

monografías

ALTA CALIDAD EN
INVESTIGACIÓN
JURÍDICA



LORENZO COTINO HUESO
JORGE CASTELLANOS CLARAMUNT
Editores

TRANSPARENCIA Y
EXPLICABILIDAD DE
LA INTELIGENCIA
ARTIFICIAL

**TRANSPARENCIA Y
EXPLICABILIDAD DE
LA INTELIGENCIA
ARTIFICIAL**

COMITÉ CIENTÍFICO DE LA EDITORIAL TIRANT LO BLANCH

- MARÍA JOSÉ AÑÓN ROIG**
*Catedrática de Filosofía del Derecho de la
Universidad de Valencia*
- ANA CAÑIZARES LASO**
*Catedrática de Derecho Civil
de la Universidad de Málaga*
- JORGE A. CERDIO HERRÁN**
*Catedrático de Teoría y Filosofía de Derecho
Instituto Tecnológico Autónomo de México*
- JOSÉ RAMÓN COSSÍO DÍAZ**
*Ministro en retiro de la Suprema Corte
de Justicia de la Nación y miembro de
El Colegio Nacional*
- EDUARDO FERRER MAC-GREGOR POISOT**
*Juez de la Corte Interamericana de Derechos
Humanos. Investigador del Instituto de
Investigaciones Jurídicas de la UNAM*
- OWEN FISS**
*Catedrático emérito de Teoría del Derecho de la
Universidad de Yale (EEUU)*
- JOSÉ ANTONIO GARCÍA-CRUCES GONZÁLEZ**
Catedrático de Derecho Mercantil de la UNED
- LUIS LÓPEZ GUERRA**
*Catedrático de Derecho Constitucional de la
Universidad Carlos III de Madrid*
- ÁNGEL M. LÓPEZ Y LÓPEZ**
*Catedrático de Derecho Civil de la
Universidad de Sevilla*
- MARTA LORENTE SARIÑENA**
*Catedrática de Historia del Derecho de la
Universidad Autónoma de Madrid*
- JAVIER DE LUCAS MARTÍN**
*Catedrático de Filosofía del Derecho y Filosofía
Política de la Universidad de Valencia*
- VÍCTOR MORENO CATENA**
*Catedrático de Derecho Procesal
de la Universidad Carlos III de Madrid*
- FRANCISCO MUÑOZ CONDE**
*Catedrático de Derecho Penal
de la Universidad Pablo de Olavide de Sevilla*
- ANGELIKA NUSSBERGER**
*Catedrática de Derecho Constitucional e
Internacional en la Universidad de Colonia
(Alemania)*
Miembro de la Comisión de Venecia
- HÉCTOR OLASOLO ALONSO**
*Catedrático de Derecho Internacional de la
Universidad del Rosario (Colombia) y
Presidente del Instituto Ibero-Americano de
La Haya (Holanda)*
- LUCIANO PAREJO ALFONSO**
*Catedrático de Derecho Administrativo de la
Universidad Carlos III de Madrid*
- CONSUELO RAMÓN CHORNET**
*Catedrática de Derecho Internacional Público y
Relaciones Internacionales de la Universidad de
Valencia*
- TOMÁS SALA FRANCO**
*Catedrático de Derecho del Trabajo y de la
Seguridad Social de la Universidad de Valencia*
- IGNACIO SANCHO GARGALLO**
*Magistrado de la Sala Primera (Civil) del
Tribunal Supremo de España*
- TOMÁS S. VIVES ANTÓN**
*Catedrático de Derecho Penal de la
Universidad de Valencia*
- RUTH ZIMMERLING**
*Catedrática de Ciencia Política de la
Universidad de Mainz (Alemania)*

Procedimiento de selección de originales, ver página web:

www.tirant.net/index.php/editorial/procedimiento-de-seleccion-de-originales

TRANSPARENCIA Y EXPLICABILIDAD DE LA INTELIGENCIA ARTIFICIAL

Lorenzo Cotino Hueso
Jorge Castellanos Claramunt
(Editores)



tirant lo blanch
Valencia, 2022

Copyright © 2022

Todos los derechos reservados. Ni la totalidad ni parte de este libro puede reproducirse o transmitirse por ningún procedimiento electrónico o mecánico, incluyendo fotocopia, grabación magnética, o cualquier almacenamiento de información y sistema de recuperación sin permiso escrito de los autores y del editor.

En caso de erratas y actualizaciones, la Editorial Tirant lo Blanch publicará la pertinente corrección en la página web www.tirant.com.

© VV.AA.

© TIRANT LO BLANCH
EDITA: TIRANT LO BLANCH
C/ Artes Gráficas, 14 - 46010 - Valencia
TELF.S.: 96/361 00 48 - 50
FAX: 96/369 41 51
Email: tlb@tirant.com
www.tirant.com
Librería virtual: www.tirant.es
ISBN: 978-84-1147-161-9
MAQUETA: Tink Factoría de Color

Si tiene alguna queja o sugerencia, envíenos un mail a: atencioncliente@tirant.com. En caso de no ser atendida su sugerencia, por favor, lea en www.tirant.net/index.php/empresa/politicas-de-empresa nuestro procedimiento de quejas.

Responsabilidad Social Corporativa: <http://www.tirant.net/Docs/RSCTirant.pdf>

Autores

(por orden de aparición)

LORENZO COTINO HUESO

*Catedrático de Derecho Constitucional
Universitat de València*

ALFRED PERIS MANGUILLOT

*Catedrático. Institut Universitari de Matemàtica Pura i Aplicada
Universitat Politècnica de València*

RAQUEL VALLE ESCOLANO

*Profesora de Derecho Constitucional
UNIR*

GABRIELE VESTRI

*Profesor Ayudante Doctor de Derecho Administrativo
Universidad de Cádiz*

MARÍA ESTRELLA GUTIÉRREZ DAVID

*Profesora de Derecho Constitucional
Universidad Complutense de Madrid*

SALVATORE PALUMBO

*Doctor por la Universitat de València
y la Università degli Studi di Palermo*

JUAN MANUEL BELLOTO GÓMEZ

Director de GuIA, OdiseIA

JAVIER PLAZA PENADÉS

*Catedrático de Derecho Civil
Universitat de València*

ADRIÁN TODOLÍ SIGNE

*Profesor Titular de Derecho del Trabajo y de la Seguridad Social
Universitat de València*

Contenido general de la obra

Bloque I. Transparencia y explicabilidad de la inteligencia artificial. Elementos conceptuales, generales y de género

- *Transparencia y explicabilidad de la inteligencia artificial y “compañía” (comunicación, interpretabilidad, inteligibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad...).* Para qué, para quién y cuánta, por Lorenzo Cotino Hueso, Catedrático de Derecho Constitucional, Universitat de Valencia.
- *Algoritmos: ¿podemos hacerlos transparentes y trazables en su proceso?*, por Alfred Peris Manguillot, Catedrático. Institut Universitari de Matemàtica Pura i Aplicada, Universitat Politècnica de València.
- *Transparencia en la inteligencia artificial y en el uso de algoritmos: una visión de género*, por Raquel Valle Escolano, Profesora de Derecho Constitucional, UNIR.

Bloque II. Transparencia y explicabilidad en el uso de la inteligencia artificial por el sector público

- *Denegación vs. Derecho de acceso al código fuente en los sistemas algorítmicos. Una perspectiva jurídico-administrativa*, por Gabriele Vestri, Prof. Ayudante Doctor de Derecho Administrativo, Universidad de Cádiz.
- *Acceso al código fuente y a los algoritmos de las Administraciones inteligentes. Lecciones a partir de experiencias comparadas*, por María Estrella Gutiérrez David, Profesora de Derecho Constitucional, Universidad Complutense de Madrid.
- *Algoritmo y transparencia de la administración pública en el ordenamiento jurídico italiano. Algunas reflexiones críticas a partir del caso “Buona Scuola”*, por Salvatore Palumbo, Doctor por la Universitat de València y la Università degli Studi di Palermo.

Bloque III. Transparencia y explicabilidad en la empresa, en la contratación y prestación de servicios de internet y en el ámbito laboral

- *Explicabilidad e inteligencia artificial responsable. Recomendaciones aterrizadas al día a día de las empresas*, por Juan Manuel Belloto Gómez, Director de GuIA, OdiselA.
- *La moderna configuración de la transparencia en la prestación de servicios y en la contratación en Internet*, por Javier Plaza Penadés, Catedrático de Derecho Civil, Universitat de València.
- *Transparencia algorítmica en el ámbito laboral*, por Adrián Todolí Signes, Prof. Titular de Derecho del Trabajo y de la Seguridad Social, Universitat de Valencia.

Índice

Contenido general de la obra	9
Presentación	17

Bloque I TRANSPARENCIA Y EXPLICABILIDAD DE LA INTELIGENCIA ARTIFICIAL. ELEMENTOS CONCEPTUALES, GENERALES Y DE GÉNERO

Transparencia y explicabilidad de la inteligencia artificial y “compañía” (comunicación, interpretabilidad, inteligibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad...). Para qué, para quién y cuánta.....	25
--	----

Lorenzo Cotino Hueso

I. LA TRANSPARENCIA, UN CONCEPTO POLISÉMICO, EVO-CADOR Y ESENCIALMENTE INSTRUMENTAL, AHORA APLICADO AL ÁMBITO DE LA INTELIGENCIA ARTIFICIAL	25
II. TRANSPARENCIA Y EXPLICABILIDAD: EL PRINCIPIO DE REFERENCIA DE LA INTELIGENCIA ARTIFICIAL CONFIABLE.....	29
III. TRANSPARENCIA Y EXPLICABILIDAD ¿PARA QUÉ? LOS MUCHOS “CAMINOS” LOS QUE LLEVAN A LA “ROMA” DE LA TRANSPARENCIA.....	31
IV. UN ESBOZO DE LAS EXIGENCIAS NORMATIVAS DE TRANSPARENCIA ALGORÍTMICA.....	36
V. ¿CUÁNTA TRANSPARENCIA?.....	38
VI. LA “TRANSPARENCIA” COMO CONCEPTO GENERAL IN-CLUSIVO Y LA MIRÍADA DE NOCIONES QUE ORBITAN A SU ALREDEDOR	41
VII. TRANSPARENCIA “TÉCNICA”, TRANSPARENCIA INTRÍN-SECA, INTERPRETABILIDAD, COMPRENSIBILIDAD, DES-COMPONIBILIDAD Y SIMULABILIDAD DEL SISTEMA DE INTELIGENCIA ARTIFICIAL	42
VIII. TRANSPARENCIA ¿PARA QUIÉN? TRANSPARENCIA “IN-TERNA” Y “EXTERNA”.....	45
1. Transparencia interna y externa	45
2. La transparencia “interna” y el futuro Reglamento UE de inte-ligencia artificial como ejemplo	46

3. La transparencia “externa” como garantía de valores, principios y derechos constitucionales	48
IX. EXPLICABILIDAD. TIPOS Y ELEMENTOS PARA LOGRARLA	50
X. INTERPRETABILIDAD, TRANSPARENCIA Y EXPLICABILIDAD	55
1. La relación de la transparencia con la explicabilidad.....	55
2. Interpretabilidad y explicabilidad ¿son lo mismo?	57
XI. AUDITABILIDAD DE LOS SISTEMAS DE IA Y CÓMO LOGRARLA	58
XII. TRAZABILIDAD (Y DOCUMENTABILIDAD)	61
XIII. TESTABILIDAD, COMPROBABILIDAD, VERIFICABILIDAD (Y REPLICABILIDAD)	62
XIV. TRANSPARENCIA ALGORÍTMICA COMO “COMUNICACIÓN”, NOTIFICACIÓN DE INTERACCIONES Y DE DECISIONES CON INTELIGENCIA ARTIFICIAL.....	63
XV. OTROS CONTENIDOS BAJO EL GENÉRICO DE LA TRANSPARENCIA: INFORMES PERIÓDICOS, CÓDIGO ABIERTO, CONTRATACIÓN PÚBLICA ABIERTA, TRANSPARENCIA DE LAS EVALUACIONES DE IMPACTO.....	65
XVI. UNA RECAPITULACIÓN, PARA CONCLUIR	67

Algoritmos: ¿podemos hacerlos transparentes y trazables en su proceso?	71
---	----

Alfred Peris Manguillot

I. EL CONCEPTO DE ALGORITMO	71
II. REDES NEURONALES Y APRENDIZAJE PROFUNDO	74
III. INTELIGENCIA ARTIFICIAL Y JUSTICIA	78
IV. ALGORITMOS MÁS ALLÁ DE LA INTELIGENCIA ARTIFICIAL.....	80
V. CONCLUSIONES	82

Transparencia en la inteligencia artificial y en el uso de algoritmos: una visión de género	85
--	----

Raquel Valle Escolano

I. INTRODUCCIÓN	85
II. LA IA Y EL USO DE ALGORITMOS: UN NUEVO PARADIGMA CON EVIDENTES VENTAJAS Y ÁREAS DE RIESGO	86
III. LA DIMENSIÓN ÉTICA COMO EJE CLAVE DE LA IA Y DE UNA GOBERNANZA MEDIADA POR ALGORITMOS	89
IV. LAS (INEVITABLES) IMPLICACIONES DE GÉNERO	91
1. La IA y el uso de algoritmos como oportunidad en materia de igualdad de género	93

2. El protagonismo de las mujeres en el campo de la inteligencia artificial. Eventuales consecuencias en el trabajo femenino	94
3. Sesgos de género que quiebran la igualdad y generan discriminación	96
V. LA TRANSPARENCIA COMO PROBLEMA DE FONDO	102
VI. A MODO DE CONCLUSIÓN	105

Bloque II

TRANSPARENCIA Y EXPLICABILIDAD EN EL USO DE LA INTELIGENCIA ARTIFICIAL POR EL SECTOR PÚBLICO

Denegación vs. Derecho de acceso al código fuente en los sistemas algorítmicos. Una perspectiva jurídico-administrativa.....	111
---	------------

Gabriele Vestri

I. INTRODUCCIÓN	111
II. EL CÓDIGO FUENTE COMO ELEMENTO DEL MODELO FUNCIONAL	113
III. EL ACCESO AL CÓDIGO FUENTE ANTE LA LEY 19/2013, DE 9 DE DICIEMBRE (LTAIBG)	119
IV. EL SESGO ALGORÍTMICO COMO BASE DE PARTIDA.....	126
V. A MODO DE CONCLUSIÓN	131

Acceso al código fuente y a los algoritmos de las Administraciones inteligentes. Lecciones a partir de experiencias comparadas	135
---	------------

María Estrella Gutiérrez David

I. INTRODUCCIÓN	135
II. LA DISCUTIDA NATURALEZA JURÍDICA DEL CÓDIGO FUENTE Y DE LOS ALGORITMOS PÚBLICOS. SU CALIFICACIÓN COMO “INFORMACIÓN PÚBLICA”	137
III. DETECCIÓN Y ESCRUTINIO PÚBLICO DE PATOLOGÍAS DE LA AUTOMATIZACIÓN Y ALGORITMIZACIÓN DE LA ACTIVIDAD ADMINISTRATIVA. EL INTERÉS PÚBLICO EN EL ACCESO.....	146
IV. EL PROBLEMA DE LOS ERRORES EN EL DISEÑO DE LOS ALGORITMOS DEL MIUR Y DE OFQUAL	149
V. EN PARTICULAR, IMPORTANCIA DE LA VALIDACIÓN DE LOS MODELOS ALGORÍTMICOS Y EL PROBLEMA DE LA AMPLIFICACIÓN DEL SESGO. DE NUEVO, OFQUAL	159
VI. SOBRE EL “VALOR IUSFUNDAMENTAL” DEL DERECHO A SABER QUÉ TRATAMIENTOS ALGORÍTMICOS APLICA LA ADMINISTRACIÓN. LECCIONES FINALES DE PARCOURSUP.....	163

