí que analizan algunas partes internas del modelo, como la interpretación de coeficientes.

---

[52] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (Ministerio del Interior y Relaciones del Reino), *Impact Assessment. Mensenrechten en Algoritmes (Evaluación de impacto. Derechos humanos y algoritmos),* julio, 2021, apartado, 2B.4 https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2021/02/25/impact-assessment-mensenrechten-en-algoritmes/IAMA.pdf

[53] M.E., Gutiérrez David, "Administraciones inteligentes y acceso al código … cit. ver págs. 55-56.

[54] *Ibídem* pág. 57.

[55] NIST, *AI Risk Management Framework… cit. apartado* 5.2.2., "la interpretatibilidad se puede diferenciar en simulabilidad, descomponibilidad y transparencia algorítmica."

[56] Señala la autora que, por ejemplo, el nodo en un árbol de decisión puede corresponderse con una descripción en lenguaje natural (todos los pacientes con presión distólica superior o igual a 150). De la misma manera, los parámetros en modelo lineal representan la relación o coeficiente entre cada variable predictora y la predicción. M.E., Gutiérrez David, "Administraciones inteligentes… cit. Remite la autora a trabajos de referencia de A. Barredo y otros, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". Information Fusion, vol. 58, en concreto pág. 90; Z. C. Lipton, "The Mythos of Model Interpretability". *ACM Queue,* vol. 16, núm. 3 (mayo-junio), 2018, pág. 12 o el ya referido ICO y Alan Turing Institute, *Explaining decisions made with Artificial Intelligence*, 2020, págs. 67-68 https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-ai/

Cuando el sistema de IA no es transparente técnicamente y es un modelo opaco, se intenta descubrir cómo funciona el modelo, o qué más puede decirnos el modelo. Aunque el sistema IA no permitan conocer con exactitud cómo funciona el modelo, puede aportar información relevante. Algunas técnicas limitan funcionalidades del modelo para intentar entenderlo o simplifican el modelo opaco. Otras técnicas permiten visualizar la influencia de unas y otras variables, describir los resultados según cada variable. También hay técnicas intentan explicar el modelo de aprendizaje a partir de ejemplos concretos.

Desde ISO se señala que se puede contar con información sobre el sistema; si es comprensible o al menos genera información comprensible al destinatario previsto, si los resultados son correctos, completos y reproducibles de manera consistente. Asimismo, se indica que puede haber variadas técnicas o estrategias de evaluación para juzgar la transparencia y la explicabilidad (que se abordan conjuntamente). De este modo hay que tener en cuenta la posibilidad de realizar evaluaciones empíricas del proceso; determinar cómo las características de entrada afectan la salida del modelo, revisar la salida de una red neuronal, que el humano inspeccione su estado interno de la red. También cabe acudir a enfoques como el muestreo de entrada aleatorio para explicación. Respecto de árboles de decisión se señala que también pueden alcanzar una gran complejidad. Se afirma desde ISO que aunque no se dé una total explicabilidad, se puede emplear una evaluación metódica y formalmente documentada de la interpretabilidad del modelo con respecto al riesgo.[57]

Transparencia ¿para quién? Transparencia “interna” y “externa”

Transparencia interna y externa.

El objeto o contenido de la transparencia algorítmica se hace depender en muy buena medida de la naturaleza del sujeto al que van destinados los datos, información o conocimiento sobre el sistema de IA. No en vano, ello es reflejo de la finalidad o función de la transparencia. A este respecto, el Gobierno de Países Bajos en estudio de impacto de algoritmos[58] afirma la distinción “entre la "transparencia interna" (es decir, la transparencia dentro de la organización y en beneficio de los auditores internos y externos, los supervisores, los jueces y los interesados (es decir, las personas identificadas o identificables cuyos datos personales son tratados por algoritmos) y la "transparencia externa" (es decir, la transparencia hacia el exterior, hacia el público en general). De hecho, se afirma que esta transparencia externa también se denomina “explicabilidad pública”[59].

La transparencia “interna” y el futuro Reglamento UE de inteligencia artificial como ejemplo

Hay muchos supuestos en los que el destinatario de la información sí que ha de contar con suficientes conocimientos técnicos, esto es no es el público general con conocimientos mínimos o inexistentes de lo que es un sistema IA. De este modo, será diferente el tipo de información si se trata de información

---

[57] Se menciona aquí de modo general y a título de cita científica, las referencias en el proceso ISO sobre IA en 2022, *Inteligencia artificial — Seguridad funcional y sistemas de IA*, apartado 8.3.

[58] Ministerie van Binnenlandse, *Impact Assessment… cit.* apartado, 2B.4.

[59] *Ibídem*, con remisión al *Toetsingskader Algemene Rekenkamer*, esto es, Marco de pruebas de algoritmos: primeros pasos. https://www.rekenkamer.nl/onderwerpen/algoritmes-digitaal-toetsingskader

para que el usuario o consumidor del sistema (y los técnicos que lo implementan) puedan manejar el sistema de IA correctamente, cumplir sus obligaciones y supervisarlo. En este punto se ha de tener en cuenta la naturaleza diferente de usuarios, importadores, distribuidores. Desde un ángulo bien diferente, también han de contar con suficientes conocimientos técnicos quienes tienen una labor de supervisión, evaluación de conformidad o auditoría (órganos supervisores, organismos notificados para la evaluación de conformidad, auditores internos, externos, órganos administrativos, etc.). En este punto, debe tenerse muy en cuenta que es muy posible que se acceda a información de alto grado de detalle e incluso de sensibilidad, si bien en general bajo fórmulas de cláusulas o deberes de confidencialidad.

Los organismos de normalización técnica NIST (EEUU), ISO, CEN CENELEC (UE) por lo general abordan la transparencia interna, no la transparencia dirigida al público en general[60]. Y en el caso del futuro Reglamento de IA de la UE[61], se contemplan importantes obligaciones de transparencia, trazabilidad (art. 12), "transparencia" (art. 13), gestión de calidad (art. 17), documentación (art. 11 y Anexo IV) y otras para los sistemas de alto riesgo. Estas exigencias son proyectables también voluntariamente a los sistemas de IA que no sean de alto riesgo (art. 69). La regla general es que la información y transparencia del futuro Reglamento de IA es "transparencia interna. De ahí que el artículo 13 sobre "transparencia" afirma que "sistemas de IA de alto riesgo se diseñarán y desarrollarán de un modo que garantice que funcionan con un nivel de transparencia suficiente para que los usuarios interpreten y usen correctamente su información de salida". Se exige "un tipo y un nivel de transparencia adecuados para que el usuario y el proveedor cumplan las obligaciones oportunas previstas". Así, estas obligaciones de transparencia por lo general no están concebidas con relación a los afectados por decisiones o funcionamiento de sistemas IA, sino esencialmente para los usuarios, otros agentes de la cadena de valor, así como supervisores y autoridades. Es por ello que se trata por lo general de información detallada y con alto grado de carácter técnico.

Lo mismo sucede respecto de la regulación de la documentación de gestión de calidad del sistema IA (art. 17) y las obligaciones de documentación técnica (arts. 11, 18 y especialmente en el Anexo IV). Se trata de una información y documentación muy amplia de carácter técnico que debe elaborarla el proveedor desarrollador del sistema IA (de alto riesgo) para mantenerla durante 10 años a disposición de los órganos supervisores (arts. 23, 50) y organismos notificados,

---

[60] Son variados los estándares existentes que hacen referencia a la transparencia algorítmica. Para EEUU cabe señalar los primeros trabajos NIST, *AI Risk Management Framework: Initial Draft*, 17 de marzo de 2022, https://www.nist.gov/document/ai-risk-management-framework-initial-draft

En el contexto de ISO cabe remitir ISO/IEC TR 24027 *Sesgo en sistemas de IA y toma de decisiones asistida por IA* ISO/IEC TR 24029-1 *Evaluación de la robustez de la red neuronal ISO 26000 Guía sobre responsabilidad social*). Y en desarrollo: ISO/IEC DIS 23894 *Gestión de riesgos*, ISO/IEC AWI TS12791 *Tratamiento del sesgo no deseado en la clasificación y tareas de aprendizaje automático de regresión* así como ISO/IEC AWI 12792 *Taxonomía de transparencia de los sistemas de IA.*

En el contexto de CEN CENELEC de la Unión Europea, se sigue en 2022 el proceso de adopción de las primeras normas, cabe destacar la CEN/CLC/JTC 21 - *Artificial Intelligence*.

[61] Se hace referencia a la Propuesta de Reglamento del Parlamento Europeo y del Consejo que se establecen normas armonizadas sobre la inteligencia artificial (Ley de Inteligencia Artificial).

autoridades notificantes y otros sujetos implicados en los procedimientos de evaluación de conformidad de los sistemas. Pero estos datos e información en principio no está destinada ni a los usuarios del sistema, ni a los afectados, ni al público en general o la comunidad investigadora. Sobre la misma en general hay un régimen de confidencialidad y acceso reservado (art. 64).

Hay que llamar la atención de un equívoco habitual. En el caso de la IA los consumidores o "usuarios" no son los interesados o afectados últimos, normalmente personas físicas o jurídicas a cuyos derechos impacta el uso de la IA. El "usuario"[62] de la IA es la entidad que utiliza el sistema de IA que un proveedor ha desarrollado (por ejemplo, adquiere el derecho de uso para su empresa). Pero el "usuario" no es el interesado a que afecta el uso del sistema y en su caso impacta en sus derechos (que sería por ejemplo, el trabajador o cliente de esa empresa).

Dicho lo anterior, en el futuro Reglamento IA sí se dan algunas obligaciones de transparencia "externa" que van dirigidas al público en general o a quienes pueden ser afectados o interesados por el suso de un sistema de IA. Así, en primer lugar, hay algunas obligaciones de "transparencia" del artículo 52, que son de obligada comunicación para que el humano conozca su interactuación con chatbots, sistemas de IA con reconocimiento de emociones o sistemas IA de adulteración de vídeos. Estas obligaciones de comunicación sí que van dirigidas a los humanos. En segundo lugar, el futuro Reglamento de IA crea un registro, la "Base de datos de la UE para sistemas de IA de alto riesgo independientes" y la información de esta base de datos "será accesible para el público" (art. 60.2º). En dicha base de datos tienen que registrarse algunos sistemas de alto riesgo[63] (art. 51 AIA) y todo el mundo puede acceder a información bastante básica que ahí consta.

La transparencia "externa" como garantía de valores, principios y derechos constitucionales

La cuestión varía y mucho en buena parte de los supuestos en los que la transparencia algorítmica es garantía de valores, principios y derechos constitucionales. En estos casos transparencia se traduce en datos, información o conocimientos dirigida al sujeto afectado por decisiones de IA y sus garantías específicas. O si se trata de información ofrecida para todo el público en general. No obstante, también respecto de estas finalidades de la transparencia algorítmica debe tenerse en cuenta a la comunidad de especialistas o científicos que sí que cuentan con alto conocimiento. Las Directrices o "Pautas" para el uso de IA en Países Bajos por el sector público (en adelante Directrices para el sector público)[64]o la guía de transparencia algorítmica pública del ICO para el Reino

---

[62] En este sentido cabe recordar la definición de "usuario" del futuro Reglamento de IA de la UE, artículo 3.1.4º "Usuario: toda persona física o jurídica, autoridad pública, agencia u organismo de otra índole que utilice un sistema de IA bajo su propia autoridad, salvo cuando su uso se enmarque en una actividad personal de carácter no profesional."

[63] En concreto, respecto de los sistemas de alto riesgo según el listado del Anexo III (art. 6. 2º), esto es, los otros sistemas de alto riesgo en razón del artículo 6.1º y Anexo II no tienen que registrarse en la base de datos.

[64] Ministerie van Justitie en Veiligheid (Ministerie van Justitie en Veiligheid), Rijksoverheid (Gobierno central), *Richtlijnen voor het toepassen van algoritmen door overheden en publieksvoorlichting over data-analyses* (*Pautas para la aplicación de algoritmos por parte de los gobiernos y educación pública sobre análisis de datos*), Directiva (Richtlijn), de 08-03-2021, https://www.rijksoverheid.nl/documenten/richtlijnen/2021/09/24/richtlijnen-voor-het-toepassen-

Unido[65] sí que están concebidas para la obligación general de informar a todo el público. En el caso de autoridades de protección de datos, se trata bien de la información general al público obligatoria, o bien de las obligaciones más concretas respecto de los afectados, especialmente por decisiones algorítmicas. El HLEG aunque hace referencia a "usuarios" [66] no parece discernir el tipo de sujeto del que se trata por lo general, tampoco la Recomendación UNESCO 2021.

En estos casos los datos, información o contenidos en que consiste la transparencia y explicabilidad no necesariamente han de ser de un carácter básico o elemental dirigidos al conocimiento medio de la ciudadanía. De hecho, en muchos supuestos, habrá de facilitarse información de alto contenido técnico similar a la dirigida a usuarios o entidades supervisoras. Ello es así por cuanto la información de profundidad será la necesaria para poder ejercer derechos por parte del interesado en cuestión, así como por los colectivos o la sociedad civil. De igual modo, la comunidad científica juega un papel muy relevante en el acceso a la información algorítmica.

Que la información deba ir dirigida a diferentes niveles de conocimiento de usuario no es problema. Como es conocido especialmente en el ámbito de la protección de datos, la información debe estar estructurada en capas, de un nivel general más accesible y simple y una capa que permite profundizar en el conocimiento[67]. Y ello debe replicarse respecto de la transparencia para el público de la IA y así se aprecia en las mejores prácticas internacionales respecto de la transparencia de la IA[68].

Explicabilidad. Tipos y elementos para lograrla

El NIST afirma que la explicabilidad va más allá de la transparencia, pero es vinculable con ella[69]. En este sentido la explicabilidad permite describir cómo el modelo genera predicciones. Supone que el usuario del modelo sepa cómo funciona y qué resultado se puede esperar según las entradas. La explicabilidad puede ser útil para mejorar el aprendizaje humano, para depurar problemas con

---

van-algoritmen-door-overheden-en-publieksvoorlichting-over-data-analyses#:~:text=Rijksoverheid-,Richtlijnen%20voor%20het%20toepassen%20van%20algoritmen,en%20publieksvoorlichting%20over%20data%2Danalyses.&text=Doel%20van%20de%20richtlijnen%20is,de%20publieksvoorlichting%20daarbij%20door%20overheden

[65] ICO (Information Comissioner Office), *What do we need to do to ensure lawfulness, fairness, and transparency in AI systems?,* ICO, 2021, https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/what-do-we-need-to-do-to-ensure-lawfulness-fairness-and-transparency-in-ai-systems/

Y de especial utilidad los estándares ICO (Information Comissioner Office), *Algorithmic transparency data standard,* ICO, 2021 (última versión julio 2022), https://www.gov.uk/government/collections/algorithmic-transparency-standard

También, de especial interés, ICO y Alan Turing Institute, *Explaining decisions … cit.*

[66] HLEG, *Directrices éticas*… cit. pág. 10.