Algoritmo y transparencia de la Administración pública en el ordenamiento jurídico italiano. Algunas reflexiones críticas a partir del caso “Buona Scuola”	169
<i>Salvatore Palumbo</i>	
I. EL CASO “BUONA SCUOLA”	169
II. USO DE SOFTWARE EN LA ADMINISTRACIÓN PÚBLICA Y PROTECCIÓN DE LA PROPIEDAD INTELECTUAL	172
III. ACCESO A LA INFORMACIÓN PÚBLICA Y LENGUAJE DE PROGRAMACIÓN	178
IV. LA TRANSPARENCIA COMO PRINCIPIO FUNDAMENTAL DE LA ACCIÓN ADMINISTRATIVA	181
V. ACCESO A LA INFORMACIÓN DE LA ADMINISTRACIÓN PÚBLICA Y PROCEDIMIENTO ALGORÍTMICO: DESDE LA LEY N°241 DE 1990 HASTA EL DECRETO LEGISLATIVO N°97 DE 2016	188
VI. CONCLUSIONES PROVISIONALES	195

Bloque III

TRANSPARENCIA Y EXPLICABILIDAD EN LA EMPRESA, LA CONTRATACIÓN Y PRESTACIÓN DE SERVICIOS DE INTERNET Y EN EL ÁMBITO LABORAL

Explicabilidad e inteligencia artificial responsable. Recomendaciones “aterrizadas” al día a día de las empresas	199
<i>Juan Manuel Belloto Gómez</i>	
I. A MODO DE INTRODUCCIÓN	199
II. EXPLICABILIDAD E INTELIGENCIA ARTIFICIAL RESPONSABLE	200
1. El porqué de la necesidad de una IA responsable, ética y normativa	200
2. La explicabilidad. Contexto y su relación con la IA responsable	201
3. Iniciativas que ejemplifican la necesidad de una IA responsable	202
4. El necesario papel de entidades independientes como OdiseIA	204
5. La necesidad de aterrizar los conceptos de la IA responsable... ..	205
6. Un ejemplo de aterrizaje. La iniciativa GuIA de OdiseIA	206
7. La explicabilidad es solo uno de los principios éticos necesarios en la IA	209
III. EXPLICABILIDAD. RECOMENDACIONES ATERRIZADAS AL DÍA A DÍA DE LAS EMPRESAS	211
1. Recomendaciones de Google, Microsoft, IBM y Telefónica	212

1.1.	Visión sectorial	212
1.2.	Explicar el modelo y también los datos que maneja	213
1.3.	Explicar lo global y lo particular	213
1.4.	Adaptar la explicación al interlocutor	214
1.5.	En todo el ciclo de vida de la solución inteligente.....	215
1.6.	Menos puede ser más	216
1.7.	Caja negra vs caja blanca	216
1.8.	Explicaciones claras y prudentes	217
1.9.	Usando causalidad, minimizando correlaciones.....	218
1.10.	La importancia del modelo de gobierno	218
2.	Recursos, tecnologías y modelos de Gobierno para gestionar-la.....	219
3.	Relación de la explicabilidad con otros principios para una IA responsable	221
IV.	EPÍLOGO.....	223

La moderna configuración de la transparencia en la prestación de servicios y en la contratación en Internet..... 225

Javier Plaza Penadés

I.	LA TRANSPARENCIA COMO VALOR Y COMO PRINCIPIO GENERAL.....	225
II.	PRINCIPIO DE TRANSPARENCIA Y PROTECCIÓN DE DATOS	226
III.	“LOS DERECHOS FUNDAMENTALES DE LA PERSONA PREVALECE SOBRE LA TECNOLOGÍA”.....	228
IV.	LA TRANSPARENCIA EN INTERNET	229
1.	Transparencia en la contratación con condiciones generales....	231
2.	Transparencia en los servicios de intermediación en línea.....	242
3.	Protección frente a los patrones oscuros “dark patterns”.....	243
V.	CONCLUSIONES	247

Transparencia algorítmica en el ámbito laboral..... 249

Adrián Todolí Signes

I.	EL USO DE LOS ALGORITMOS EN EL TRABAJO	249
II.	EL DEBER DE TRANSPARENCIA ALGORÍTMICA EN EL ÁMBITO LABORAL.....	253
1.	Derecho de información de los representantes de los trabajadores sobre los algoritmos: el nuevo art. 64.4 ET	253
2.	Concepto de algoritmo.....	256
III.	COMPARATIVA ENTRE LOS DERECHOS DE INFORMACIÓN DE ESTATUTO Y LA REGULACIÓN EUROPEA.....	257
IV.	CONTENIDO DE LA INFORMACIÓN A ENTREGAR.....	259

V. DERECHOS DE INFORMACIÓN YA EXISTENTES QUE AFECTAN A LOS ALGORITMOS.....	263
VI. CONCLUSIONES: TRASCENDENCIA DEL NUEVO DERECHO DE INFORMACIÓN.....	264

Presentación

Ciertamente es un orgullo haber liderado y coordinado la mejor monografía en español sobre la transparencia y la explicabilidad de los sistemas de inteligencia artificial (IA), un tema esencial para la confiabilidad y garantías de la tecnología disruptiva por excelencia en este siglo XXI que es el más invocado en todo documento o declaración sobre IA y que está en primera línea de las políticas de ética y regulación. Ya en 2018 con Andrés Boix tuvimos ocasión de lanzar el primer congreso en España sobre transparencia algorítmica¹ y la primera monografía sobre regulación de sistemas de IA². Y es un placer académico formar parte de un equipo líder en España en la investigación jurídica de la IA, precisamente en la Facultad de Derecho de la Universidad de Valencia, la primera en España en el famoso ránking de Shanghai en 2022. También cabe señalar que los coordinadores de esta obra y casi la mitad de sus autores, formamos parte de OdiselA (Observatorio del Impacto Ético y Social de la Inteligencia Artificial), la asociación más importante en la materia en España y en la que Lorenzo Cotino coordina las áreas jurídicas y, como se expone, Juan Manuel Belloto dirige la emblemática y pionera iniciativa GuIA, que tanto tiene que ver con la transparencia algorítmica.

Se cuenta en esta ocasión con un panel de lujo. Nueve autores, siete de ellos catedráticos o profesores de la universidad española, ocho doctores. La perspectiva jurídica es la que prima, no en vano, siete de los autores son expertos en Derecho constitucional, administrativo, laboral o civil. En todo caso, se cuenta con la especial aportación de dos expertos de otras disciplinas (Alfred Peris Manguillot, Catedrático de Matemàtica Pura i Aplicada y Juan Manuel Belloto Gómez,

¹ *Transparencia digital, algorítmica y big data. V Congreso Internacional del avance del Gobierno Abierto*. III Congreso de Buen gobierno y transparencia de la Comunitat Valenciana, <http://www.derechotics.com/congresos/2018-transparencia-big-data>

² BOIX PALOP, Andrés y COTINO HUESO, Lorenzo (coords.), *Monográfico Derecho Público, derechos y transparencia ante el uso de algoritmos, inteligencia artificial y big data RGDA Iustel*, n° 50, febrero 2019. Acceso en https://www.iustel.com/v2/revistas/detalle_revista.asp?id=1

Director de la iniciativa GuIA, pionera en España). Ello enriquece sin duda este libro.

La obra se ha distribuido en tres bloques respecto de la transparencia y explicabilidad de la inteligencia artificial. Los elementos conceptuales, generales y de género (I) las garantías y particularidades de la transparencia y explicabilidad en el uso de la inteligencia artificial por el sector público (II) y, en el ámbito privado, el impulso de la transparencia y la explicabilidad sociedad civil y su proyección en la empresa, en la contratación y prestación de servicios de internet y en el ámbito laboral.

Así, Lorenzo Cotino, uno de los coordinadores y de quienes suscriben estas palabras, inicia la obra con un estudio en el que se pretende asentar el concepto de la transparencia y explicabilidad de la inteligencia artificial. Para ello se expone cómo es el principio unánimemente reconocido en los documentos de la llamada *ética de la AI*, así como se describen las (insuficientes) las obligaciones normativas que imponen la misma en España. Asimismo, se reflexiona y se delimita el contenido de la transparencia en razón de sus finalidades. Para ello es esencial cuestionarse el para qué, para quién y cuánta transparencia algorítmica. Sobre estas premisas, el grueso de este estudio que inicia el libro es adentrarse en los conceptos de transparencia y explicabilidad y, con ellos, toda la y “compañía” de nociones que orbitan e integran estas nociones y amplían su campo denotativo. Así, se exponen —con apoyo de los documentos de referencia— los conceptos de comunicación, interpretabilidad, inteligibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad y otros más de interés a la materia.

Alfred Peris Manguillot, Catedrático de Matemàtica Pura i Aplicada, se cuestiona si podemos hacer los algoritmos transparentes y trazables en su proceso?, para ello explica el concepto de algoritmo, las redes neuronales y aprendizaje profundo y sus capas más ocultas y el enorme reto matemático que supone lograr la comprensión del funcionamiento de estos sistemas. Este primer bloque general lo cierra Raquel Valle Escolano cuando aborda los riesgos, pero también oportunidades que implica la IA para la igualdad, las respuestas frente a los mismos desde el punto de vista ético y jurídico. Sobre estas bases

se señala a la transparencia como problema de fondo en el que se pueden lograr las más importantes respuestas.

El segundo bloque está centrado en las particularidades del sector público y se tiene la suerte de contar con quienes más y mejor han trabajado este tema en España. El mismo lo inicia Gabriele Vestri, quien ya ha publicado excelentes estudios sobre la transparencia algorítmica y la contratación pública de algoritmos, en esta obra se centra en el derecho de acceso a la información pública que suponen los sistemas de IA e incluso el acceso a su código fuente. La polémica sentencia de 2021 que ratifica la previa denegación del CTBG estatal al código fuente del sistema BOSCO es criticada, además de analizado todo el marco jurídico aplicable con especial atención del fenómeno de los sesgos de los sistemas IA. Se propone incluso ir más allá de facilitar el acceso al código, sino dar acceso al sistema funcional mismo. María Estrella Gutiérrez trabaja desde hace años muy exhaustivamente el derecho de acceso a la información pública en clave nacional y comparada, así como es autora del trabajo más prolijo sobre transparencia algorítmica, con un cononcimiento técnico que no es nada habitual entre juristas³. Pues bien, en esta ocasión ofrece el trabajo de Derecho comparado más profundo hasta la fecha y del que todos debemos aprender. El mismo permite al lector adentrarse en la naturaleza del código fuente y de los algoritmos, las posibilidades de que la ciudadanía detecte y control las posibles patologías del uso público de sistemas IA. Descubir posibles sesgos es un elemento de gran relevancia a tener en cuenta para considerar, incluso, la naturaleza de derecho fundamental del acceso a esta información administrativa. Y en la misma línea del excelente trabajo de Gutiérrez cabe calificar la aportación del doctor italiano Salvatore Palumbo, a quien he tenido el gusto de codirigir su tesis doctoral sobre el uso público de IA. Su estudio se adentra en la transparencia del algoritmo público italiano a partir de un interesante supuesto de aquel país.

³ M.E. Gutiérrez David, “Administraciones inteligentes y acceso al código fuente y los algoritmos públicos. Conjutando riesgos de cajas negras decisionales”, en *Derecom*, n.º 31, págs. 19-105, <http://www.derecom.com/derecom/>

El tercer y último bloque del libro se dedica a la Transparencia y explicabilidad en el contexto empresarial, la contratación y prestación de servicios de internet y en el ámbito laboral. Juan Manuel Belloto es el responsable del proyecto *GuIA*, la mayor iniciativa desde la sociedad civil en España sobre el uso responsable de los sistemas de IA a través de la mencionada asociación OdiseIA. Desde 2021, Lorenzo Cotino ha tenido la suerte de coordinar la parte jurídica de la primera entrega de este ambicioso proyecto en 2022. En su estudio se expone esta iniciativa, el necesario papel de entidades independientes como OdiseIA y cómo “aterrizar” los principios éticos de la IA de modo concreto y tecnológicamente en el seno de las empresas. A la vista de su directo conocimiento en el proyecto *GuIA*, se detallan las recomendaciones y acciones de las grandes tecnológicas (*Google*, *Microsoft*, *IBM* y *Telefónica*) para hacer efectivos estos principios.

Javier Plaza en su estudio hace un recorrido la sobre moderna configuración de la transparencia en la prestación de servicios y en la contratación en Internet, centrándose en la transparencia en la contratación con condiciones generales y en los servicios de intermediación en línea, así como la Protección frente a los patrones oscuros “dark patterns”. Se subraya la importancia del Reglamento 2019/1150 de la Unión Europea, que ciertamente es bastante desconocido para la mayoría de los juristas. Y cierra esta obra Adrián Todolí Signes, que es la persona de referencia académica en España sobre los sistemas de IA y algoritmos en el ámbito laboral. De hecho, ha participado en la comisión de expertos que en mayo de 2022 ha realizado la *Guía práctica y herramienta sobre la obligación empresarial de información sobre el uso de algoritmos en el ámbito laboral*⁴. Así las cosas, su contribución revisa la trascendencia creciente del uso de los algoritmos en el trabajo, el deber de transparencia algorítmica en el ámbito laboral y en particular el derecho de información de los representantes de los trabajadores sobre los algoritmos en razón del nuevo art. 64.4 Estatuto de los Trabajadores (en la llamada *ley rider*). Todolí se detiene en elementos relevantes como concepto de algoritmo a los efectos de esta normativa, así como se efectúa una comparativa con la regulación

⁴ Ministerio de Trabajo y Economía Social, *Información algorítmica en el ámbito laboral*, mayo 2022, https://www.lamoncloa.gob.es/serviciosdeprensa/notas-prensa/trabajo14/Documents/2022/100622-Guia_algoritmos.pdf

europea, se examina cuál es la información que se debe facilitar para una reflexión final sobre la gran importancia de este derecho de información en el contexto laboral. Se cierra así esta monografía que ha de ser la de referencia en nuestro país y en español en la materia y que es ciertamente un honor haber podido impulsar y coordinar.

Esta publicación ha sido realizada con el apoyo económico de la Conselleria de Participación, Transparencia, Cooperación y Calidad Democrática. El contenido de la publicación es responsabilidad exclusiva de la Universitat de Valencia y no representa necesariamente a la Generalitat Valenciana.

Los coordinadores queremos agradecer en particular a Joaquín Martín Cubas su liderazgo y continuo apoyo desde la dirección de la Cátedra PAGODA (Cátedra de Gobierno Abierto, Participación y Open Data) en la Facultad de Derecho, con el constante apoyo de la Consellería de Participación, Transparencia, Cooperación y Calidad Democrática. Se trata precisamente de una Consellería fuertemente implicada con la transparencia algorítmica. Baste señalar la más que novedosa regulación de 2022 que va a suponer las más altas cotas de publicidad activa sobre algoritmos e inteligencia artificial pública no sólo en España sino en toda Europa. Joaquín Martín, Andrés Boix y quienes suscriben hemos logrado conformar un equipo humano con decenas de actividades, seminarios y publicaciones. Es un orgullo presentar esta obra que pasar a ser ineludible en nuestro país gracias a la calidad de sus autores y sus estudios.

También resulta imprescindible mencionar los proyectos en el marco de los cuales esta obra ha llegado a buen puerto. Así, cabe mencionar los proyectos del Ministerio de Ciencia e Innovación, Retos “Derechos y garantías frente a las decisiones automatizadas...” (RTI2018-097172-B-C21) y en el ámbito valenciano el proyecto de excelencia “Algorithmic law”, Prometeo/2021/009, 2021-24 y el proyecto de investigación de grupos emergentes CIGE/2021/123 “Garantías, límites constitucionales y perspectiva ética ante la transformación digital: Big data, inteligencia artificial y robótica”, ambos de la Conselleria de Innovación, Universidad, Ciencia y Sociedad Digital de la Generalitat Valenciana.

De igual modo, cabe mencionar la importante iniciativa *ValGRAI: Valencian Graduate School and Research Network of Artificial Intelligence*, a la que pertenece uno de los coordinadores de la obra.

Lorenzo Cotino Hueso
Jorge Castellanos Claramunt
Valencia, Julio de 2022

Bloque I
Transparencia y explicabilidad
de la inteligencia artificial.
Elementos conceptuales,
generales y de género

Transparencia y explicabilidad de la inteligencia artificial y “compañía” (comunicación, interpretabilidad, inteligibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad...).

Para qué, para quién y cuánta

LORENZO COTINO HUESO¹
Catedrático de Derecho Constitucional
Universidad de Valencia

I. 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 transparencia². 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.

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, “Algorithmic law” (Prometeo/2021/009, 2021-24) Generalitat Valenciana” y estancia de personal investigador en empresa Generalitat Valenciana (AEST/2021/012). Asimismo, investigación realizada en el marco de *ValGRAI: Valencian Graduate School and Research Network of Artificial Intelligence*.

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

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 organizativo³. 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 principales 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 contextos⁵. 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

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, pp. 33-90, p. 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, pp. 797-838, pp. 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, pp. 293-304, en, p. 294 y a lo largo de todo su estudio con relación a dichas disciplinas.

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, 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 internacional⁷, 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, etc⁸..

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

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, pp. 1069-1100, en concreto, pp. 1078-1079.

⁸ *Ibidem*, p. 1079.

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 reconocerse¹¹. 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 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”.

⁹ P. Dyrberg, “El acceso público a los documentos y las autoridades comunitarias”, en *Revista de Derecho Comunitario Europeo*, n° 2, vol I, julio/diciembre de 1997, pp. 377-411, p. 377.

¹⁰ G. Arena, “Transparencia administrativa y democracia”, en *Revista Vasca de Administración Pública* n° 37, 1993, pp. 9-20, p. 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, pp. 15-18.

II. 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 84¹². 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 básicos¹³. 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 afecta-

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, pp. 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, p. 98.

Remite también a Hagendorff, T. “The Ethics of AI Ethics: An Evaluation of Guidelines”, n.º. 30 *Minds and Machines* 99., p. 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, p. 4, <https://dash.harvard.edu/handle/1/42160420>

dos y finalmente sólo un 3% con obligaciones de contratación pública abierta cuando se use por el sector público.

Como en otro lugar he analizado¹⁴, la Declaración del Parlamento UE¹⁵ 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 transparencia¹⁷ se añadiría a los otros cuatro principios básicos.

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, p. 10,

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.

III. 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 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 de¹⁹:

- la *accountability* y el principio democrático obligan a que se pueda controlar la actuación de programadores, funcionarios

<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, pp. 689-707, pp. 699-700, <https://doi.org/10.1007/s11023-018-9482-5>

¹⁸ HLEG, *Directrices éticas...*, cit., p. 10.

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, pp. 917-952, <http://links.uv.es/BmO8AU7>.

- y analistas de bajo nivel que “traducen” la ley en un código, muchas veces sin capacidad técnica ni jurídica para hacerlo²⁰;
- 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 datos²¹ y la doctrina²².
 - La transparencia también es esencial para poder controlar errores, discriminación y sesgo algorítmicos públicos y privados²³.

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, pp. 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: apro-

- Además, respecto del uso de la IA en el ámbito sancionador, policial y judicial²⁴, además, será especialmente importante la transparencia para garantizar la individualización del uso de la IA respecto del afectado y que se 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 UNESCO²⁵ (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 colectiva²⁶. 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 declaracio-

ximación 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.

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., p. 31.

nes y manifiestos que hemos hecho desde la Red DAIA²⁷ o los mejores trabajos en nuestro país como los del monográfico que tuve ocasión de coordinar con Boix²⁸, destacando muy especialmente respecto de la transparencia los trabajos de Gutiérrez²⁹, Vestri³⁰ y Cerrillo³¹, 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 funcionamiento del sistema, especialmente para comprender la cadena de causalidades³². En ocasiones, la única posibilidad de poder corregir un problema técnico del sistema o de los

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*, n° 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. Conjuran-do riesgos de cajas negras decisionales”, en *Derecom*, n° 31, pp. 19-105, ver pp. 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*, n° 56, Zaragoza, 2021, pp. 368-398.

³¹ A. Cerrillo i Martínez, “La transparencia de los algoritmos que utilizan las administraciones públicas”, *Anuario de Transparencia Local*, n° 3, 2020, pp. 41-78.

³² L. Ortiz de Zárate Alcarazo, “Explicabilidad (de la inteligencia artificial)”. En *Eunomia. Revista en Cultura de la Legalidad*, n° 22, 2022, pp. 328-344, p. 334 <https://doi.org/10.20318/eunomia.2022.6819>

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, n° 39), para la generación de confianza en los sistemas de IA³³. 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 Plan³⁵. 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.

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

³⁴ HLEG, *Directrices éticas...*, cit., p. 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*, n° 50, febrero 2019. Acceso en https://www.iustel.com/v2/revistas/detalle_revista.asp?id=1

IV. 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 consideren “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 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, etc³⁷.

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

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.

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 apartado 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.”

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.

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 caso, tendrá derecho a conocer los algoritmos que afectan a la toma de decisiones laborales⁴¹, 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.

V. ¿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 responsabilidad⁴². Y además se dificulta el control de calidad. Por el contrario,

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

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 Bajos⁴³, 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 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

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.

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 aspectos⁴⁴. 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 transparencia⁴⁶.

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.

⁴⁴ *Ibidem*, p. 29.

⁴⁵ <https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBD HA: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 n° 4, mayo 2020. <https://diariolaley.laleynext.es/Content/Documento.aspx?params=H4sIAAAAAA AAEAMtMSbF1CTEAAmMDSwNjM7Wy1KLizPw8WyMDIwMDEyNzkEB mWqVLfnJIZUGqbUIRaSoApoQJizQAAAA=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>

VI. 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, audibilidad, 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.

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 interpretabilidad 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 modelo⁴⁹.

⁴⁷ NIST, *AI Risk Management Framework...*, cit. apartado 5.3.3, pp. 12-13.

⁴⁸ *Ibidem*, apartados 5.2.2. pp. 10 y ss.

⁴⁹ *Ibidem* pp. 12 y ss., en concreto, p. 13.

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 pertinentes⁵⁰.

El HLEG acude a un concepto inclusivo⁵¹, *Transparencia = “Incluidas la trazabilidad, la explicabilidad y la comunicación”*. Su check list o lista de control de “Transparencia” incluye la *trazabilidad, 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.

VII. 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

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. pp. 18 y 22.

código fuente, cómo se entrena el algoritmo, así como los datos utilizados, las variables de entrada, los 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 algorítmica⁵³. 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 funcionamiento⁵⁴. 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* algorítmica⁵⁵:

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

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 pp. 55-56.

⁵⁴ *Ibidem*, p. 57.

⁵⁵ NIST, *AI Risk Management Framework...*, cit. apartado 5.2.2., “la interpretabilidad se puede diferenciar en simulabilidad, descomponibilidad y transparencia algorítmica.”

Descomponibilidad. Gutiérrez supone que los componentes del modelo (inputs, parámetros y cálculo) de pueden descomponer y permiten una explicación intuitiva⁵⁶.

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.

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

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, p. 90; Z. C. Lipton, “The Mythos of Model Interpretability”. *ACM Queue*, vol. 16, núm. 3 (mayo-junio), 2018, p. 12 o el ya referido ICO y Alan Turing Institute, *Explaining decisions made with Artificial Intelligence*, 2020, pp. 67-68 <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-ai/>

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 riesgo⁵⁷.

VIII. TRANSPARENCIA ¿PARA QUIÉN? TRANSPARENCIA “INTERNA” Y “EXTERNA”

1. *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 algoritmos⁵⁸ 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.

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

2. 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 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 general⁶⁰. Y en el caso

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*.

del futuro Reglamento de IA de la UE⁶¹, 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, 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).

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).

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 uso 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 interacción 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 riesgo⁶³ (art. 51 AIA) y todo el mundo puede acceder a información bastante básica que ahí consta.

3. 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

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.

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 Unido⁶⁵ 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.

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

⁶⁵ ICO (Information Commissioner 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 Commissioner 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., p. 10.

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 conocimiento⁶⁷. 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 IA⁶⁸.

IX. EXPLICABILIDAD. TIPOS Y ELEMENTOS PARA LOGRARLA

El NIST afirma que la explicabilidad va más allá de la transparencia, pero es vinculable con ella⁶⁹. 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 mejo-

⁶⁷ En general puede seguirse, AEPD, *Guía para el cumplimiento del deber de Informar*, AEPD, 2018, pp. 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., p. 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. pp. 13 y ss.

rar el aprendizaje humano, para depurar problemas con 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 impacantes que “debería garantizarse que se proporcione una explicación satisfactoria” (nº 40)⁷⁰.

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.

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).

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 a la 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

⁷¹ HLEG, *Directrices éticas...*, cit., p. 18.

⁷² *Ibidem*, p. 22.

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., p. 22.

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 Gutiérrez⁷⁵, 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.

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

⁷⁵ M.E., Gutiérrez David, “Administraciones inteligentes...”, cit., p. 55.

Ortiz de Zárate señala en general que las explicaciones tienen que ser lo más objetivas posibles y deben generar un consenso más o menos sólido⁷⁶. 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 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 IA⁷⁷. 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 Zárate⁷⁸ 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

⁷⁶ L. Ortiz de Zárate Alcarazo, “Explicabilidad (de la inteligencia artificial)...”, cit., p. 338.

⁷⁷ *Idem*, 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., p. 338.

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 organización⁷⁹.

X. INTERPRETABILIDAD, TRANSPARENCIA Y EXPLICABILIDAD

1. *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 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 ex-

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.

⁸⁰ *Ibidem*, n° 75.

plicabilidad, 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 transparencia⁸².

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 sistema⁸³.

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.

⁸¹ *Risk Management Framework...*, cit. apartado 5.2.1, p. 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., p. 334.

⁸⁴ Ministerie van Binnenlandse, *Impact Assessment...*, cit.

⁸⁵ *Ibidem*, 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>

2. 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 algoritmo⁸⁶. Para el organismo de referencia de EEUU, la 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 determinada⁸⁷.

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

⁸⁶ *Risk Management Framework...*, cit. apartado 5.2.2, p. 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.

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 *Ibidem*, 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”.

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 solo⁸⁹. 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 explicar⁹⁰.

XI. 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

⁸⁸ L. Ortiz de Zárate Alcarazo, “Explicabilidad (de la inteligencia artificial)...”, cit., p. 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*, p. 113.

⁹¹ HLEG, *Directrices éticas...*, cit., p. 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.”

de datos, validación y verificabilidad (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

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., p. 6.

⁹³ HLEG, *Directrices éticas...*, cit., p. 48. Respecto de auditabilidad ver lista de control, p. 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.”

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 públicos⁹⁷. 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 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.

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

⁹⁷ Directrices para el sector público, Rijksoverheid, *Richtlijnen voor het toepassen van algoritmen...*, cit., p. 26.

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.

XII. 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.

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.”

XIII. 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 “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.

⁹⁹ Directrices para el sector público, Rijksoverheid, *Richtlijnen voor het toepassen van algoritmen*..., cit., p. 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.

XIV. 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 humanos¹⁰². Siguiendo el mencionado estudio de referencia de Harvard, la comunicación se desdobra 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 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

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”.

“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 desdobra en un “principio de notificación” de la adopción de decisiones por un sistema de IA¹⁰⁴, 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 transparencia¹⁰⁶. 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.

¹⁰⁴ *Ibidem*, “Notification when Interacting with an AI.”

¹⁰⁵ *Idem*.

¹⁰⁶ HLEG, *Directrices éticas...*, cit.

También, en “comunicación” el HLEG incluye cuestiones variadas aunque relacionadas con la transparencia. Podrían agruparse 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.
- 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.

XV. 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 tam-

bién 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 afirmaciones¹⁰⁸.

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, etc¹⁰⁹.

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 por el sector público de sistemas de IA bien merece un análisis jurídico detallado¹¹¹.

¹⁰⁷ Fjeld, et. al., “Principled Artificial Intelligence..., cit., p. 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), p. 13 (principio 6.7.), entre otros.

¹¹⁰ Access Now, “Human Rights in the Age of Artificial Intelligence”, Access Now, 2018, n° 9, p. <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf>

¹¹¹ Para ello, entre otros G. Vestri, “Sistemas de inteligencia artificial en la contratación pública: entre códigos fuente y datos abiertos”, *Actualidad administrativa*, N° 12, 2021. También, J. Miranzo Díaz, “Inteligencia artificial y contratación pública”, en I. Martín (Dir.), *Administración electrónica, transparencia y contratación pública*, INAP, 2020, pp. 105-142.

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).

XVI. 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 unánimemente 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, importa-

dores, 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 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 “inter-na”, 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 denominación 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.

Algoritmos: ¿podemos hacerlos transparentes y trazables en su proceso?

ALFRED PERIS MANGUILLOT¹

*Catedrático. Institut Universitari de Matemàtica Pura i Aplicada
Universitat Politècnica de València*

I. EL CONCEPTO DE ALGORITMO

Posiblemente la primera idea que viene a la cabeza sobre lo que es un algoritmo es la de una máquina despiadada que, mediante la inteligencia artificial, dictamina determinadas conclusiones, que pueden ser percibidas como arbitrarias o sin justificación, y que cada vez en mayor grado van a afectar a nuestras vidas y a todo lo que nos rodea. Ciertamente buena parte de lo anterior es cierto, pero quedarse con esa idea es simplificar sobremanera lo que significan los algoritmos, así como el modo en que afectan a nuestra vida.

El concepto de algoritmo consiste en un conjunto de instrucciones o reglas definidas y no-ambiguas, ordenadas y finitas que permite, típicamente, solucionar un problema, realizar un cómputo, procesar datos y llevar a cabo otras tareas o actividades. Por tanto estamos ante una noción muy amplia, y a la vez muy antigua (el origen del propio término “algoritmo” lo podemos situar en el griego antiguo *arithmos*, aunque también influyó un nombre propio: Al-Khwarizmi, de Persia (780-850 DC)). Sin duda alguna un algoritmo no va a tener “piedad”, porque no está entre sus funciones (a no ser que de algún modo se programe para ello), y también es cierto que nuestras vidas van a ser altamente influenciadas por los algoritmos, que no olvidemos son programados por personas.

¹ Transcripción ampliada de mi intervención realizada sobre “Transparencia y auditabilidad de algoritmos” en la sesión del 26 de abril de 2022 del IX Congreso del avance del Gobierno Abierto de la Comunitat Valenciana (València, 26 y 27 de abril de 2022).

Por tanto, y ante lo irremediable del avance tecnológico y sus consecuencias, debemos prepararnos para generar unos algoritmos trazables, justos, y explicables en sus conclusiones y procesos.

Esto conlleva a día de hoy una cierta incompatibilidad: Los algoritmos de altas prestaciones para afrontar problemas de una gran complejidad utilizan herramientas de inteligencia artificial que no son trazables. La explicabilidad es un reto en los fundamentos matemáticos de la inteligencia artificial donde el objetivo es llegar a comprender cómo los modelos de aprendizaje automático llegan a las predicciones cuando se trata de explicar modelos complejos.

En definitiva, con los algoritmos nos encontramos ante la dicotomía de tener que elegir entre alta prestación pero nula o complicada explicabilidad, o sacrificar el rendimiento en pos de una mayor trazabilidad, explicabilidad y justicia. Por supuesto, esa dicotomía solo se presenta en problemas de gran complejidad, y según cual sea el objetivo de dichos problemas convendrá decantarse por un mayor acercamiento al rendimiento o a la trazabilidad.