[67] En general puede seguirse, AEPD, *Guía para el cumplimiento del deber de Informar*, AEPD, 2018, págs. 5 y ss. https://www.aepd.es/es/media/guias/guia-modelo-clausula-informativa.pdf

[68] Así, las Directrices para el sector público de Países bajos para el caso de información sobre decisiones automatizadas se afirma que la información "está "estratificada" (Rijksoverheid, *Richtlijnen voor het toepassen van algoritmen*… cit. pág. 32. Y especialmente, los estándares del ICO para el sector público están claramente estructurados en estos dos niveles (ICO, *Algorithmic transparency data standard*… cit.)

[69] NIST, *AI Risk Management Framework… cit. apartado 5.3.3.* págs. 13 y ss.

los sistemas de IA y los datos de entrenamiento. “Los sistemas explicables pueden depurarse y supervisarse más fácilmente, y se prestan a una documentación, una auditoría y una gobernanza más completas.” Y además la explicabilidad puede ser muy útil “para abordar los requisitos de transparencia”, pero advierte el NIST que “la transparencia no garantiza la explicabilidad, especialmente si el usuario no comprende los principios técnicos”. El NIST también apunta diversos riesgos de la explicabilidad: falta de coherencia en la explicación, incorrección humana de inferir cómo funciona el modelo. Como regla general “cuanto más opaco es un modelo, menos explicables se considera”.

Desde ISO se viene a definir la explicabilidad como la propiedad de un sistema de IA para expresar factores importantes que influyen en los resultados del sistema de IA de una manera comprensible para los humanos.

Para la Recomendación UNESCO 2021, la explicabilidad es esencialmente “la inteligibilidad de la entrada, salida y funcionamiento de cada componente algorítmico y la forma en que contribuye a los resultados de los sistemas” (nº 40). Su relación con la transparencia se da por cuanto los resultados del sistema IA “deberían aspirar a ser comprensibles y trazables” y los algoritmos que los generan “sean explicables”. Se añade además respecto de los sistema de IA más impactantes que “debería garantizarse que se proporcione una explicación satisfactoria” (nº 40).[70]

Para el Grupo de altos expertos de la Comisión Europea (HLEG) (nº 53) “la explicabilidad concierne a la capacidad de explicar tanto los procesos técnicos de un sistema de IA como las decisiones humanas asociadas”. Añade que “la explicabilidad técnica requiere que las decisiones que adopte un sistema de IA sean comprensibles para los seres humanos y estos tengan la posibilidad de rastrearlas”[71]. El HLEG se plantea que la precisión del sistema en ocasiones ha de ser a costa de la explicabilidad. “Además, puede que sea necesario buscar un equilibrio entre la mejora de la explicabilidad de un sistema (que puede reducir su precisión) o una mayor precisión de este (a costa de la explicabilidad).”

El punto de partida es que “El grado de necesidad de explicabilidad depende en gran medida del contexto y la gravedad de las consecuencias derivadas de un resultado erróneo o inadecuado” nº 53). En esta dirección, se detalla la explicabilidad según el “impacto significativo en la vida de las personas”. Así la explicabilidad se traduce en “una explicación adecuada del proceso de toma de decisiones del sistema de IA” que debe “adaptarse al nivel de especialización de la parte interesada” (nº 77) [72]. En otros apartados se vincula la explicabilidad con información, sin la cual “no es posible impugnar adecuadamente una decisión” (nº 53).

---

[70] Resulta de interés reproducir la afirmación desde la Recomendación UNESCO 2021, nº 40 “La explicabilidad supone hacer inteligibles los resultados de los sistemas de IA y facilitar información sobre ellos. La explicabilidad de los sistemas de IA también se refiere a la inteligibilidad de la entrada, salida y funcionamiento de cada componente algorítmico y la forma en que contribuye a los resultados de los sistemas. Así pues, la explicabilidad está estrechamente relacionada con la transparencia, ya que los resultados y los subprocesos que conducen a ellos deberían aspirar a ser comprensibles y trazables, apropiados al contexto. Los actores de la IA deberían comprometerse a velar por que los algoritmos desarrollados sean explicables. En el caso de las aplicaciones de IA cuyo impacto en el usuario final no es temporal, fácilmente reversible o de bajo riesgo, debería garantizarse que se proporcione una explicación satisfactoria con toda decisión que haya dado lugar a la acción tomada, a fin de que el resultado se considere transparente.

[71] HLEG, *Directrices éticas*… cit. pág. 18.

[72] *Ibídem*, pág. 22.

Resulta también de interés que la noción que se maneja de explicabilidad que se identifica con la "transparencia del modelo de negocio". Así, se debe dar al explicabilidad "explicaciones sobre la medida en que el sistema de IA condiciona e influye en el proceso de toma de decisiones de la organización, sobre las decisiones de diseño del sistema y sobre la lógica subyacente a su despliegue"[73]. En otros pasajes, la explicabilidad es "crucial" para la confianza y se traduce "los procesos han de ser transparentes, que es preciso comunicar abiertamente las capacidades y la finalidad de los sistemas de IA y que las decisiones deben poder explicarse —en la medida de lo posible— a las partes que se vean afectadas por ellas de manera directa o indirecta" (nº 53).

Para los sistemas IA más opacos o de caja negra, debe acudirse "a medidas relacionadas con la explicabilidad (por ejemplo, la trazabilidad, la auditabilidad y la comunicación transparente sobre las prestaciones del sistema)". (nº 53).

Resulta también de interés hacer referencia a la "explicabilidad colegiada". El Gobierno de Países Bajos afirma que la "explicabilidad consiste en poder explicar los resultados de los análisis de datos y cómo se han producido y/o pueden interpretarse en un lenguaje comprensible. Comparable a la explicación que recibimos de una persona que toma la misma decisión. Esto se refiere tanto a la explicabilidad intercolegial como a la explicabilidad para los implicados."[74] En concreto, señala que "la explicabilidad requiere que una organización tenga una visión y un control internos sobre el proceso de desarrollo y que los colegas, los equipos y los departamentos participen en el proceso de desarrollo y se expliquen unos a otros lo que están haciendo y las decisiones que están tomando. Esto se refiere a la explicabilidad colegiada."

Siguiendo a Gutiérrez[75], si la transparencia intrínseca del modelo es una característica pasiva que permite al observador humano comprender o entender el sistema, la *explicabilidad* es una característica *activa* del modelo que se refiere a la capacidad de generar una explicación sobre el comportamiento del modelo a partir de los datos utilizados, de los resultados obtenidos y del proceso completo de la toma de decisión en función de la audiencia o perfil de los destinatarios a los que se dirige la explicación. Las explicaciones son el medio a través del cual pueden explicarse las decisiones de un modelo de *machine learning* de una forma clara, comprensible, transparente e interpretable. Por tanto, si la interpretabilidad es el objetivo final a conseguir, las explicaciones son herramientas para conseguir la interpretabilidad del modelo.

Ortiz de Zárate señala en general que las explicaciones tienen que ser lo más objetivas posibles y deben generar un consenso más o menos sólido[76]. Ya para la IA, teniendo en cuenta si el destinatario del mensaje son los expertos o la ciudadanía, señala que las 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

---

[73] "Además, debería ser posible disponer de explicaciones sobre la medida en que el sistema de IA condiciona e influye en el proceso de toma de decisiones de la organización, sobre las decisiones de diseño del sistema y sobre la lógica subyacente a su despliegue (garantizando así la transparencia del modelo de negocio)." nº 77, HLEG, *Directrices éticas*… cit. pág. 22.