Por ejemplo, en problemas de diagnosis médica “in vitro” posiblemente se prefiera una gran precisión en el diagnóstico con el uso de herramientas avanzadas de inteligencia artificial donde un gran número de variables entran en juego, y una amplia base de datos nutre el aprendizaje. Según el Instituto de Medicina de la Academia Nacional de Ciencias de Ingeniería y Medicina de EEUU “los errores de diagnóstico contribuyen a aproximadamente el 10 por ciento de las muertes de los pacientes” y también representan del 6 al 17 por ciento de las complicaciones hospitalarias. Es importante tener en cuenta que el desempeño del médico generalmente no es la causa directa de los errores de diagnóstico. De hecho, los investigadores atribuyen la causa de los errores de diagnóstico a una variedad de factores que incluyen: colaboración e integración ineficientes de las tecnologías de la información en salud, fallos de comunicación entre los médicos, los pacientes y sus familias, y un sistema de trabajo sanitario que, por diseño, no es adecuadamente favorable al proceso diagnóstico.

Para brindar un contexto adicional, una revisión de 25 años de pagos de reclamaciones por negligencia en EEUU realizada por investigadores del Hospital Johns Hopkins mostró que las reclamaciones por errores de diagnóstico tenían una mayor incidencia en entornos

de pacientes ambulatorios (68.8 por ciento) que en pacientes hospitalizados (31.2 por ciento). Sin embargo, los que ocurrieron en un entorno hospitalario tenían aproximadamente un 11,5 por ciento más de probabilidades de ser letales. Los pagos totales durante el período de 25 años ascendieron a una suma sustancial de casi 39.000 millones de dólares. Para abordar estos desafíos, muchos investigadores y empresas están aprovechando la inteligencia artificial para mejorar los diagnósticos médicos².

Cuando tratamos problemas complejos que afectan a derechos (permisos penitenciarios, riesgo de comisión de delito, etc.) la situación es mucho más delicada y requiere de la búsqueda de un equilibrio adecuado entre rendimiento y trazabilidad/justicia (sacrificando prestaciones algorítmicas) en el cual entraremos con más detalle en la siguiente sección.

Volvamos a la amplitud del concepto de algoritmo, y más concretamente a aquellos algoritmos que, o bien por la simplicidad del problema, o bien por un diseño adaptado al problema que se va a tratar, no requieren de fases de “caja negra” o aprendizaje profundo. Pensemos, por ejemplo, cómo actúa un simple termostato (de frío o calor) donde el algoritmo hace que el aparato se active únicamente cuando estamos fuera del umbral de temperatura deseado. Entre esa situación tan simple y los algoritmos que precisan de herramientas avanzadas de inteligencia artificial, con sus fases “ocultas”, hay un amplísimo margen donde el foco debe ponerse en la transparencia de los procesos de ejecución y de las ponderaciones utilizadas, si fuera el caso. Este tipo de algoritmos son los que vamos a llamar “deterministas”, y son aquellos donde (si se conocen los elementos suficientes) es posible su trazabilidad completa, podemos explicar cada paso de su ejecución, y a igualdad de “inputs” producen el mismo “output”. Esta definición no es completamente rigurosa, pero sacrificamos algo de rigor para una mejor visualización del concepto. Ahora bien, que exista la posibilidad de todo lo anterior en un algoritmo determinista no significa que sea público o auditable. Por tanto se debe enfatizar el uso

² D. Faggella, “Machine Learning for Medical Diagnostics - 4 Current Applications”, 2020. <https://emerj.com/ai-sector-overviews/machine-learning-medical-diagnostics-4-current-applications/>

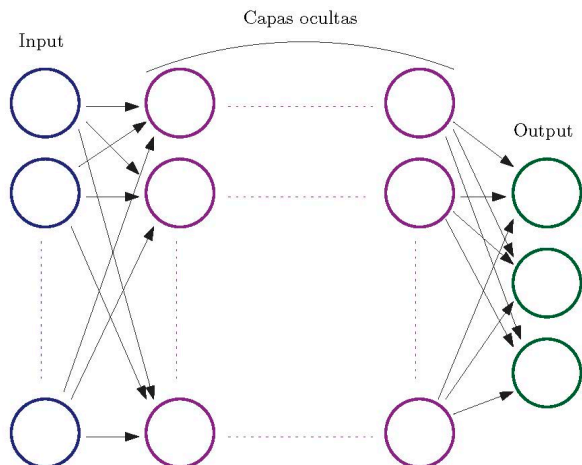
de algoritmos deterministas, y a la vez que quienes los utilizan (administraciones públicas, etc.) pongan todos los elementos que permitan su auditabilidad de forma transparente. Esto puede suponer a veces un conflicto con la protección de datos si no se hace adecuadamente, pero tampoco tiene que ser la excusa para no hacerlo, trabajando en la anonimización conveniente para cumplir con ambos aspectos: Auditabilidad y protección de datos.

II. REDES NEURONALES Y APRENDIZAJE PROFUNDO

En esta sección intentaremos dar una visión de las redes neuronales y el aprendizaje automático profundo. Una red neuronal artificial consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales. La información de entrada (input) atraviesa la red neuronal (donde se somete a diversas operaciones) produciendo unos valores de salida (output). Algunas características generales de una red neuronal son las siguientes.

- Aprenden y se forman a sí mismas, en lugar de ser programadas de forma explícita.
- Para el aprendizaje automático se minimiza una función de coste.
- Se entrenan con unos modelos o ejemplos.
- El nivel de profundidad de la red, como por ejemplo el número de capas ocultas, determina si entramos en lo que se conoce como aprendizaje profundo.

En la siguiente figura mostramos una red neuronal artificial con varias capas ocultas:



Las redes neuronales de aprendizaje profundo, que involucran varias capas ocultas, son las que ofrecen una mayor dificultad (por no decir imposibilidad) de ser trazables. Aquí hay que tener en cuenta que, a diferencia de los algoritmos deterministas, igualdad de inputs no tienen por qué producir el mismo output. Pongamos, por ejemplo, una máquina que utiliza una red neuronal de aprendizaje automático jugando al ajedrez o al go (juego que se origina en China hace más de 2500 años y se consideró una de las cuatro artes esenciales de la antigüedad China). Ante posiciones idénticas del adversario la misma máquina puede proponer distintas jugadas de respuesta. Esto ocurre porque, en el proceso de minimización de la función coste a través de las capas ocultas, llega a una solución mediante un camino desconocido, que es lo suficientemente buena y de un alto nivel de profundidad, pero no tiene por qué ser la mejor jugada posible (es lo que llamaríamos llegar a un mínimo local de la función de coste que es muy bueno, pero no tiene por qué ser un mínimo absoluto). El go es uno de los juegos más complejos que existen, incluso más que el ajedrez, utilizando unas reglas muy simples. En 2016, *AlphaGo* de Google, desarrollado por *DeepMind*, fue el primer programa informático basado en redes neuronales de aprendizaje profundo capaz de derrotar a un campeón mundial humano en go. Cabe tener en cuenta que este avance en el campo de la inteligencia artificial llegó casi 20 años después de que los grandes maestros de ajedrez, como Kasparov, fueran derrotados

por programas como Deep Blue (IBM). En 2017 se presentó *AlphaGo Zero*, una versión sin datos humanos y más sólida que cualquier otra versión humana anterior que derrotara al campeón de go. Su entrenamiento consistió en jugar juegos contra sí mismo, de manera que *AlphaGo Zero* superó con creces a la primera versión de *AlphaGo* en tres días al ganar 100 juegos a 0. *AlphaGo* fue entrenado inicialmente para imitar el juego humano, tratando de igualar los movimientos de los jugadores expertos de juegos históricos registrados, utilizando una base de datos de alrededor de 30 millones de movimientos. Una vez que había alcanzado un cierto grado de habilidad, era entrenado aún más al realizar un gran número de partidas contra otras instancias de sí mismo, usando aprendizaje por refuerzo para mejorar su juego. La versión posterior *AlphaZero* de *DeepMind* ya era capaz de derrotar a los mejores programas informáticos de ajedrez, go y shogi. Por ejemplo, *AlphaZero* dominó completamente en un torneo de ajedrez a *Stockfish* después de solo 4 horas de autoaprendizaje, sin acceso a libros de apertura o bases de datos de tablas de finales. Una característica de este tipo de aprendizaje profundo es que no es entrenado de forma humana por grandes jugadores (como sí se hacía con *Deep Blue* con grandes maestros del ajedrez) que introduzcan jugadas respuesta a determinadas posiciones, o aperturas conocidas, sino que es puro autoaprendizaje. De esta manera genera nuevas jugadas que nunca habían sido utilizadas. Dicho de otro modo, es creativo.

Dejamos a un lado los juegos de mesa, y volvemos a las aplicaciones de la inteligencia artificial en campos prácticos, como la predicción de riesgos en préstamos bancarios. Prever el riesgo crediticio es el principal motivo de preocupación del sector bancario en la mayoría de los países del mundo. Este riesgo de crédito surge cuando el prestatario no devuelve el dinero obtenido. El prestamista utiliza agencias de crédito que investigan el historial crediticio del consumidor, aunque las desviaciones en el comportamiento del consumidor y la situación económica han hecho que la información de los departamentos de crédito no sea fiable en lo que respecta al comportamiento futuro del consumidor. Cada vez más se utilizan herramientas avanzadas de inteligencia artificial que analizan inmensas bases de datos y, para un consumidor concreto, ofrecen un diagnóstico de riesgo crediticio. El problema lo volvemos a encontrar en que estas herramientas de inteligencia artificial no ofrecen una explicación de las razones que

llevan a denegar o no un crédito a una determinada persona. Se puede producir un trato desigual, un impacto desigual, o ambos. Como trato desigual entenderíamos que la decisión que recibe una persona sobre un préstamo cambia con su información sobre determinado atributo sensible, y por impacto desigual cuando el resultado de la decisión beneficia o perjudica desproporcionadamente a los miembros de ciertos grupos de valores de atributos sensibles (por ejemplo, ser funcionario o no serlo).

Otra aplicación de la inteligencia artificial que está experimentando un gran desarrollo es la del reclutamiento de talentos por parte, especialmente, de grandes empresas. Según la Oficina de Estadísticas Laborales de EEUU durante la década que va de 2016 a 2026, la fuerza laboral en EEUU aumentará en 11,5 millones. Este aumento equivale aproximadamente a una tasa de crecimiento anual del 0,7 por ciento. Sin embargo, junto con este crecimiento anticipado, se espera que ciertas industrias experimenten desafíos de contratación. Por ejemplo, según una encuesta de CompData de 2016 de 30000 organizaciones, las industrias de la hostelería, la atención médica y la banca y las finanzas experimentaron las tasas de rotación laboral más altas, que oscilan entre el 18 y casi el 30 por ciento. Hoy en día, el aprendizaje automático está emergiendo como una estrategia para ayudar a los empleadores a realizar de manera más eficiente la búsqueda y contratación de talentos. La adquisición exitosa de talento es una tarea exigente que requiere la capacidad de descubrir y evaluar las habilidades y experiencias de las candidaturas. La inteligencia artificial está participando en una nueva era impulsada por la tecnología de reclutamiento y contratación al aumentar la capacidad y ahorrar tiempo a los empleadores. Las herramientas de contratación de inteligencia artificial que utilizan análisis predictivos para recomendar candidatos pueden poner a los usuarios por delante de su competencia. El análisis predictivo también puede resultar útil para la retención de buenas contrataciones. Por ejemplo, la capacidad de determinar la probabilidad de que una candidatura encaje en el trabajo de una empresa al principio del proceso de contratación podría mejorar las posibilidades de que esa persona permanezca en la empresa. A pesar del potencial de la automatización para mejorar la eficiencia y la capacidad de adquisición de talento, existen algunos desafíos a considerar. Para ofrecer valor, la inteligencia artificial requiere acceso a grandes

conjuntos de datos de calidad. Por lo tanto, es posible que la tecnología no resulte tan útil para las empresas que revisan un conjunto más pequeño de currículos o para aquellas que no contratan regularmente. Además, algunas empresas esperan que la inteligencia artificial pueda eliminar los sesgos en el proceso de contratación. Sin embargo, como hemos comentado, es importante señalar que las herramientas de aprendizaje automático también son capaces de aprender los sesgos humanos. Dado que cada empresa es única, hay varios factores que deben tenerse en cuenta al evaluar los beneficios de integrar la inteligencia artificial en el proceso de selección y contratación³.

III. INTELIGENCIA ARTIFICIAL Y JUSTICIA

Si el impacto de la inteligencia artificial y los algoritmos en la justicia es un hecho, y al parecer va a ser cada mayor, parece claro que hay que trabajar en que dicho impacto, cuyo objetivo es la eficiencia, vaya de la mano de la explicabilidad y que se eviten sesgos en los datos que nutren los algoritmos así como en sus procesos.

Si nos fijamos en COMPAS, que es una de las varias herramientas de inteligencia artificial para evaluación de riesgo que se usan en el sistema de justicia penal de EEUU, destacamos a continuación algunas características que lo definen:

- Se entrena con los datos históricos de los acusados.
- Luego predice la probabilidad de que un acusado sea detenido por un otro delito durante el período de espera del juicio.
- Los jueces no siempre siguen estas recomendaciones, pero las evaluaciones de riesgo resultan influyentes.
- Por ley la raza no se incluye en la toma de datos, pero un estudio previo⁴ demuestra a la vista de los resultados que sí influye.

³ K. Sennaar, “Machine Learning for Recruiting and Hiring – 6 Current Applications”, 2019. <https://emerj.com/ai-sector-overviews/machine-learning-for-recruiting-and-hiring/>

⁴ J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “Machine bias risk assessments in criminal sentencing”, *ProPublica*, May, 23, 2016.

Esto motivó que diversos grupos de investigadores intentaran generar algoritmos que mejoraran COMPAS en los aspectos de eliminar sesgos de raza, a costa de sacrificar el rendimiento. Un equipo del *Massachusetts Institute of Technology* intentó mejorar COMPAS con el mismo conjunto de datos que *ProPublica* utilizó en su análisis⁵. A partir de más de 7.200 perfiles con nombre, edad, raza y la puntuación de riesgo de COMPAS, se muestrearon aleatoriamente a 500 acusados blancos y negros. Observaron que, si el umbral de alto riesgo era un 7/10, en primer lugar cabía mejorar el umbral. Esto falló, incluso utilizando la proporción de *Blackstone* (el 57% de los liberados fueron rearrestados, y el 17% fueron encarcelados innecesariamente). Además concluyeron que, aunque se pudiera mejorar el sistema predictivo, al final alguien tiene que decidir dónde poner el umbral de alto riesgo. Analizaron entonces con más detalle la muestra de 500 acusados y, a pesar de que la raza es un ejemplo de una categoría protegida en Estados Unidos, lo que significa que la discriminación sobre esa base es ilegal, sin embargo sobre aquellos que fueron liberados un 65% de blancos fueron posteriormente rearrestados, mientras que el porcentaje fue solo de un 32% en los negros. Por otra parte, observando los porcentajes de los encarcelados innecesariamente (es decir, aquellos que dieron un alto perfil de riesgo, eran encarcelados a espera de juicio, y posteriormente declarados no culpables) resultó ser de un 11% en blancos y de un 32% en negros. La única solución posible era plantear distintos umbrales para blancos y negros lo cual significaba serios problemas legales. El problema raíz, en definitiva, era que los algoritmos de aprendizaje automático se entrenan con datos generados con historiales de exclusión y discriminación. ¿Significa esto que hay que abandonar toda esperanza de generar algoritmos a la vez eficientes y libres de sesgos en ámbitos de aplicación como el de COMPAS? Esto puede generar controversia, pero las dificultades inherentes no deberían ser un muro infranqueable que aparte todo intento de mejorar COMPAS. Se está trabajando en metodologías que alcancen un adecuado equilibrio entre precisión y justicia/reducción

⁵ K. Hao and J. Stray, "Can you make AI fairer than a judge? Play our courtroom algorithm game", *MIT Technology Review*, October, 17, 2019.

de sesgos⁶. Aunque es complicado describir aquí los aspectos técnicos de dichos algoritmos, sí que podemos apuntar que introducen en aquellas categorías sensibles un parámetro que pondera entre precisión y sesgo, y que calcula para distintos valores del parámetro la precisión, el grado de injusticia, y como afecta a los grupos de categoría sensible protegida (no blancos) y no protegida (blancos).

También cabe resaltar que no todos los algoritmos de inteligencia artificial que se generan en el ámbito judicial padecen de sesgos relevantes, y posiblemente el software *Prometea* de Argentina sea un buen ejemplo de ello. Algunas de sus características son las siguientes:

- Utiliza inteligencia artificial, pero no aprendizaje profundo.
- Se utiliza fundamentalmente para casos contencioso-administrativos (solicitudes de vivienda digna, bonificaciones en empleo público, ejecución de multas no pagadas, denegación de licencias de taxi por antecedentes penales, etc.).
- Está permitiendo la descongestión judicial, fundamentalmente para casos más automatizables y que se precise una uniformidad en las decisiones.

Esto vuelve a incidir en cuestiones planteadas anteriormente: si podemos generar algoritmos que sean trazables, que no conlleven pasos intermedios de “caja negra”, y que estén libres de sesgos (o que estos sean detectables y se puedan corregir), parece aconsejable seguir trabajando en ello⁷.

IV. ALGORITMOS MÁS ALLÁ DE LA INTELIGENCIA ARTIFICIAL

Vamos a centrarnos en algoritmos que no utilizan inteligencia artificial, que son del tipo determinista, y que incluso matemáticamente son bastante simples: Esencialmente medias ponderadas donde

⁶ ver, e.g., E. Carrizosa, M. Galvis-Restrepo, and D. Romero-Morales, “On clustering categories of categorical predictors in generalized linear models”, *Expert Systems with Applications*, 2021, p. 115245.

⁷ J. G. Corvalán, “Inteligencia artificial: retos, desafíos y oportunidades-Prometea: la primera inteligencia artificial de Latinoamérica al servicio de la justicia”, *Revista de Investigações Constitucionais*, v. 5, n. 1, 2018, pp. 295-316.

distintos factores y variables determinan el resultado final, cada uno de ellos tiene su peso, e incluso podemos agruparlos en distintas dimensiones para extraer información concreta de los resultados por dimensiones. A modo de ejemplo consideraremos el Visor de Espacios Urbanos Sensibles (VEUS), que responde a un encargo de la Conselleria d'Habitatge i Arquitectura Bioclimàtica realizado en el Instituto Valenciano de la Edificación y la Universitat Politècnica de València para definir políticas que evalúen y prioricen las intervenciones necesarias en los entornos urbanos⁸. En el VEUS se otorgan unas valoraciones a las secciones censales de la Comunitat Valenciana utilizando medias ponderadas. Incluye una colección de cuatro factores y nueve variables que son evaluadas de 0 a 3 para cada sección censal. Cabe destacar que todo es absolutamente público (pesos de cada factor y variable, criterios de valoración), y está basado en informes expertos y en datos facilitados fundamentalmente por el Instituto Nacional de Estadística. La fórmula que ofrece el índice de vulnerabilidad de cada sección censal viene dada por:

$$IV(i) = \left(\frac{100}{3}\right) * \left(\frac{PF * \left(\frac{\sum_j pf_j * f_j(i)}{\sum_j pf_j} \right) + PV * \left(\frac{\sum_k pv_k * v_k(i)}{\sum_k pv_k} \right)}{PF + PV} \right)$$

Donde,

- $IV(i)$ es el índice de vulnerabilidad de la sección censal “i”.
- PF es el peso global de los factores considerados.
- PV es el peso global de las variables consideradas.
- pf_j es el peso del factor j-ésimo.
- pv_k es el peso de la variable k-ésima.
- $f_j(i)$ es el valor que toma el factor j-ésimo en la sección censal “i”.
- $v_k(i)$ es el valor que toma la variable k-ésima en la sección censal “i”.

⁸ B. Serrano, R. Temes, A. Peris, A. Moya, R. De León, B. Martín y B. Menéndez, “Visor de Espacios Urbanos Sensibles de la Comunitat Valenciana 2020”, Generalitat Valenciana, Abril de 2020. <https://calab.es/observatorio-del-habitat/veus/>

Cada uno de los factores y variables se sitúa entre 0 y 3, de manera que el resultado final de índice de vulnerabilidad está entre 0 y 100. Además, se pueden agrupar variables en tres dimensiones: Residencial, Socioeconómica y Sociodemográfica, lo cual permite analizar situaciones de vulnerabilidad de forma separada en cada una de esas dimensiones y así establecer tipos de vulnerabilidad, siendo la máxima (vulnerabilidad integral) cuando una determinada sección censal supera el percentil 66 (está en el primer tercio de vulnerabilidad) en cada una de las tres dimensiones. Esto facilita poder visualizar fácilmente las áreas donde urge más intervenir. Los datos son volcados al Visor Cartográfico de la Generalitat Valenciana⁹, que ofrece un análisis muy detallado.

Lo descrito en el tipo de algoritmo utilizado en el VEUS es algo muy común en diversas instancias públicas que se apoyan en algoritmos basados en medias ponderadas, como es el caso del Sistema de Seguimiento Integral en los casos de Violencia de Género (Sistema VioGén), de la Secretaría de Estado de Seguridad del Ministerio del Interior, cuyo objetivo es integrar toda la información de interés que se estime necesaria para hacer una predicción del riesgo de violencia de género y, atendiendo al nivel de riesgo, realizar seguimiento y protección a las víctimas en todo el territorio nacional. Aquí las ponderaciones utilizadas en las variables que integran el algoritmo no son públicas. Otro ejemplo, en este caso en Cataluña, es el algoritmo RisCanvi, que calcula el riesgo de reincidencia de los presos y sirve de apoyo para la concesión de permisos penitenciarios. En este caso hay mayor información pública en lo que a datos se refiere.

V. CONCLUSIONES

En las secciones anteriores hemos podido exponer algunos aspectos de los algoritmos, especialmente en lo que se refiere a su tipo, uso e impacto en nuestras vidas.

Por una parte tenemos los algoritmos que utilizan herramientas avanzadas de inteligencia artificial, especialmente aprendizaje profun-

⁹ <https://visor.gva.es/visor/>

do, que incluyen capas ocultas en el proceso las cuales imposibilitan la trazabilidad y pueden tener sesgos en los datos que los nutren. A pesar de las dificultades inherentes, se está trabajando en mejorar este tipo de algoritmos de modo que, en la búsqueda de un equilibrio adecuado entre precisión del algoritmo y corrección de sesgos. Además, y esto supone un enorme reto matemático-computacional, hay un objetivo puesto en la construcción paralela de algoritmos trazables y explicables que puedan ofrecer unas prestaciones como los de aprendizaje profundo. Luego parece recomendable seguir incidiendo en esas líneas de trabajo, dado que la inteligencia artificial irremediamente va a seguir formando parte de nuestras vidas *in crescendo*.

Por otra parte se ha descrito lo que entendemos como algoritmos deterministas, trazables, y que permiten su auditabilidad. A diferencia de los anteriores, si se dispone de suficiente información, podemos explicar las conclusiones a las que llegan. Se hace preciso, evidentemente, poder acceder a los datos que los nutren para una correcta auditabilidad. Además este tipo de algoritmos son capaces de ofrecer una solución satisfactoria sin necesidad de recurrir a herramientas avanzadas de inteligencia artificial. Es, en definitiva, aconsejable su uso. Sería preferible que este tipo de software fuera desarrollado por entidades públicas (universidades, centros públicos de investigación) o, en su defecto, que la empresa desarrolladora facilite todos los detalles para su auditabilidad posterior por terceros. Finalmente, todo algoritmo que utilice una serie de ponderaciones, estas se deben realizar en colaboración con expertos o basadas en estudios especializados, y ser sometidas a sucesivas actualizaciones.

Transparencia en la inteligencia artificial y en el uso de algoritmos: una visión de género

RAQUEL VALLE ESCOLANO

Directora del Máster Universitario en Dirección en la Gestión Pública

Profesora de Derecho Constitucional

UNIR

I. INTRODUCCIÓN

En una sociedad formada por mujeres y hombres, en la que la igualdad entre unas y otros, pese a los indiscutibles avances realizados, es todavía una quimera, es obvio que la irrupción de la inteligencia artificial (IA) y la utilización de algoritmos para la gobernabilidad, se encuentra permeada por numerosas e importantes implicaciones que tienen relación con el género. Tal y como señala el Instituto Europeo de Igualdad de Género¹, si bien los sistemas de IA pueden ofrecer nuevas oportunidades para la igualdad entre hombres y mujeres, su uso también plantea nuevos desafíos y crea el riesgo de implantar nuevos patrones de desigualdad de género, o de reforzar algunos ya existentes. En estas páginas se mencionarán y abordarán de forma necesariamente breve algunas de estas cuestiones, subrayando la imperiosa necesidad de que, en el diseño y la implementación de esta materia, que, pese a avanzar a grandes pasos, se encuentra todavía en un estadio incipiente, se elaboren estrategias y se aborden medidas que, dentro de las dimensiones éticas de la IA, introduzcan parámetros igualitarios y combatan las habituales discriminaciones de género.

¹ EUROPEAN INSTITUTE OF GENDER EQUALITY (EIGE), *Artificial intelligence, platform work and gender equality*, 2021, <https://eige.europa.eu/publications/artificial-intelligence-platform-work-and-gender-equality-report>

II. LA IA Y EL USO DE ALGORITMOS: UN NUEVO PARADIGMA CON EVIDENTES VENTAJAS Y ÁREAS DE RIESGO

La inteligencia artificial es un término comprensivo que engloba un conjunto de tecnologías de rápida evolución, que sin duda puede generar un amplio abanico de beneficios económicos y sociales en múltiples sectores y actividades de la sociedad actual y futura. Por su parte, un algoritmo es un conjunto de instrucciones o reglas definidas y no-ambiguas, ordenadas y finitas, diseñado y utilizado en IA, que permite solucionar un problema, realizar un cómputo, procesar datos y llevar a cabo otras tareas o actividades, tales como la toma de decisiones utilizando datos. Como señala Cotino, los algoritmos existen y se han utilizado desde hace décadas; la razón por la que ahora centran la atención es que algoritmos avanzados de aprendizaje automático —*advanced machine learning algorithms*— están revolucionando y automatizando la vida y la sociedad². La IA y el uso de algoritmos, ha abierto la puerta a numerosas aplicaciones que pueden mejorar la vida de las personas, facilitar las políticas públicas y con ello incrementar la eficiencia³, en un segmento que abarca desde la realización de cálculos más rápidos y con menores costes, hasta la realización de valoraciones sobre la realidad o la adopción de decisiones diferentes a las que habrían adoptado o generalmente adoptan los seres humanos⁴.

² L. Cotino Hueso, “Riesgos e impactos del Big Data, la inteligencia artificial y la robótica: enfoques, modelos y principios de la respuesta del derecho”, *Revista general de Derecho administrativo*, núm. 50, 2019, p. 3, <https://www.dropbox.com/s/010qzjnh7mwwk3t/cotinoiustel.pdf?dl=0>

³ E. Ruvalcaba-Gómez, “Inteligencia artificial en los gobiernos locales de México: análisis de percepción de los responsables de TIC”, VVAA, *Inteligencia artificial y ética en la gestión pública*, Centro Latinoamericano de Administración para el Desarrollo (CLAD), 2021, pp. 111-138. <https://clad.org/wp-content/uploads/2021/03/Libro-7-Inteligencia-artificial-y-%C3%A9tica-en-la-gesti%C3%B3n-p%C3%BAblica.pdf>

⁴ 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”, *Revista de Derecho Público: Teoría y Método*, vol. 1, pp. 223-270, p. 230.

El Libro Blanco sobre la inteligencia artificial de la UE señala, en esta línea, que la IA tiene un gran potencial capaz de cambiar nuestras vidas, a través de aspectos tan relevantes como una mejor atención sanitaria (por ejemplo, mediante el logro de diagnósticos mucho más precisos, o permitiendo una mejor prevención de las enfermedades), un aumento de la eficiencia de la agricultura, optimizando la contribución a la mitigación del cambio climático, la mejora de la eficiencia de los sistemas de producción a través de un mantenimiento predictivo, el aumento de la seguridad de los europeos, y toda una serie de transformaciones adicionales que a día de hoy tan solo podemos intuir⁵. De igual modo, el Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial, también subraya los relevantes beneficios de la IA, cuando establece que, mediante la mejora de la predicción, la optimización de las operaciones y de la asignación de los recursos y la personalización de la prestación de servicios, la inteligencia artificial puede facilitar la consecución de resultados positivos desde el punto de vista social y medioambiental, así como proporcionar ventajas competitivas esenciales a las empresas y la economía europea. Esto es especialmente necesario en sectores de gran impacto como el cambio climático, el medio ambiente y la salud, el sector público, las finanzas, la movilidad, los asuntos internos y la agricultura⁶. En particular, en el sector público, se ha puesto de manifiesto cómo el futuro de la IA y la gobernanza de los algoritmos, constituye un proceso abierto capaz de generar una gobernanza inteligente, entendida como un nuevo paradigma basado en la disponibilidad de cada vez más tecnologías

⁵ COMISIÓN EUROPEA, *Libro Blanco sobre la inteligencia artificial -un enfoque europeo orientado a la excelencia y la confianza*, 2020, https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf

⁶ COMISIÓN EUROPEA, *Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial)*, 2021, https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF

basadas en datos y una orientación hacia la provisión de mejores servicios públicos para generar valor público^{7/8/9}.

No obstante, los mismos factores y técnicas que potencian los beneficios socioeconómicos de la IA, también pueden generar nuevos riesgos potenciales o consecuencias negativas para personas o grupos concretos o para la sociedad en su conjunto. Algunos de los más destacados se vinculan con la responsabilidad, la transparencia, la sustitución de la intervención humana y la consiguiente pérdida de control de la tecnología (asociada a la transición de una tecnología administrada por humanos a una IA independiente), la opacidad en la toma de decisiones, la seguridad, la protección de datos y la privacidad de los ciudadanos, entre otros^{10/11}. En este sentido, una de las relaciones más completa y sistemática de ventajas y eventuales peligros de toda esta serie de tecnologías es, sin duda, la que lleva a cabo Cotino, aportando cifras representativas de los beneficios e impacto económico de la IA, de los *big data* o macrodatos y de la robótica, así como, en el otro extremo, detallando toda una serie de riesgos, peligros e impactos asociados a tales realidades, que exigen respuestas jurídicas¹².

⁷ J. I. Criado, J. R. Gil-García, “Creating public value through smart technologies and strategies: from digital services to artificial intelligence and beyond”, *International Journal of Public Sector Management*, 32(5), 2019, <https://www.emerald.com/insight/content/doi/10.1108/IJPSM-07-2019-0178/full/pdf?title=creating-public-value-through-smart-technologies-and-strategies-from-digital-services-to-artificial-intelligence-and-beyond>

⁸ J. I. Criado, “Inteligencia artificial: madurez tecnológica, adopción e innovaciones en la gestión pública”, VVAA, *Inteligencia artificial y ética en la gestión pública*, Op. Cit., pp. 11-44.

⁹ A. Meijer, “Datapolis: A Public Governance Perspective on ‘Smart Cities’”, *Perspectives on public Management and Governance*, 1(3), 2017, pp. 195-203.

¹⁰ B.W. Wirtz, J.C. Weyerer, C. Geyer, “Artificial Intelligence and the Public Sector-Applications and Challenges”, *International Journal of Public Administration*, 42(7), 2019, pp. 596-615.

¹¹ UNESCO, *Artificial Intelligence and Gender Equality. Key findings of UNESCO’s Global Dialogue*, 2020, <https://unesdoc.unesco.org/ark:/48223/pf0000374174>

¹² L. Cotino Hueso, “Riesgos e impactos del Big Data, la inteligencia artificial y la robótica: enfoques, modelos y principios de la respuesta del derecho”, cit.