[74] Ministerie van Binnenlandse, *Impact Assessment… cit.* pág. 46. De igual modo en Directrices para el sector público, Rijksoverheid, *Richtlijnen voor het toepassen van algoritmen…* cit. pág. 24.

[75] M.E., Gutiérrez David, "Administraciones inteligentes… cit. pág. 55.

[76] L. Ortiz de Zárate Alcarazo, "Explicabilidad (de la inteligencia artificial)… cit. pág. 338.

que ser rigurosas y cumplir una serie de requisitos para asegurar que la función explicativa se cumple.

Asimismo, en el caso de la explicabilidad de la AI Ortiz recuerda que no existe una única explicación que sirva para explicar todos los usos de la IA[77]. En este sentido, considero que la explicabilidad debe ser adaptativa a la finalidad de uso del sistema y riesgos, así como a los fines propios que tiene la explicabilidad en su contexto particular. Es a partir de un análisis de cada caso concreto cuando cabe concretar la explicabilidad según sus tipos.

Según Ortiz de Zárate[78] puede variar el nivel o grado de la explicación; la forma, pues la explicación puede ser descriptiva (cómo funciona la tecnología de IA) o pueden ser explicaciones justificativas, esto es, que explican los criterios que se utilizan para alcanzar una explicación, es decir, el porqué. También las explicaciones pueden ser locales o completas. Son locales en los casos en los que la mejor opción sea explicar una parte del sistema que sea representativa de forma muy precisa. Son completas cuando se entiende que es más conveniente dar explicaciones de la totalidad del sistema.

Desde HLEG se establece un *check list* para la transparencia que integra el de la trazabilidad, explicabilidad y la comunicación. Para el cumplimiento de la explicabilidad, las cuestiones se centran en evaluar si las decisiones y resultados del sistema de IA son "comprensibles" "sobre las razones por las que un sistema adoptó una decisión determinada"; que se pueda facilitar a los usuarios explicación de resultados específicos. Si el sistema se diseñó "desde el principio la interpretabilidad". Es interesante también la cuestión de si se "ha investigado y tratado de utilizar el modelo más sencillo e interpretable posible para la aplicación en cuestión".

La lista de control del HLEG incluye también si se puede analizar sus datos relativos a la formación del modelo y los ensayos realizados y si se puede modificar y actualizar estos datos a lo largo del tiempo. También se incluye evaluar si tras el desarrollo del modelo es posible examinar su interpretabilidad o si dispone de acceso al flujo de trabajo interno del modelo.

Cabe destacar que en el apartado de explicabilidad la lista de evaluación del HLEG propone a evaluar si las decisiones del sistema de IA afectan a las decisiones de la organización. Se incluye también el cuestionamiento de por qué se desplegó el sistema IA, cuál es el modelo, cómo crea valor para la organización[79].

Interpretabilidad, transparencia y explicabilidad

La relación de la transparencia con la explicabilidad

Respecto de la relación de la transparencia con la explicabilidad, el HLEG afirma que la transparencia "guarda una relación estrecha con el principio de

---

[77] *Ídem*, remite en particular al trabajo de V. Arya y otros. *One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques*, 2019. https://arxiv.org/abs/1909.03012

[78] L. Ortiz de Zárate Alcarazo, "Explicabilidad (de la inteligencia artificial)… cit. pág. 338.

[79] "¿Ha evaluado en qué medida la decisión del sistema influye en los procesos de adopción de decisiones de la organización?

¿Ha evaluado por qué se desplegó ese sistema en particular en esa área concreta?

¿Ha evaluado el modelo de negocio del sistema (por ejemplo, de qué modo crea valor para la organización)?", HLEG, *Directrices éticas*… cit.

explicabilidad e incluye la transparencia de los elementos pertinentes para un sistema de IA: los datos, el sistema y los modelos de negocio."[80]

El NIST considera que "la explicabilidad está relacionada con la transparencia. Por lo general, cuanto más opaco es un modelo, menos explicable es. Sin embargo, la transparencia no garantiza la explicabilidad, especialmente si el usuario no comprende los principios técnicos."[81] (5.2.1). También desde la estandarización de ISO se insiste en la necesidad de separar estos conceptos que obviamente están relacionados y que normalmente quedan englobados bajo la "transparencia". Es más, recuerda que es posible que un nivel deseable de explicabilidad a menudo se puede lograr sin un alto nivel de transparencia. [82]

Ortiz de Zárate se centra en el concepto de explicabilidad, no obstante, afirma que la transparencia sería la habilidad para hacer visibles las componentes de un sistema de IA y sería condición necesaria, pero no suficiente para se cumpliera con el principio de explicabilidad, pues para ello se requiere un ejercicio de interpretación posterior que haga comprensible para los seres humanos todo lo que tiene lugar dentro del sistema[83].

Desde el Gobierno de Países Bajos se afirma que "es importante distinguir entre "transparencia técnica" y "explicabilidad. Se apunta que "la transparencia técnica no siempre da lugar a la explicabilidad. El proceso de razonamiento de un algoritmo no siempre es fácil de entender y comprender". Por ello, "en muchos casos, la explicabilidad es más importante que la transparencia técnica."[84]. Y es que puede ser difícil, incluso para los expertos, obtener una visión suficiente del algoritmo y su funcionamiento sobre la base de la transparencia técnica. También se afirma la "transparencia externa", esto es, la información que se ha de dar público en general y señala que a esta transparencia externa también se le denomina "explicabilidad pública"[85].

Interpretabilidad y explicabilidad ¿son lo mismo?

La diferencia entre la interpretabilidad y la explicabilidad no es tampoco muy clara. Para el NIST la interpretabilidad se refiere al significado del resultado de un algoritmo en el contexto de su finalidad. Se centraría más en la comprensión de los resultados que arroja el sistema en razón de sus finalidades, así como la posibilidad de que el usuario del sistema IA puede valorar el cumplimiento de las finalidades del sistema. Por su parte, la explicabilidad sería más relativa a la representación de los mecanismos subyacentes al funcionamiento de un algoritmo[86]. Para el organismo de referencia de EEUU, la

---

[80] *Ibídem,* nº 75.

[81] *Risk Management Framework… cit.* apartado 5.2.1.,pág. 11.

[82] Se menciona de modo general las referencias en el proceso ISO sobre IA en 2022, Inteligencia artificial — Seguridad funcional y sistemas de IA, apartado 8.3.

[83] L. Ortiz de Zárate Alcarazo, "Explicabilidad (de la inteligencia artificial)… cit. pág. 334.

[84] Ministerie van Binnenlandse, *Impact Assessment… cit.*

[85] *Ibídem,* apartado 2.4 B, con remisión al *Toetsingskader Algemene Rekenkamer*, esto es, Marco de pruebas de algoritmos: primeros pasos. https://www.rekenkamer.nl/onderwerpen/algoritmes-digitaal-toetsingskader

[86] *Risk Management Framework… cit.* apartado 5.2.2.,pág. 11: "La interpretabilidad trata de cubrir un déficit de significado. Aunque la explicabilidad y la interpretabilidad a menudo se utilizan indistintamente, la explicabilidad se refiere a una representación de los mecanismos subyacente al funcionamiento de un algoritmo, mientras que la interpretabilidad se refiere al significado de su resultado en el contexto de su propósito funcional diseñado. El supuesto subyacente es que las percepciones de riesgo se derivan de la falta de capacidad para dar sentido o contextualizar adecuadamente los resultados de los modelos.

explicabilidad trata de proporcionar una descripción de cómo el modelo genera las predicciones y se vincula a la percepción que tiene el usuario de cómo funciona el modelo y qué puede esperar para una entrada determinada[87].