Entre estas debilidades o riesgos figuran, precisamente, los sesgos de género asociados a la información que sirve de base a la IA, así como la operativa en el diseño, objetivos, implementación y funcionamiento de los algoritmos que realizan los cálculos operacionales y llegan a adoptar decisiones que nos afectan diariamente en cuestiones en muchos casos fundamentales.

III. LA DIMENSIÓN ÉTICA COMO EJE CLAVE DE LA IA Y DE UNA GOBERNANZA MEDIADA POR ALGORITMOS

Existe un claro imperativo para que en el campo de la IA se adopten y fortalezcan principios éticos sólidos y, en relación al tema que nos ocupa, que posibiliten cambios en materia de género dirigidos a la igualdad y a la no discriminación. La equidad constituye una preocupación cada vez más importante, ya que los modelos algorítmicos y de aprendizaje automático se utilizan, como ya se ha señalado, para respaldar la toma de decisiones en aplicaciones de alto riesgo, como préstamos hipotecarios, contratación y sentencias de prisión¹³.

Como señala Cotino¹⁴, el punto de partida y premisa ética de la IA en las numerosas declaraciones y los diferentes documentos que la abordan, no es otro que la dignidad y los derechos fundamentales. El autor destaca asimismo la importancia fundamental que se ha concedido a dicha cuestión a nivel europeo, creando un diseño de la misma *made in Europe*, con sello propio frente al mundo, y subrayando su papel, en una concepción que va más allá de la estricta regulación de aspectos concretos, tal y como puso de manifiesto el Supervisor Europeo de Protección de Datos¹⁵, al señalar que en el entorno digital

¹³ R. K. Bellamy et al.. “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias”, *IBM Journal of Research and Development*, vol. 63, núm. 4/5, 2019, pp. 1-4, <https://arxiv.org/abs/1810.01943>

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”, *Revista catalana de dret públic*, núm. 58, 2019, pp. 29-48.

¹⁵ Supervisor Europeo de Protección de Datos, Dictamen 4/2015, *Hacia una nueva ética digital. Datos, dignidad y tecnología*, p. 4 y ss., <https://edps.europa.eu/sites/>

actual no basta con respetar la ley, sino que es preciso tener en cuenta la dimensión ética del tratamiento de datos.

Una buena recopilación de los principios éticos que debe respetar la IA y el uso de algoritmos, es la que lleva a cabo *AI ETHICS LAB*¹⁶, que los enumera y agrupa, tal y como puede apreciarse en la tabla 1, en relación con cada uno de los grandes ejes que les sirven de marco.

En particular, el *Informe UNESCO sobre Inteligencia artificial e igualdad de género* insta a reforzar las consideraciones de género y a abordarlas de manera monográfica en el terreno de las implicaciones éticas, señalando que la preocupación por la igualdad de género en ocasiones se incluye como un subconjunto, dentro de las consideraciones generales por el sesgo en el desarrollo de algoritmos, en el conjunto de datos utilizados para su formación, y en su uso en la toma de decisiones. Sin embargo, el género es una cuestión mucho más relevante y amplia, que incluye, entre otras materias, el empoderamiento de las mujeres vinculado a su representación y a su presencia y protagonismo entre las personas que desarrollan y programan en IA, los investigadores y líderes de empresas, así como el acceso a la educación inicial y las oportunidades de formación para las mujeres, todo lo cual requiere la consideración de la igualdad de género como un principio independiente del resto de consideraciones de corte ético que se manejan en IA¹⁷.

edp/files/publication/15-09-11_data_ethics_es.pdf

¹⁶ AI Ethics Lab, *A Guide to the Dynamics of AI Principles: Our Toolbox for Understanding & Using AI Principles*, 2020, <https://aiethicslab.com/big-picture/>

¹⁷ UNESCO, *Artificial Intelligence and Gender Equality*, cit.

Autonomía	No hacer daño	Beneficio	Justicia
Poder para decidir Control humano Supervisión humana Transparencia (para entender) Apertura (para entender) Explicabilidad Libertad Derechos fundamentales Privacidad personal Protección de la privacidad Valores humanos	Control de los riesgos Seguridad Capacidad de precaución Protección de datos Privacidad (para evitar el daño) Explicabilidad Transparencia (para evitar el daño) Reproducibilidad Precisión Fiabilidad Despliegue responsable Prevenir la carrera armamentista	Promover el bienestar Beneficio para la sociedad Generación de beneficios netos Sostener el planeta Impacto Eficacia Explicabilidad Excelencia científica Diseño centrado en el usuario Enfocado a las personas, como prioridad	Justicia Derechos fundamentales Igualdad No discriminación Evitar sesgos Inclusividad Diversidad Neutralidad de los datos Datos representativos Beneficio y prosperidad compartidos Impacto social y económico Evitar la disparidad Mitigar la desvertebración social Preservar la solidaridad Explicabilidad Transparencia (para la rendición de cuentas) Apertura (para la rendición de cuentas) Rendición de cuentas Auditabilidad Fiabilidad Inclusividad Transparencia judicial Gobernanza abierta

Tabla 1. Principios de la Inteligencia Artificial. Ethics Lab (2020)

IV. LAS (INEVITABLES) IMPLICACIONES DE GÉNERO

La igualdad entre mujeres y hombres es un principio jurídico universal reconocido en diversos textos internacionales sobre derechos humanos, así como un eje esencial en el ámbito de los derechos de la ciudadanía en la Unión Europea, que ha permitido desarrollar un acervo comunitario sobre igualdad de género de gran amplitud e importante calado, con un claro reflejo en el panorama legal de nuestro país. De este modo, la igualdad entre los sexos y la no discriminación

tienen una vigencia formal indiscutible, al punto que el art. 14 sirve para inaugurar a nivel constitucional la enumeración de los derechos fundamentales y libertades públicas, sin perjuicio de constatar que existen numerosas e importantes brechas que cuestionan la igualdad real, generando inaceptables situaciones de discriminación en numerosos, variados y significativos ámbitos de nuestra sociedad.

De ahí que nuestra Ley Orgánica 3/2007, de 22 de marzo, para la igualdad efectiva de mujeres y hombres, enuncie y desarrolle toda una serie de criterios generales de actuación de los poderes públicos, que coadyuven a los objetivos de la norma, con un planteamiento basado en la transversalidad, que lleva al legislador a declarar que el principio de igualdad de trato y oportunidades entre mujeres y hombres informará, con carácter transversal, la actuación de todos los poderes públicos, con objeto de contribuir a eliminar desigualdades de género, corregir procedimientos y métodos de trabajo e impulsar tendencias de cambio social. Las Administraciones públicas lo integrarán, de forma activa, en la adopción y ejecución de sus disposiciones normativas, en la definición y presupuestación de políticas públicas en todos los ámbitos y en el desarrollo del conjunto de todas sus actividades.

La cuestión es que, cuando irrumpen cambios y transformaciones sociales que generan situaciones nuevas, hay que estar vigilantes, ya que al surgir de los cimientos y del engranaje de la sociedad actual, las desigualdades de género existentes suelen traspasarse a las nuevas realidades. Esto es precisamente lo que ocurre con el uso de algoritmos para la toma de decisiones en el gobierno y las políticas públicas, y en general en el ámbito de la inteligencia artificial, terrenos en los que se advierte la existencia de importantes implicaciones relacionadas con el género. En este sentido, mujeres y hombres tienen que tener, en un espacio tan transcendental para posibilitar avances en la sociedad, y susceptible de transformar muchas de las pautas y reglas básicas jurídicas, técnicas, científicas, de las políticas públicas o en definitiva, de la gobernabilidad, iguales derechos, y una presencia, protagonismo y representación similares, de modo que la IA y el uso de algoritmos no se conviertan en cuestiones en las que, como en tantas otras, los intereses, deseos y expectativas de las mujeres se vean cercenados, mientras que en las reglas de juego y la toma de decisiones exista un predominio claramente masculino.

En esta línea, podríamos resaltar varios aspectos, en la materia analizada, cuyo análisis es interesante en clave de género, como son las siguientes: la IA y el uso de algoritmos como oportunidad en materia de igualdad de género; el protagonismo de las mujeres en el campo de la inteligencia artificial. Eventuales consecuencias en el trabajo femenino; y sesgos de género en la construcción y manejo de algoritmos que quiebran la igualdad y generan discriminación. A todos ellos haremos referencia a continuación.

1. La IA y el uso de algoritmos como oportunidad en materia de igualdad de género

No podemos renunciar a concebir teóricamente cualquier situación novedosa de avance, tendente a mejorar el bienestar y las condiciones de vida de la ciudadanía, como lo es la implantación de la IA y el uso de algoritmos en la gobernabilidad, como una oportunidad asimismo de progreso en la línea de consecución de una sociedad más igualitaria, más inclusiva y más justa.

La propia Estrategia para la IA en I+D+I de España, formulada en 2019, establece, al enunciar sus prioridades, que serán enmarcadas en la nueva Estrategia Española de Ciencia, Tecnología e Innovación (EECTI) 2021-2028 y que deberán ser desarrolladas en iniciativas y actividades definidas y financiadas a través de los Planes Estatales de Ciencia, Tecnología e Innovación (PECTI), que es una condición en el desarrollo de las tecnologías y aplicaciones de la IA ligada a esta estrategia evitar el sesgo negativo y los prejuicios de los que adolece nuestra sociedad, como el sesgo de género¹⁸.

Los gobiernos son conscientes de esa responsabilidad frente a los desafíos de las nuevas tecnologías en general y la IA en particular. La brecha de género en el ámbito científico y la discriminación de género en el ámbito tecnológico son factores de vulnerabilidad de nuestra economía y nuestra sociedad. Por ello, en la actualidad existe en los diversos países una mayor sensibilización, que en ocasiones lleva a

¹⁸ MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES, *Estrategia Española de I+D+I en Inteligencia Artificial*, 2019, p. 9, https://portal.mineco.gob.es/RecursosArticulo/mineco/ministerio/ficheros/201202_ENIA_V1_0.pdf

prácticas y recomendaciones, tanto del sector público y organizaciones internacionales, como de determinadas organizaciones de índole privada, que tratan de sensibilizar sobre la necesidad de evitar el sesgo negativo y los prejuicios de género u otras formas de discriminación en la IA y sus prácticas.

De igual modo, a nivel internacional, organizaciones internacionales instan en documentos y declaraciones diversas a todos los actores involucrados en el ámbito de la IA (compañías, líderes, actores sociales, etc.) a combatir sesgos de género y a desarrollar soluciones igualitarias y éticamente comprometidas en este campo. Podríamos decir que las consideraciones de género están comenzando, aun tímidamente, a irrumpir en la agenda digital y en las nuevas tecnologías.

Finalmente, y en sentido positivo, algoritmos y herramientas informáticas han comenzado también a aplicarse para conseguir mayores grados de igualdad y combatir la discriminación. Valga, por todos, como ejemplo, la experiencia de uso de algoritmos para integrar la perspectiva de género en la administración de la justicia, auspiciada por *Iniciativa Latinoamericana de Datos Abiertos* (ILDA), que aplicó esta herramienta, en un proyecto interesante y muy completo, a la estandarización de datos sobre feminicidios y sistematización de la información contenida en las sentencias relacionadas con este delito y en materia de violencia de género¹⁹.

2. *El protagonismo de las mujeres en el campo de la inteligencia artificial. Eventuales consecuencias en el trabajo femenino*

La posición tradicional de mujeres y hombres ante la realidad digital, preámbulo o área íntimamente relacionada con la inteligencia artificial, no ha sido en absoluto igualitaria.

Numerosos estudios han abordado, en este sentido, aspectos como la brecha digital de género, examinando al tiempo las oportunidades

¹⁹ H. Fallas, C. A. Contreras, *Algoritmos para integrar la perspectiva de género en la administración de la Justicia*, Iniciativa Latinoamericana de Datos Abiertos (ILDA), 2022, <https://datosabiertos.org/wp-content/uploads/2022/05/Algoritmos-para-integrar-la-perspectiva-de-ge%CC%81nero-en-la-Justicia.pdf>

y los desafíos para la igualdad de género en los mercados laborales actuales, profundamente transformados por la irrupción de la inteligencia artificial (IA) y el trabajo en plataformas digitales. Algunas de las conclusiones, altamente coincidentes en todos los casos, alcanzadas por los mismos, hacen referencia a las siguientes cuestiones^{20/21}:

- Las mujeres hacen un uso cada vez más frecuente de las tecnologías digitales, pero este mayor acceso a los servicios y productos no significa su inclusión digital. La brecha digital de género persiste con fuerza, confirmando las diferencias en el nivel de competencias digitales entre hombres y mujeres, así como los diferentes usos de las tecnologías, que se mantienen a lo largo de las distintas etapas de la vida. Para abordarlas es imprescindible considerar la intersección del género con otras variables como el nivel educativo, la edad, la dimensión rural/urbana, o el nivel socioeconómico. Su superación es todavía un reto que requiere políticas, estrategias y medidas decididas y novedosas, dado que se trata de una brecha extraordinariamente resistente.
- La transformación del trabajo relacionada con la IA tiene impactos desiguales de género en el mercado de trabajo. Así, aunque determinadas tareas habitualmente realizadas por hombres en el transporte, almacenamiento y las actividades de fabricación pueden automatizarse, el riesgo de pérdida de empleo de las mujeres es mayor, debido a la automatización, que podría eliminar muchos trabajos administrativos²². En otro orden de cosas, la demanda de profesionales de IA está creciendo rápidamente, y se trata de empleos bien pagados. Sin embargo, existe una disparidad de género considerable: los hombres dominan el diseño de la IA, la programación y la gestión, mientras que las mujeres realizan tareas en ocasiones invisibilizadas, en en-

²⁰ M. Sainz, L. Arroyo, *Mujeres y digitalización. De las brechas a los algoritmos*. Instituto de la Mujer y para la Igualdad de Oportunidades, Ministerio de Igualdad, 2020, DOI: 10.30923/MujDigBreAlg-2020

²¹ EUROPEAN INSTITUTE OF GENDER EQUALITY (EIGE), *Artificial intelligence, platform work and gender equality*, 2021, cit.

²² G. Lordan, *What Work Disappears? Automation and the changing nature of work*, Publications Office of the European Union, 2019, <https://op.europa.eu/en/publication-detail/-/publication/1246be88-329a-11ea-ba6e-01aa75ed71a1/>

tornos menos privilegiados. Los factores explicativos de dicha realidad no parecen mejorar, y están relacionados con estereotipos de género y la brecha de género en habilidades digitales, la falta de equilibrio trabajo-vida, la brecha salarial entre varones y mujeres o la existencia de entornos de trabajo fuertemente dominados por hombres. Finalmente, la tecnología basada en IA puede reforzar los estereotipos de género y permitir nuevas formas de violencia de género, la ciberviolencia²³.

- Las evidencias sugieren además que el techo de cristal constituye un importante freno a la permanencia y desarrollo profesional de las mujeres en el sector TIC.
- Por lo que respecta a los trabajos que se llevan a cabo a través de plataformas tecnológicas, con gran impacto en las últimas décadas, los estudios han demostrado que no son inmunes a las desigualdades entre mujeres y hombres, sino que, por el contrario, reflejan claramente las tendencias generales del mercado laboral tradicional. De ahí la necesidad de abordar las desigualdades de género en este terreno, de modo que su desarrollo no contribuya a mayores grados de segmentación del mercado laboral ni de descualificación, en perjuicio de las trabajadoras.

Finalmente, la práctica totalidad de estudios analizados coinciden en subrayar la importancia de las competencias digitales para la empleabilidad, así como en situar la IA como tecnología clave para determinar cómo serán los empleos de futuro. De ahí la necesidad de invertir la situación actual, fomentando la formación continuada de las mujeres en competencias digitales.

3. Sesgos de género que quiebran la igualdad y generan discriminación

El diseño e implantación de sistemas de IA tiene ante sí, por lo relativamente reciente de su existencia, un desafío ético que conviene analizar, que no es otro que el trato desigualitario que el manejo de datos conlleva en relación a determinados colectivos, que se incor-

²³ EIGE, *Cyber violence against women and girls*, 2017, <https://eige.europa.eu/publications/cyber-violence-against-women-and-girls>

pora en los algoritmos, de forma más o menos inconsciente. Campos clasifica tales sesgos en estadístico, en función de cómo se obtiene la información, que puede provocar errores de medición o similares; cultural, derivado de factores como el lenguaje manejado en cada caso o los estereotipos y códigos presentes en cada sociedad a lo largo del tiempo; cognitivo, que nos identifica y depende de nuestra personalidad; o el sesgo de confirmación, que avala nuestras opiniones y creencias, sesgando la información en función de las mismas²⁴. Para combatir todos ellos, y su influencia en los patrones de reconocimiento, en la recolección de datos y en la propia programación, es necesario que las organizaciones responsables en materia de IA articulen vías para evitar que tales sesgos y prejuicios penetren en su funcionamiento.

En concreto, en este ámbito adquieren una extraordinaria relevancia las desigualdades y discriminaciones de género, que actúan, entre otras vías, a través de los estereotipos, del sesgo en la recogida y manejo de datos, o del protagonismo secundario que aún hoy poseen las mujeres en muchos espacios económicos, sociales o culturales.

De hecho, en la actualidad el sesgo de género en el desarrollo de algoritmos, en los conjuntos de datos utilizados para su construcción, y en su uso en la toma de decisiones, se considera un área crítica y emergente en el binomio igualdad de género-IA, que requiere una actuación rigurosa e inmediata²⁵. En esta línea, resulta muy revelador un trabajo monográfico realizado en la Universidad de California Berkeley, por el Centro de Género, Equidad y Liderazgo (EGAL) de la *Haas School of Business*, orientado a mitigar el sesgo en inteligencia artificial. El estudio proporciona un marco sobre cómo el sesgo se manifiesta en conjuntos de datos y algoritmos. Para ello, analiza y descompone conceptos para eliminar las brechas del campo de la IA desde un prisma técnico, tras poner de manifiesto que el sesgo en la IA es un problema generalizado y crítico.

²⁴ C. Campos Acuña, “Herramientas para una gobernanza ética de la inteligencia artificial: una visión directiva”, VVAA, *Inteligencia artificial y ética en la gestión pública*, Op. Cit., p. 75-110, p. 92.

²⁵ UNESCO, *Artificial Intelligence and Gender Equality*, cit.

Es interesante desgranar algunas de las consideraciones que el trabajo recoge en relación con los sesgos y la inadecuada construcción y utilización de los algoritmos, entre ellas las siguientes²⁶:

- Los conjuntos de datos utilizados para construir los algoritmos son fundamentales: los sistemas de IA toman decisiones basadas en dicha información, y son varios los puntos por donde puede penetrar el sesgo. Dichos conjuntos de datos se generan a partir de la actividad diaria de las personas (por ejemplo, consumo, comportamiento, condiciones de salud), y se recolectan a través de las diversas plataformas, tecnológicas o de otro modo. Se supone que los datos son precisos, pero existen importantes lagunas (incluyendo poca o ninguna información proveniente de ciertos individuos y comunidades), además de que la misma contiene aspectos claramente discriminatorios —por raza, cuestiones económicas, etc.—, y sesgos de género.
- Por otro lado, la influencia humana no se puede eliminar de los datos. En muchos casos, son personas quienes deciden qué, dónde y cómo se almacenan los datos recogidos y categorizados, así como parámetros para los distintos conjuntos de datos. Finalmente, los datos también se etiquetan, una actividad con una importante carga de subjetividad.
- Un conjunto de datos sesgado puede no ser representativo de la sociedad al sobre o sub-representar ciertas identidades en un contexto particular. En otras ocasiones, tales datos sesgados pueden ser exactos, pero representativos de una sociedad injusta, en la medida en que reflejan la discriminación que efectivamente están sufriendo determinados grupos de personas.
- El sesgo también puede penetrar en un algoritmo en diversos momentos. Así, puede hacerlo cuando se define el propósito de un modelo de IA y las limitaciones bajo las que opera; o bien al seleccionar las entradas que el algoritmo debe considerar para encontrar patrones y sacar conclusiones, ya seleccionando datos y/o *proxies*/variables.

²⁶ G. Smith, I. Rustagi, *Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook*, The Center for Equity, Gender and Leadership at the Haas School of Business, University of California, Berkeley, 2020.

- Finalmente, existe la posibilidad de errores en las predicciones y de sesgos si se utiliza un sistema de IA en un contexto diferente o para una población diferente a partir del cual se desarrolló originalmente o si se aplica para casos distintos a aquellos para los cuales fue originalmente desarrollado. Y ello puede hacerse tanto errónea como intencionalmente, por individuos u organizaciones que intervienen en estos procesos.

Cuando alguna de las situaciones anteriores sucede, un algoritmo puede dar lugar a resultados técnicamente inexactos, en forma de predicciones y/o decisiones que afectan a ciertos subconjuntos de la población. También un algoritmo puede contribuir a la discriminación, independientemente de la calidad de los datos utilizados, dependiendo de cómo se evalúe.

Vemos, pues, que el riesgo para la creación e implementación de soluciones y decisiones automatizadas, provenientes del uso de algoritmos, es altísimo, dado que en ciertas ocasiones se excluye a distintos sectores de la población (por ejemplo, a las personas LGBTIQ+), o se manejan datos parciales e inexactos sobre un grupo, o sacados de contexto, aumentando con ello el riesgo de su vulnerabilización. De ahí la necesidad de asegurar que los datos utilizados son representativos de la población a la que se dirige el sistema. Y en el caso de las mujeres, existe en general una infrarrepresentación en multitud de ámbitos, generada por aspectos diversos, como las discriminaciones históricas, la ausencia de puesta en valor de su protagonismo, trabajo o intervención en multitud de áreas y contextos, la falta de estadísticas que segreguen datos en función del sexo, etc.

Ante este estado de cosas, diversos estudios y documentos monográficos en la materia^{27/28/29}, han realizado interesantes recomendaciones, que atañen a diversos órdenes de cuestiones:

- En relación a los equipos de trabajo, se subraya la necesidad de construir equipos diversos y multidisciplinares a la hora de tra-

²⁷ *Ibidem.*

²⁸ UNESCO, *Artificial Intelligence and Gender Equality*, cit.

²⁹ R. Avila, A. Brandusescu, J. Ortiz Freuler, D. Thakur, *Policy Brief W20 Argentina: Artificial Intelligence: Open Questions about Gender Inclusion*, World Wide Web Foundation, Geneva, 2018, <http://webfoundation.org/docs/2018/06/AI-Gender.pdf>

bajar en algoritmos y sistemas de IA. Al tiempo, es imprescindible promover una cultura de la ética y la responsabilidad en relación con la IA, que llegue a todos los actores relacionados con la misma, de modo que les aliente a priorizar las consideraciones de equidad en cada paso del proceso de desarrollo del algoritmo. Por otro lado, dado que suprimir completamente el sesgo de la IA puede no ser factible, las organizaciones deben mantener un estándar de explicabilidad en torno al funcionamiento de sus modelos, así como de transparencia alrededor de sus posibles deficiencias.

- Por lo que respecta a los modelos de IA, es necesario implementar desarrollos de datos responsables, con controles y equilibrios capaces de crear, si es necesario, nuevos conjuntos de datos, así como adaptar los existentes. Para ello hay que establecer políticas y prácticas que permitan el desarrollo responsable de algoritmos, que verifiquen y mitiguen el sesgo en cada etapa del proceso de desarrollo de algoritmo. Esto implica dotar a los equipos de marcos éticos que prioricen la equidad al definir los objetivos de los algoritmos, asegurando que los conjuntos de datos utilizados son desarrollados y etiquetados responsablemente, y garantizar que las variables no perjudiquen a ciertas comunidades. Es necesario concienciar y sensibilizar, de modo que las organizaciones que emplean IA aborden los problemas de falta de representatividad y sesgo, y que incrementen la recopilación y el uso de datos interseccionales, especialmente de grupos subrepresentados. En particular, y en cuanto a la corrección del sesgo de datos, siempre que sea posible, hay que comprometerse a usar conjuntos de datos que hayan sido desarrollados con un prisma de igualdad de género (pese a que ello pudiera suponer un aumento de costes), detectar y corregir el sesgo de datos y garantizar que los mismos representan efectivamente a la población a la que el algoritmo va a afectar.
- Finalmente, en el ámbito de la responsabilidad social y del liderazgo, se propone establecer modelos de gobierno corporativo para una IA responsable y poner en marcha políticas internas para mitigar sesgos, en este caso de género, en todas las fases del proceso. Para ello, establecer estructuras de gobernanza y liderazgo de cariz ético en la IA es un primer paso. Es necesi-

rio introducir postulados de responsabilidad social corporativa capaces de promover cambios de gran alcance en materia ética en la IA, involucrando y educando en la materia a científicos/as, ingenieros/as, empresarios/as y estudiantes. Se insta asimismo a todos los actores implicados a utilizar su voz e influencia para avanzar en los cambios legislativos que hagan posible una IA responsable.

Hemos visto, por tanto, que los sesgos presentes en la información que habitualmente se maneja, perjudican y discriminan en materia de género, y no son más que un reflejo de las dinámicas sociales y del desequilibrio en el protagonismo y poder en la toma de decisiones, que provoca una infrarrepresentación de las mujeres en multitud de ámbitos. Los sesgos de género en el diseño y creación de tecnologías reproducen así los sesgos existentes en nuestra sociedad, confirmando, por otro lado, la escasa presencia de mujeres en el diseño y creación de tecnología. Es necesario, por tanto, identificarlos y combatirlos, ya que la IA y el uso de algoritmos van a ver intensificado su uso de forma importante en las próximas décadas, tanto para la predicción como para la toma de decisiones.

Y por otro lado, tal y como subrayan Sainz y Castaño³⁰, la escasez de estudios que aborden los sesgos de género —por ejemplo, poca presencia de mujeres entre las poblaciones consideradas— en la generación, tratamiento y análisis de datos masivos (*big data*), pone de manifiesto la urgente necesidad de examinar las deficiencias de la calidad de la información que se toma como referencia para el diseño de políticas de impacto social, como puede ser la salud, la educación, o el consumo.

Más allá de ello, desde planteamientos feministas se ha argumentado también que los modelos matemáticos no siempre son capaces de ofrecer una explicación a la dimensión política, social o económica de fenómenos tan complejos como la discriminación o la desigualdad de género. Reducir la realidad social a estadísticas y modelos, provoca en ocasiones que ciertas relaciones de poder no se hagan explícitas, lo que restaría validez a su utilización en clave de género. Los estudios

³⁰ M. Sainz, L. Arroyo, *Mujeres y digitalización. De las brechas a los algoritmos*, cit., p. 107.

de *big data* no suelen llegar a penetrar en determinadas consideraciones de lo social, por lo que, pese a reconocer su extraordinaria capacidad de facilitar la vida de las personas, en ocasiones tienden a la simplificación y presentan importantes limitaciones a la hora de, por ejemplo, en las dinámicas de género, dar cuenta de elementos como el poder, en tanto que elemento explicativo y constitutivo de lo social, ni de la lógica de las carencias. Por ello es necesario también un desarrollo de los mismos a partir de tales consideraciones críticas, de modo que puedan ser capaces de integrar y abordar todas las dimensiones interseccionales que condicionan la discriminación de las mujeres.

V. LA TRANSPARENCIA COMO PROBLEMA DE FONDO

Hemos visto, pues, la importancia creciente que están adquiriendo los modelos de *machine learning* o aprendizaje automático, fundamentalmente a través de las decisiones que los mismos adoptan, a partir de los algoritmos con los que estos sistemas se construyen. Parece inevitable que, cada vez más, factores clave de nuestras vidas puedan ser medidos y evaluados por un algoritmo. Y aquí aparece el problema o cuestión fundamental, y es que los datos que alimentan a dichos algoritmos pueden contener sesgos que inciden en la decisión tomada, lo que finalmente tiene impacto en las personas y en sus derechos.

Para contrarrestar ese hándicap, se propone como solución más clara la transparencia. Una necesidad de transparencia en la toma de decisiones que constituye hoy en día una preocupación esencial de gobiernos y organizaciones internacionales. La transparencia de datos y de algoritmos, como subraya Sangüesa³¹, implica la capacidad de saber qué datos se utilizan, cómo se utilizan, quiénes los utilizan, para qué los utilizan y cómo se llega a partir de los datos a tomar las decisiones que afectan a la esfera vital de quien reclama esta transparencia.

³¹ R. Sangüesa, "Inteligencia artificial y transparencia algorítmica: "It's complicated", BiD: textos universitaris de biblioteconomia i documentació, núm. 41 (deseembre), 2018, p. 2, <http://bid.ub.edu/es/41/sanguesa.htm>, DOI: <http://dx.doi.org/10.1344/BiD2018.41.12>

La forma más usual a la que se ha recurrido en la práctica para implantar dicha transparencia ha sido la vía regulatoria, que tiene claros defensores, por todos Boix Palop³², que argumenta que asumir el valor normativo del código fuente integrado en la adopción de decisiones —dado que los algoritmos y programas son empleados como herramientas decisorias en cuanto a la concurrencia o no de ciertas circunstancias de hecho, asociadas a ciertas consecuencias jurídicas a los mismos—, lleva de forma inevitable a deducir consecuencias jurídicas, centradas en la necesidad de regular tales algoritmos cuando son empleados por la Administración Pública con funciones de predefinición y programación normativa. Otros posicionamientos, sin embargo, critican el enfoque normativo, al considerar que el mismo ignora una premisa básica, y es que la transparencia no es solamente un problema legal, sino también un problema tecnológico, y al desconocer tal complejidad, dicho planteamiento genera un marco regulatorio insuficiente y rígido³³.

Lo cierto es que, yendo más allá de dicho debate, es sumamente interesante la idea, apuntada por Sangüesa³⁴, de que la solución, para llegar a conocer el núcleo central de las decisiones hoy en día tomadas por mecanismos de inteligencia artificial, no radica solamente en los datos y en los algoritmos, sino en el contexto de prácticas sociales y políticas que los rodean. Y ello resultaría plenamente aplicable a las implicaciones que el uso de los algoritmos y los requerimientos de transparencia observan en materia de género. El problema, en este caso, no reside solo en desentrañar el funcionamiento último de tales algoritmos, sino en decodificar y combatir un conjunto de prácticas que han dado como resultado que el protagonismo y papel de las mujeres, así como sus deseos, expectativas, oportunidades, rasgos definitorios, y en definitiva, todo aquello que las caracteriza y mediante lo cual se expresan, en ocasiones esté silenciado, opacado y/o deficientemente interpretado en los datos disponibles, que sirven de base a los algoritmos, y con ello a la toma de decisiones, muchas de ellas

³² A. Boix Palop, A., “Los algoritmos son reglamentos...”, cit.

³³ C. Araya Paz, “Transparencia algorítmica ¿un problema normativo o tecnológico?”, *CUHSO (Temuco)*, núm. 31(2), 2021, pp. 306-334, <https://dx.doi.org/10.7770/cuhs0-v31n2-art2196>

³⁴ R. Sangüesa, “Inteligencia artificial y transparencia algorítmica...”, cit.

fundamentales, en los ámbitos público y privado. No hay que olvidar la insistencia de la Ley Orgánica 3/2007, para la igualdad efectiva de mujeres y hombres, en la necesidad de que se incluya sistemáticamente la variable de sexo en las estadísticas, encuestas y recogida de datos que se lleven a cabo; o de que se establezcan o incluyan en las operaciones estadísticas nuevos indicadores que posibiliten un mejor conocimiento de las diferencias en los valores, roles, situaciones, condiciones, aspiraciones y necesidades de mujeres y hombres, su manifestación e interacción en la realidad que se vaya a analizar; o de que se analicen otras variables cuya concurrencia resulta generadora de situaciones de discriminación múltiple en los diferentes ámbitos de intervención.