Como señala Zárate, se da incluso un uso indistinto entre interpretabilidad y explicabilidad, para hacer referencia a la "habilidad para explicar o presentar sistemas de IA en términos comprensibles para los humanos"[88]. Recuerda esta autora que para algunos autores "explicabilidad" e "interpretabilidad" son términos equivalentes, al igual "comprensibilidad", "inteligibilidad" y "legibilidad" entienden que no solo[89]. Para otros, sólo son conceptos próximos, de modo que la interpretabilidad tendría que ver con la capacidad de un sistema de IA de ser comprensible para los humanos, mientras la explicabilidad añadiría que el contenido de la explicación se corresponde con la realidad que trata de explicar.[90]

Auditabilidad de los sistemas de IA y cómo lograrla

Para el HLEG la "auditabilidad" implica "la capacidad de un sistema de IA de someterse a la evaluación de sus algoritmos, datos y procesos de diseño" y recuerda que "garantizar la existencia de mecanismos de trazabilidad y registro desde las primeras fases de diseño del sistema de IA puede favorecer la auditabilidad del sistema." (nº 148, p. 48).[91]

En Países Bajos se insiste en que "Los algoritmos deben ser auditables por diseño", bajo "el principio básico de que el algoritmo debe ser rastreable y verificable". Que el sistema sea auditable implica cumplir requisitos para el algoritmo en términos de reconocimiento de datos, validación y verificabilidad

---

La interpretabilidad del modelo se refiere a la medida en que un usuario puede determinar el cumplimiento de esta función y las consiguientes implicaciones de este resultado sobre otras decisiones consecuentes para ese usuario. Las interpretaciones suelen estar contextualizadas en términos de valores y reflejan la sencillez, distinciones categóricas. Por ejemplo, una sociedad puede valorar la privacidad y la seguridad, pero los individuos pueden tener diferentes determinaciones de los umbrales de seguridad. Los riesgos para la interpretabilidad pueden ser a menudo de la interpretación que pretenden los diseñadores de los modelos, aunque esta sigue siendo un área de investigación abierta. La prevalencia de diferentes interpretaciones puede ser fácilmente medido con instrumentos psicométricos."

[87] *Ibídem*, 5.2.1 Explicabilidad. "La explicabilidad trata de proporcionar una descripción programática, a veces causal, de cómo el modelo se generan predicciones percepción que tiene el usuario de cómo funciona el modelo qué resultado que se puede esperar para una entrada determinada".

[88] L. Ortiz de Zárate Alcarazo, "Explicabilidad (de la inteligencia artificial)… cit. pág. 334. Remite en particular a los trabajos de, Doshi-Velez, con cita de F. Doshi-Velez, y B. Kim, *Towards a rigorous science of interpretable machine learning*, 2017, arXiv preprint arXiv:1702.08608.

[89] Menciona en esta dirección a autores como Cabitza o Lou.

[90] Señala en este caso a Markus et al., "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies", en *Journal of Biomedical Informatics*, pág. 113.

[91] HLEG, *Directrices éticas*… cit. pág. 48, nº 148: "La auditabilidad se refiere a la capacidad de un sistema de IA de someterse a la evaluación de sus algoritmos, datos y procesos de diseño. Constituye uno de los siete requisitos que debería cumplir cualquier sistema de IA fiable. Esto no implica necesariamente que siempre deba disponerse de forma inmediata de la información sobre los modelos de negocio y la propiedad intelectual del sistema de IA. El hecho de garantizar la existencia de mecanismos de trazabilidad y registro desde las primeras fases de diseño del sistema de IA puede favorecer la auditabilidad del sistema."

(del algoritmo). En esta línea se detallan los elementos de auditabilidad del sistema. [92]

El HLEG en su lista de control implica valorar "¿Ha establecido mecanismos para facilitar la auditabilidad del sistema por parte de agentes internos o independientes (garantizando, por ejemplo, la trazabilidad y registro de los procesos y resultados del sistema de IA)?"[93] Por su parte, la Recomendación UNESCO 2021 vincula la transparencia y la explicabilidad a la responsabilidad, rendición de cuentas y fiabilidad (nº 41)[94] y, por ello, a la auditabilidad y trazabilidad de los sistemas. En ese sentido apuesta por "mecanismos adecuados de supervisión, evaluación del impacto, auditoría y diligencia debida" y "garantizar la auditabilidad y la trazabilidad (del funcionamiento) de los sistemas de IA"[95]

Por cuanto a la trazabilidad, el Gobierno de Países Bajos señala que "en cuanto a la transparencia, el término trazabilidad también es importante. La trazabilidad significa que está claro cómo un algoritmo ha llegado a un determinado resultado", añade que "La trazabilidad desempeña un papel importante en los principios FAIR. Los principios FAIR se refieren a la trazabilidad, la accesibilidad, la interoperabilidad y la reutilización."[96] En Países Bajos para el sector público se han desarrollado intensamente los elementos de auditabilidad de los sistemas de IA públicos[97]. Y como puede seguirse, ello tiene clara relación con la transparencia. Como principio, se afirma "el principio básico de que el algoritmo debe ser rastreable y verificable". Así, se afirma los modelos, algoritmos, datos y decisiones con un impacto significativo deben ser documentados y registrados para que puedan ser verificados posteriormente. Esto implica un proceso de I+D exhaustivo y una documentación en la que se pueda rastrear el uso de los algoritmos en la producción.

Y ello se traduce en buena medida en que debe darse un diseño transparente (en este caso para el sector público). Como medidas y focos de atención se subraya que los algoritmos deben diseñarse y gestionarse de forma que sean accesibles para los controladores y supervisores y, preferiblemente, también para los expertos y los ciudadanos. Esto implica que el algoritmo -para el sector público- no debe ser confidencial; no patentado y si el algoritmo no está

---

[92] Así se afirma que la "transparencia requiere a su vez que el algoritmo sea explicable y auditable (proceso). A efectos de esta justificación mediante auditorías, se establecen requisitos para el algoritmo en términos de reconocimiento de datos, validación y verificabilidad (del algoritmo). Esto ya debería tenerse en cuenta a la hora de adquirir sistemas o desarrollar algoritmos. Los algoritmos deben ser *auditables por diseño*." Directrices para el sector público, Rijksoverheid, *Richtlijnen voor het toepassen van algoritmen…* cit. pág. 6.

[93] HLEG, *Directrices éticas…* cit. pág. 48. Respecto de auditabilidad ver lista de control pág. 41.

[94] "41. La transparencia y la explicabilidad están estrechamente relacionadas con las medidas adecuadas de responsabilidad y rendición de cuentas, así como con la fiabilidad de los sistemas de IA."

[95] "43. Deberían elaborarse mecanismos adecuados de supervisión, evaluación del impacto, auditoría y diligencia debida, incluso en lo que se refiere a la protección de los denunciantes de irregularidades, para garantizar la rendición de cuentas respecto de los sistemas de IA y de su impacto a lo largo de su ciclo de vida. Dispositivos tanto técnicos como institucionales deberían garantizar la auditabilidad y la trazabilidad (del funcionamiento) de los sistemas de IA, en particular para intentar solucionar cualquier conflicto con las normas relativas a los derechos humanos y las amenazas al bienestar del medio ambiente y los ecosistemas."

[96] Ministerie van Binnenlandse, *Impact Assessment… cit.* 2.4 b).

[97] Directrices para el sector público, Rijksoverheid, *Richtlijnen voor het toepassen van algoritmen…* cit. pág. 26.

publicado, debe ser posible auditarlo. Asimismo, el algoritmo debe estar documentado de modo que el código describe lo que ocurre y existe una documentación técnica que describe el desarrollo del modelo. A lo anterior se añade que en lo posible hay que utilizar algoritmos de código abierto o que se pongan a disposición como código abierto.

Lo anterior se afirma como directrices ha de ser "sobre la base de una justificación adecuada" Y ello incluiría justificar las decisiones, como la elección de los algoritmos específicos y los datos utilizados. Así las cosas, la documentación técnica ha de describir por qué se ha elegido el algoritmo final y qué datos se han utilizado de manera bien fundamentada. se afirman también principios como que "siempre que sea posible, utilice métodos sencillos en lugar de complejos. Esto beneficiará la explicabilidad, la auditabilidad y la mitigación de riesgos." De ahí que haya que explicar por qué se han elegido técnicas de inteligencia artificial, por ejemplo, porque producen resultados mucho mejores que un método simple.

Relacionado con la transparencia, se señala como criterio que hay que proporcionar una descripción detallada del modelo y su funcionamiento, junto con una validación verificable de que el código se ajusta a la especificación. Para ello la documentación técnica describe detalladamente el funcionamiento del modelo. También es importante documentar en el código lo que hace o tiene que hacer una rutina para que otro analista pueda revisar el código adecuadamente. Igualmente se incluye que se registre las observaciones, como las desviaciones en los datos o los resultados inesperados o inexplicables.

Asimismo se añaden no pocos elementos por cuanto a la verificación de resultados. También se señala que se debe garantizar que los resultados del modelo son reproducibles.

Trazabilidad (y documentabilidad)

La "trazabilidad" es para el HLEG uno de los 7 elementos clave (nº 4) y esencialmente implica el deber de documentar de manera rigurosa los conjuntos de datos y los procesos y algoritmos utilizados con la finalidad de identificar los motivos de una decisión errónea por el sistema y prevenir futuros errores. Se recuerda que la trazabilidad "facilita la auditabilidad y la explicabilidad" (nº 76)[98].

A tal fin, deben documentarse los métodos utilizados para diseñar y desarrollar programación o la forma en que se creó el modelo; los datos de entrada que se recopilaron y seleccionaron, los métodos empleados para ensayar y validar el sistema algorítmico; los escenarios o casos de uso; los ensayos y la validación; los resultados del sistema algorítmico así como otras posibles decisiones que se producirían en casos diferentes.

Testabilidad, comprobabilidad, verificabilidad (y replicabilidad)

En Países Bajos para el sector público se detallan también el concepto de testabilidad. Así, bajo el ámbito de "testabilidad" se introduce la

[98] HLEG, *Directrices éticas*… cit. nº 76: "Trazabilidad. Los conjuntos de datos y los procesos que dan lugar a la decisión del sistema de IA, incluidos los relativos a la recopilación y etiquetado de los datos así como a los algoritmos utilizados, deberían documentarse con arreglo a la norma más rigurosa posible con el fin de posibilitar la trazabilidad y aumentar la transparencia. Esto también es aplicable a las decisiones que adopte el sistema de IA. Esto permitirá identificar los motivos de una decisión errónea por parte del sistema, lo que a su vez podría ayudar a prevenir futuros errores. La trazabilidad, por tanto, facilita la auditabilidad y la explicabilidad."

"comprobabilidad", que “consiste en poder probar realmente los resultados”[99]. La comprobabilidad va más allá de la explicabilidad, que sólo describe los resultados en lenguaje comprensible, pero no los comprueba. Así, el análisis de datos debe organizarse de forma que se pueda comprobar el método de análisis de datos, los algoritmos utilizados, los conjuntos de datos y el procesamiento real. Y se apuesta especialmente por la transparencia, verificabilidad y cuando se aplica el análisis de datos para la toma de decisiones.

Por su parte, la “verificabilidad” iría vinculada a las actividades de autoridades supervisión y los tribunales. Se remite en este sentido al *Marco Normativo para la Auditoría de Algoritmos y el Tribunal de Cuentas holandés* (AR)[100]. Por cuanto a la verificabilidad, se identifica con la posibilidad de probar un algoritmo, se afirma que los algoritmo en sí mismos deberían ser verificables. Aunque se señala que es diferente a la transparencia, pero se afirma que “transparencia y la verificabilidad están relacionadas”.

El estudio de referencia desde Harvard aúna los principios de "verificabilidad y replicabilidad"[101]. Para garantizar que los sistemas de IA funcionan como deberían se afirma que un experimento de IA debe "mostrar el mismo comportamiento cuando se repite en las mismas condiciones" y proporcionar suficientes detalles sobre sus operaciones para que pueda ser validado.

Transparencia algorítmica como “comunicación”, notificación de interacciones y de decisiones con inteligencia artificial

La “comunicación” es una noción que aparece de un modo u otro vinculada en los documentos básicos de la IA. Como ahora se expone, aglutina diversos elementos, en algunos casos heterogéneos. El HLEG hace referencia a la “comunicación” esencialmente como “derecho a saber que [las personas] están interactuando con un sistema de IA” (nº 78) y para ello los sistemas de IA se deben identificar como tales a los humanos.[102] Siguiendo el mencionado estudio de referencia de Harvard, la comunicación se desdobla en primer lugar, como un principio de "notificación cuando se interactúa con un sistema de IA"[103], de modo que “los seres humanos siempre deben ser conscientes cuando se relacionan con la tecnología en lugar de hacerlo directamente con otra persona”. Ello es

---

[99] Directrices para el sector público, Rijksoverheid, *Richtlijnen voor het toepassen van algoritmen…* cit. pág. 30.

[100] Cabe remitir al documento ECLI:NL:RVS:2017:1259, https://www.recht.nl/rechtspraak/uitspraak/?ecli=ECLI:NL:RVS:2017:1259, en particular los apartados 14.3 y 14.4. Se trata del Asesoramiento no solicitado del Consejo de Estado sobre los efectos de la digitalización en las relaciones entre el Estado y la ley", Documentos Parlamentarios II, 2017/18, 26643, n.º 557.

[101] J. Fjeld, et. al., “Principled Artificial Intelligence… cit.

[102] HLEG, *Directrices éticas*… cit. nº 78: “Comunicación. Los sistemas de IA no deberían presentarse a sí mismos como humanos ante los usuarios; las personas tienen derecho a saber que están interactuando con un sistema de IA. Por lo tanto, los sistemas de IA deben ser identificables como tales. Además, cuando sea necesario, se debería ofrecer al usuario la posibilidad de decidir si prefiere interactuar con un sistema de IA o con otra persona, con el fin de garantizar el cumplimiento de los derechos fundamentales. Más allá de lo expuesto, se debería informar sobre las capacidades y limitaciones del sistema de IA a los profesionales o usuarios finales; dicha información debería proporcionarse de un modo adecuado según el caso de uso de que se trate y debería incluir información acerca del nivel de precisión del sistema de IA, así como de sus limitaciones.”

[103] J. Fjeld, et. al., “Principled Artificial Intelligence… cit. “Notification when AI Makes a Decision about an Individual Regular Reporting Requirement”.

especialmente respecto de “interacciones de chatbots, los sistemas de reconocimiento facial, los sistemas de puntuación de crédito y, en general, "cuando los sistemas de aprendizaje automático se utilizan en la esfera pública"". Igualmente se afirma un “principio de interacción” para advertir que existe, por ejemplo, un sistema de reconocimiento facial.

A este respecto cabe destacar el artículo 52 del futuro Reglamento de IA de la UE por cuanto la comunicación de interacciones viene recogida especialmente como Título IV y artículo 52: “Obligaciones de transparencia para determinados sistemas de IA”. Se dispone que “Los proveedores garantizarán que los sistemas de IA destinados a interactuar con personas físicas estén diseñados y desarrollados de forma que dichas personas estén informadas de que están interactuando con un sistema de IA” (1º). Asimismo se debe informar del funcionamiento del sistema -y obviamente de su existencia- a las personas “expuestas” a “un sistema de reconocimiento de emociones o de un sistema de categorización biométrica” 2º). También se debe conocer cuando un “sistema de IA que genere o manipule contenido de imagen, sonido o vídeo que se asemeje notablemente a personas, objetos, lugares u otras entidades o sucesos existentes, y que pueda inducir erróneamente a una persona a pensar que son auténticos o verídicos (ultrafalsificación), harán público que el contenido ha sido generado de forma artificial o manipulado.”

En segundo lugar, la comunicación se desdobla en un “principio de notificación” de la adopción de decisiones por un sistema de IA[104], que vendría acompañado en su caso de la opción de no utilizar esos sistemas o aportar la información a los mismos. Se conecta especialmente respecto de las decisiones sobre individuos y las garantías frente a las mismas, ya se trate del artículo 22 RGPD o decisiones que impacten en otros derechos fundamentales. Sin entrar ahora en detalles, esta “notificación” y la información aparejas informar “sobre los datos personales utilizados en el proceso de toma de decisiones, "acceso a los factores, la lógica y las técnicas que produjeron el resultado" de un sistema de IA y, en general, "cómo se llega a los procesos de toma de decisiones automatizados y de aprendizaje automático".[105]

Además, el HLEG maneja el concepto de “comunicación” integrado en general en la transparencia[106]. Sin embargo, se trata de una noción muy variada y heterogénea. Así, lista de evaluación incluye en comunicación comprobar que se ha informado a los usuarios (finales) que están interactuando con un sistema de IA. También si se se ha etiquetado su sistema como de IA. Desde el punto de vista de las decisiones adoptadas, se introducen cuestiones sobre si se informa a de las razones y criterios subyacentes a los resultados del sistema de IA.

También, en “comunicación” el HLEG incluye cuestiones variadas aunque relacionadas con la transparencia. Podrían agruparte unas en tener en cuenta a los destinatarios o afectados finales del sistema (aunque se denominan usuarios) a los efectos de adaptar el sistema, ver cómo les afecta y, sobre todo que la información que se les facilite sobre el mismo sea eficaz. Así, se señala la evaluación de:

-si hay procesos que tengan en cuenta las opiniones de los usuarios y que utilicen dichas opiniones para adaptar el sistema.

---

[104] *Ibídem*, “Notification when Interacting with an AI.”

[105] *Ídem*.

[106] HLEG, *Directrices éticas*… cit.

-Si se informa hacia otras audiencias, hacia terceros o hacia el público en general.

Según el caso de uso, ¿ha tenido en cuenta la psicología humana y sus posibles limitaciones, como el riesgo de confusión, el sesgo de confirmación o la fatiga cognitiva?

Por otra parte, en el apartado de “comunicación” ciertamente se incluyen obligaciones de transparencia y obligaciones de información sobre:

- los riesgos potenciales o percibidos, como la posible existencia de sesgos;

-si se ha dejado claro el propósito del sistema de IA y quién o qué podrá beneficiarse del producto o servicio que ofrezca éste;

- información clara sobre los escenarios de utilización del producto, con información sea comprensible y adecuada para los usuarios.

- las características, limitaciones y posibles carencias del sistema de IA ya para los encargados de su despliegue en un producto o servicio así como para usuarios finales o consumidores.

Otros contenidos bajo el genérico de la transparencia: informes periódicos, código abierto, contratación pública abierta, transparencia de las evaluaciones de impacto

Para el estudio de referencia que analiza prácticamente todos los documentos de IA bajo el genérico de la transparencia se incluye también el “principio de "información periódica”[107]. Así, los sistemas deben revelar sistemáticamente información importante sobre su uso. Se conecta también con la finalidad desarrollar métricas comparables a nivel internacional para medir la investigación, el desarrollo y el despliegue de la IA y reunir las pruebas necesarias para respaldar estas afirmaciones[108].

El principio de "datos y algoritmos de código abierto" se vincula al concepto ya conocido de código abierto y al desarrollo de algoritmos comunes y de la investigación y colaboración abiertas para apoyar el avance de la tecnología, evitar monopolios, etc.[109]

Asimismo, se afirma como principio inclusivo de la transparencia a la “contratación pública abierta”. Se sigue en este sentido a Access Now cuando recomienda que: "Cuando un organismo gubernamental pretenda adquirir un sistema de IA o sus componentes, la contratación debe realizarse de forma abierta y transparente, de acuerdo con las normas de contratación pública. Esto incluye la publicación del propósito del sistema, los objetivos, los parámetros y otra información para facilitar la comprensión del público. La contratación debe incluir un periodo para los comentarios del público, y los Estados deben ponerse en contacto con los grupos potencialmente afectados cuando sea pertinente para garantizar la oportunidad de hacer aportaciones."[110] El tema de la contratación

---

[107] Fjeld, et. al., “Principled Artificial Intelligence… cit. pág. 47.

[108] En esta dirección de mencionan los documentos básicos de OCDE, Ministros de Comercio y Ministros de Economía Digital del G20, Access Now, Universidad de Montreal o Amnistía Internacional.

[109] Se recuerda en este sentido la Declaración de Montreal, Universidad de Montreal (n 34) pág. 13 (principio 6.7.), entre otros.

[110] Access Now, “Human Rights in the Age of Artificial Intelligence”, Access Now, 2018, nº 9, pág. https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

por el sector público de sistemas de IA bien merece un análisis jurídico detallado[111].

Desde la Recomendación UNESCO 2021 se afirma el principio de la transparencia de las evaluaciones de impacto (nº 51 y 53). Así, se afirma que "deberían aplicarse protocolos de transparencia ejecutables, que correspondan al acceso a la información, incluida la información de interés público en poder de entidades privadas" (nº 51) y en concreto, "las evaluaciones del impacto ético deberían ser transparentes y abiertas al público, cuando proceda" (nº 53).

Una recapitulación, para concluir

Este estudio ha pretendido dar un exhaustivo acercamiento conceptual a la transparencia y explicabilidad de la IA. Quien suscribe lleva más de dos décadas dedicado a la investigación sobre transparencia y sobre el ámbito digital y los últimos años sobre inteligencia artificial. Y lo cierto es que me he remontado a reflexiones de hace veinte años para recordar que la transparencia es un concepto polisémico, evocador y esencialmente instrumental. Y ello vuelve a apreciarse claramente en la novedosa proyección de la exigencia de transparencia ahora en el disruptivo ámbito de la inteligencia artificial. Sobre esta base, se ha apuntado cómo unánimamente los documentos de referencia de la llamada *ética de la IA* proclaman la transparencia -y explicabilidad- como el principio más consensuado, eso sí, con muchos contenidos y significados bajo dicho *paraguas*.

Siendo que la transparencia es esencialmente un instrumento para lograr unas finalidades, el estudio se cuestiona el ¿para qué? de la transparencia y los muy variados motivos o finalidades que conducen a la "Roma" de la transparencia de la IA. Con carácter general, esencialmente la transparencia algorítmica es un elemento instrumental que favorece la necesitada confiabilidad en la IA. Y sobre todo puede discernirse, por un lado, que la transparencia algorítmica es la herramienta básica de la garantía de toda una serie de principios democráticos y derechos fundamentales. Así sucede especialmente respecto de la que luego se define como "transparencia externa". Por otro lado, la transparencia concentra otro tipo de finalidades cuando se trata de la transparencia "interna", que es pieza esencial para el conocimiento y comprobación del buen funcionamiento del sistema de IA por los sujetos de la cadena de valor (usuarios del sistema, importadores, distribuidores, etc.), así como todos aquellos que tienen que verificar, comprobar el mismo (autoridades, evaluadores, etc.). Estas dos grandes fuentes de la necesidad de la transparencia de la IA no son en modo alguno compartimentos estancos. No obstante, unas y otras finalidades sí que tienen en general derivadas importantes, como por ejemplo, la mayor vinculación a derechos fundamentales de la transparencia externa, que también queda especialmente vinculada a garantías del así uso público de algoritmos.

Este estudio aproximativo no pretendía acometer un análisis normativo. No obstante, se ha hecho un seguimiento básico de las actuales exigencias normativas de transparencia algorítmica. Desde las normas ya más clásicas vinculadas a las decisiones automatizadas y protección de datos (art. 22 RGPD

[111] Para ello, entre otros G. Vestri, "Sistemas de inteligencia artificial en la contratación pública: entre códigos fuente y datos abiertos", *Actualidad administrativa*, Nº 12, 2021. También, J. Miranzo Díaz, "Inteligencia artificial y contratación pública", en I. Martín (Dir.), *Administración electrónica, transparencia y contratación pública*, INAP, 2020, págs. 105-142.

y afines), a las normas de transparencia recientes como la ley valenciana de 2022 o las diversas que regulan el uso público de IA. Esperemos que siguiendo la senda del apartado XVIII de la Carta de Derechos digitales haya una clara mejora regulatoria en este sentido. De igual modo, a lo largo del estudio se han visto aun someramente las exigencias de transparencia "interna", que esencialmente vienen impuestas por la normalización técnica de la IA y, sobre todo, por el futuro Reglamento de IA de la UE con sus muchas exigencias de transparencia, documentación técnica, etc. También se ha descrito cómo recientemente el ámbito laboral (la llamada *ley rider)* o el de las plataformas (*Digital Services Act*) se suman a las regulaciones que exigen la transparencia de los algoritmos.

También se han formulado algunos elementos orientativos para determinar ¿cuánta transparencia? Así, se afirma que el grado de transparencia habrá de ser mayor a impacto y riesgo por el uso de la IA. También habrá de ser mayor según la mayor opacidad y el tipo de algoritmo. Asimismo, a mayor participación, control y supervisión humana en las decisiones de la IA es posible exigir menor transparencia y explicabilidad del sistema de IA utilizada. Asimismo y como principio, también mayor transparencia se ha de dar cuando haya un uso de algoritmos por el poder público. Lamentablemente, el muy mejorable artículo 41 Ley 40/2015 y otras regulaciones no parece ir en esa dirección y parece que el sector público no tenga casi interés en exigirse garantías de transparencia y explicabilidad cuando es él quien utiliza los algoritmos y sistemas de IA.

Una vez realizadas las reflexiones y análisis anteriores, el estudio se ha detenido en la "transparencia" algorítmica como concepto general inclusivo y la miríada de nociones que orbitan a su alrededor, a saber: *trazabilidad, explicabilidad, interpretabilidad, comprensibilidad, inteligibilidad, legibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad, descomponibilidad, verificabilidad, replicabilidad, comunicación, código abierto*, así como referencias a la *transparencia técnica en sentido estricto*, o la importante diferencia que se ha seguido entre *transparencia externa e interna*.

Se ha iniciado este análisis conceptual partiendo de la llamada transparencia "técnica" y la transparencia intrínseca. Se trataría de algún modo en el grado de comprensibilidad y la interpretabilidad de un modelo específico por sí mismo. No obstante, cuando el sistema de IA no es transparente técnicamente y es un modelo opaco, hay que acudir a variadas técnicas o estrategias de evaluación para juzgar la transparencia y la explicabilidad.

El estudio se ha cuestionado ¿para quién? Es la transparencia algorítmica. De ahí se ha discernido de las ya mencionadas transparencia "interna" y "externa" y su reflejo en el futuro Reglamento UE de inteligencia artificial como ejemplo y la transparencia "externa" esencialmente como garantía de valores, principios y derechos constitucionales.

Una vez abordado esencialmente el concepto de transparencia algorítmica y sus elementos afines, el estudio se centra en el fundamental concepto de la explicabilidad de la IA, la pareja de baile esencial de la transparencia. De hecho, en no pocas ocasiones la explicabilidad es más importante que la transparencia técnica o que la facilitación de toda una serie de datos o información sobre el sistema IA. Se ha profundizado en el contenido de la explicabilidad y los medios para lograrla. Y ello ha llevado a apreciar los vínculos del *triángulo* de interpretabilidad, transparencia y explicabilidad. De un lado, se analiza la relación de la transparencia con la explicabilidad, para luego cuestionarse si la

interpretabilidad y explicabilidad son lo mismo. De ahí resulta una transición entre los tres conceptos que los hace prácticamente inescindibles.

A partir de ahí, el estudio recorre no pocos conceptos también esenciales para el genérico de la transparencia algorítmica. Así, la fundamental noción de la auditabilidad de los sistemas de IA y se apuntan algunos elementos de cómo lograrla. Los algoritmos deben ser auditables por diseño, rastreables y verificables. Ello conduce necesariamente a explicar otros conceptos como los de trazabilidad (y documentabilidad), testabilidad, comprobabilidad, verificabilidad (y replicabilidad). Finalmente, se han examinado las exigencias de "comunicación", un elemento también incluido en la transparencia de la IA. Se trata de la obligación de comunicar al humano que se interactúa con un sistema de IA, informar si éste evalúa las emociones, genera audios o vídeos falsos o manipulados, toma decisiones respecto de nosotros o, en general, que los sistemas de IA se usan en la esfera pública. La "comunicación" de la IA también incluyen una miscelánea de obligaciones de información en diversos documentos.

Este exhaustivo recorrido por todas las facetas de la amplia denotación de la transparencia de la IA concluye describiendo con otros contenidos que se suelen también incluir en este amplio concepto: obligación de informes periódicos, código abierto, contratación pública abierta de la IA, o transparencia de las evaluaciones de impacto.

Así las cosas, se ha intentado aportar luz, o al menos conocimiento respecto de un tema que es elemento esencial del futuro que nos aguarda. Por razones de extensión, se dejan para futuras publicaciones la determinación concreta de todos los elementos respecto de los cuales debe darse la transparencia y explicabilidad algorítmica, en razón de las finalidades y la concreta regulación. En cualquier caso, éstas y otras muchas otras cuestiones de interés se abordan en esta monografía por excelentes autores que sin duda será de igual o mayor interés para el lector.