Las prácticas sociales y políticas constituyen, por tanto, aspectos clave a integrar en materia de transparencia algorítmica y en general en el uso de la IA, y sin embargo no han sido sino hasta fechas muy recientes, objeto de consideración, ya que no existía en los grandes centros de desarrollo de estas tecnologías una sensibilidad que fuera más allá de las cuestiones técnicas³⁵, que capaz de establecer marcos más amplios para afrontar y mitigar los riesgos generados por la IA, en los que, junto a los aspectos técnicos, se valoraran otros de índole ética y legal, en una perspectiva mucho más compleja e interdisciplinaria. Estas consideraciones llevarían a proponer, tal y como ya ocurre por ejemplo en el campo de la medicina o de la bioética, la existencia de prácticas novedosas que incorporaran en todos los pasos del diseño técnico, conceptos éticos.

Lo cierto es que, tal y como establece Beck³⁶, la inteligencia artificial, y mucho más en sus perfiles actuales, aun relativamente incipientes, constituye una tecnología de riesgo, que logra un impacto sistémico y distribuye sus riesgos y su incremento de forma desigual en la población. Y siguiendo la pauta más habitual, los colectivos hasta ahora más afectados por decisiones guiadas por sistemas de intelligen-

³⁵ D. Boyd, K. Crawford, "Six Provocations for Big Data", *A decade in Internet time: Symposium on the dynamics of the Internet and society*, 21 September 2011, Oxford Internet Institute, 2011. <http://dx.doi.org/10.2139/ssrn.1926431>

³⁶ U. Beck, *La sociedad del riesgo: hacia una nueva modernidad*, Paidós Ibérica, Barcelona, 2022.

cia artificial son precisamente los más débiles³⁷. En este marco cabría considerar las desigualdades de género, que hoy en día subsisten, todavía con fuerza en nuestra sociedad, y que provocan que el sesgo del uso de algoritmos actúe en mayor medida en perjuicio de sus derechos e intereses de las mujeres en general, e incluso, en los casos de ciertos colectivos de mujeres, discapacitadas, migrantes, o pertenecientes a minorías étnicas, entre otras, podrían eventualmente dichos sesgos afectarles en mayor medida.

VI. A MODO DE CONCLUSIÓN

La IA y el uso de algoritmos ha generado un importante salto cualitativo, con un potencial indiscutible para mejorar nuestras vidas. Las posibilidades de la inteligencia artificial para afrontar con eficiencia los importantes retos que tienen ante sí las sociedades contemporáneas, es, por tanto, indiscutible. Predecir, analizar y evaluar, en ámbitos trascendentales, como la economía, la medicina, las cuestiones sociales, entre otros; prestar mejores servicios públicos; o servir de apoyo a las políticas públicas y a la toma de decisiones, son todas ellas facetas que pueden procurar un mayor bienestar a la ciudadanía. Por supuesto, del mismo modo hay que analizar sus puntos oscuros o eventuales riesgos, identificarlos y combatirlos. Para ello, lo primero es apuntalar el edificio de la IA con toda una serie de consideraciones éticas que sean respetadas, de modo que no se pierda de vista que lo primero son las personas y sus derechos.

Entre dichas consideraciones éticas ocupan un lugar indiscutible las que tienen que ver con la igualdad de género. Porque todas las implicaciones que la IA y el uso de algoritmos poseen, sin duda afectan de forma diferencial a mujeres y hombres. En este sentido, es necesario subrayar que la IA es, en primer lugar, una oportunidad, que puede ser utilizada para reducir las discriminaciones de género y lograr mayores cotas de igualdad. Pero no es menos cierto que algunas consecuencias de esta nueva realidad no pueden ser evaluadas en términos tan positivos, siendo una de ellas las transformaciones que

³⁷ V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, St. Martin's Press, New York, 2018.

se han operado en el mercado de trabajo. Existe, en este sentido, una brecha digital, que provoca que el protagonismo de las mujeres en los ámbitos tecnológicos sea muy escaso, una situación que no parece que se altere con el paso de los años.

Por otro lado, pese a estar protagonizada e implementada por ordenadores y sistemas informáticos, lo cierto es que ni la inteligencia artificial ni el uso de algoritmos para predecir el comportamiento humano y, lo que es más importante, para la toma de decisiones, constituyen realidades neutras y asépticas, ajenas a la realidad social, económica y cultural en la que se insertan y desarrollan. En este sentido, en nuestra sociedad subsisten a día de hoy importantes brechas de género desigualitarias y situaciones discriminatorias, que los conjuntos de datos en que se basan y el propio comportamiento de los sistemas de IA no hacen más que replicar. Por eso es fundamental trabajar en la línea de que los sistemas de IA no reproduzcan prejuicios y discriminaciones de género. Al contrario, si se diseñan adecuadamente, los sistemas de IA deberían contribuir a reducir los prejuicios y la discriminación estructural existente.

Con tal objeto, las exigencias de transparencia y las consideraciones éticas, deben de ocupar un lugar esencial a la hora de establecer los perfiles y requerimientos de la IA y del uso de algoritmos, de modo que la desigualdad entre hombres y mujeres, los prejuicios, los estereotipos y los sesgos de género presentes en la información no penetren en el ámbito técnico, que no es en absoluto un espacio objetivo ni neutral, sino un reflejo de la sociedad en que vivimos. Siguiendo esta línea argumental, ya destacaba Surden³⁸ que los valores sociales presentes bajo el pretendido manto de objetividad de la tecnología, se hallaban en muchos casos incrustados en el diseño de los sistemas tecnológicos, y prevenía frente a la posibilidad de que estos pudieran ser contrarios a la igualdad, principios constitucionales y derechos humanos. Y es innegable, además, que las mujeres no tienen una presencia ni un protagonismo en la información y en los datos disponibles en las distintas áreas, similar a la que tienen los varones, lo que sin duda

³⁸ H. Surden, *Values Embedded in Legal Artificial Intelligence*, U of Colorado Law Legal Studies Research Paper No. 17-17, <http://dx.doi.org/10.2139/ssrn.2932333>

tiene un claro reflejo en los algoritmos y por ello en la realización de predicciones y en una toma de decisiones que no resulta igualitaria.

La extraordinaria relevancia y el impacto de estas realidades, que cada vez más van a ir moldeando nuestras vidas, están en la base de la absoluta necesidad de estudios rigurosos que sensibilicen en la materia y que sustenten reflexiones y recomendaciones útiles para perfeccionar el diseño de los sistemas de IA en general y de los algoritmos en particular, en clave de género, integrando de ese modo los derechos, intereses, opiniones, posibilidades, expectativas y aspiraciones de ambos sexos, en un plano igualitario.

Por otro lado, la histórica exclusión de las mujeres de las áreas tecnológicas, no hace sino incrementar exponencialmente los riesgos de una realidad sesgada. Para que su desarrollo personal y su crecimiento profesional sean completamente integrados, es imprescindible incrementar sustancialmente la participación de las mujeres en el diseño, implementación, evaluación y debate sobre ética y normas de la próxima generación de aprendizaje automático y de las tecnologías impulsadas por la IA. Solo la inclusión significativa de las mujeres en todas las etapas puede dar como resultado políticas y tecnologías que hagan realidad la igualdad digital³⁹.

³⁹ R. Avila, A. Brandusescu, J. Ortiz Freuler, D. Thakur, *Policy Brief W20 Argentina: Artificial Intelligence: Open Questions about Gender Inclusion*, cit.

Bloque II
**Transparencia y explicabilidad en
el uso de la inteligencia artificial
por el sector público**

Denegación *vs.* Derecho de acceso al código fuente en los sistemas algorítmicos. Una perspectiva jurídico-administrativa

GABRIELE VESTRI¹

*Prof. Ayudante Doctor de Derecho Administrativo²
Director del Observatorio Sector Público e Inteligencia Artificial
Universidad de Cádiz*

I. INTRODUCCIÓN

Los avances tecnológicos representan, hoy en día, un verdadero reto que desafía la forma de entender las relaciones entre ciudadanos y, desde luego, entre los administrados y las Administraciones públicas. El sector público español —como el de otras muchas realidades públicas europeas y mundiales— se encuentra en un proceso de transición digital o no tanto de transformación ya que en este último caso supondríamos un proceso de cambio que suele ser lento y ante el cual, tanto el usuario como las personas que ejercen su labor en el seno de la Administración pública, se van preparando gradualmente. Por esta razón en este trabajo, de manera voluntaria, utilizamos la expresión transición digital y no el término de transformación digital. El concepto de transición se realiza a través de un proceso comparativamente más rápido y que, en ocasiones, puede llegar a ser repentino. El actual escenario digital y las herramientas algorítmicas y de inteligencia artificial (IA) representan el fruto de acontecimientos que han adelantado considerablemente la implementación de sistemas disruptivos cambiando de hecho el paradigma de muchos sectores que hasta el momento nos permitían sentirnos, en cierta medida, acostumbrados.

¹ Se trata de la ampliación escrita de la ponencia impartida en el seno del *Congreso Transparencia, aplicabilidad y auditabilidad de la inteligencia artificial y los algoritmos* y organizado por la Universidad de Valencia, Generalitat Valenciana de los días 26-27 de abril de 2022, titulada: “*Derecho de acceso al algoritmo público: lecciones aprendidas del caso bono eléctrico*”.

² Acreditado a Profesor Titular de Universidad.

En este escenario, el derecho, especialmente administrativo, está llamado a responder jurídicamente a los retos aludidos anteriormente. Así, conceptos jurídico-administrativos, como transparencia, acto administrativo, motivación, discrecionalidad administrativa etc., —la enumeración podría continuar— se encuentran ante un obligado cambio de perspectiva. En este sentido, los conceptos mencionados, deben ser tratados hoy según las nuevas tecnologías disruptivas. Así, la doctrina jurídico-administrativa y el derecho público en general, lo que nos ocupa en este estudio, debaten sobre la transparencia algorítmica, sobre el valor de la decisión de la máquina como elemento instrumental del acto administrativo, sobre la necesidad, o no, de introducir el sistema algorítmico en la motivación del acto administrativo o sobre la limitación de la discrecionalidad administrativa debida al uso de algoritmos y de algoritmos basados en IA.

A la luz de este interesante y constante debate, las disciplinas jurídicas están llamadas a un (re)planteamiento de ciertos escenarios que, precisamente, a causa de los actuales sistemas tecnológicos, parecen cambiar los paradigmas conocidos hasta el momento. El derecho administrativo, como decíamos, es uno de los protagonistas de este cambio ya que la aplicación de tecnología disruptiva a los procesos administrativos puede encontrar una repercusión directa en los derechos y en las garantías de los administrados (entre otros los artículos 14, 18 y 24 de la Constitución Española). Asimismo, es indudable el cambio de gestión pública al que está asistiendo —y participando— el sector público español. Un modelo que en palabras recientes de Ramió camina hacia una burocracia elástica. Resulta muy interesante la visión del autor que menciona lo que es totalmente visible y que en ocasiones no todos nos atrevemos a comentar:

“En la gestión pública moderna la excepcionalidad es cada vez más ordinaria. Es habitual que surjan nuevos problemas totalmente imprevistos (...), nuevas demandas ciudadanas, nuevas oportunidades de gestionar de manera diferente de lo habitual o para abrir nuevos horizontes de políticas y servicios públicos, o de exigencia de mejora de los procesos internos de gestión. En resumen: la gestión de la novedad va dejando de ser novedosa y va a pasar a formar parte de la rutina de la gestión pública”³.

³ C. Ramió. *Burocracia inteligente. Guía para transformar la Administración pública*, Catarata, 2022, p. 65.

En definitiva, lo que aparentemente es el futuro de la gestión de las Administraciones públicas es, en cambio, una realidad inclusive en aquellos temas como los algoritmos, los códigos fuente o los sistemas de IA que parecen mantener todavía ciertas características de novedad. Lo que sí es trascendental en esta materia es la respuesta jurídica que debe ofrecer el derecho. Es aquí donde se hace patente la necesidad de una Administración pública flexible, precisamente “elástica”.

En este estudio, nos interesa analizar el acceso —o la denegación— a la información del código fuente que, en el sector público, utilizan determinados sistemas tecnológicos y que finalmente colaboran en la toma de decisiones administrativas relevantes.

Debemos reseñar que cuando hablamos de código fuente no hacemos referencia necesariamente a un sistema de inteligencia artificial, sino que nos referimos a modelos cuyos códigos fuente permiten el funcionamiento de un sistema algorítmico más complejo.

II. EL CÓDIGO FUENTE COMO ELEMENTO DEL MODELO FUNCIONAL

Ahora bien, el código fuente es uno de los elementos que componen un modelo funcional. Así, debemos pensar en el código fuente según lo que Arnalboldi define como:

“La serie de instrucciones que componen a un programa en lenguaje de programación. Pero en este primer estado no es directamente ejecutable por la computadora, sino que debe ser traducido a otro lenguaje y de esa forma será más fácil para la máquina interpretarlo y llevar adelante las instrucciones que el mismo contiene”⁴.

En otras palabras, el código fuente representa una interconexión entre elementos. Una fórmula de la que el usuario no conoce las instrucciones o, como señala Soriano Arnaz, identifica la traducción informática de las instrucciones que permiten *operacionalizarlas* en código (lenguaje informático)⁵. Como probablemente se habrá podido

⁴ F.A. Arnalboldi, *Código fuente*. Publishing Platform, 2004, pp. 1-44.

⁵ A. Soriano Arnaz, “Decisiones automatizadas y discriminación: aproximación y propuestas generales” *Revista General de Derecho Administrativo*, núm. 56,

observar, en un modelo funcional predefinido, el solo conocimiento de un código informático puede no ser suficiente para que el mismo funcione y opere según lo deseado. En este sentido y a sabiendas de que la conformación técnica de un modelo algorítmico —y de IA— es más compleja, podemos resumir los que los ingenieros computacionales nos han transmitido para la comprensión básica de un modelo funcional.

Así, debemos de entender un modelo funcional según el esquema básico que someramente proponemos:

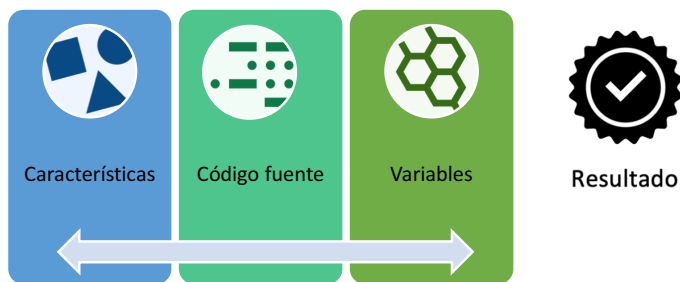


Fig. 1. Esquema básico de modelo funcional. Elaboración propia según conversación mantenida con Javier Jiménez de la Jara, Ingeniero computacional de la Universidad de Cádiz, especialista en *machine learning*

De esta manera, para que el modelo funcione necesita de, por lo menos, tres elementos: las características, el código fuente y las variables. Todos ellos mantienen una interrelación. A modo de ejemplo: podríamos crear un modelo funcional mediante el cual reconocer la tipología de un animal, supongamos, el gato persa. En este caso, el modelo tendría una serie de características (la forma de la cara del animal, el pelaje, color de los ojos, etc.). Asimismo, el modelo debería tener unas variables específicas que identifiquen y pormenorizen esta raza (color del pelo, los tipos de gato persa, el tamaño, las extremidades, etc.). El código fuente actuaría como factor multiplicador. En otras palabras, serviría para “conectar” las características con las variables y de esta manera otorgar un resultado óptimo. En el ejemplo,

el modelo podría establecer que las fotos que se le proponen al sistema representan un gato persa himalayo por tener una determinada forma de la cara, el pelo del cuerpo blanco, nariz negra, ojos azules, etc., excluyendo por ende otras tipologías de gato persa como el *solid*, *silver*, *golden smoke*, *tabby* etc.

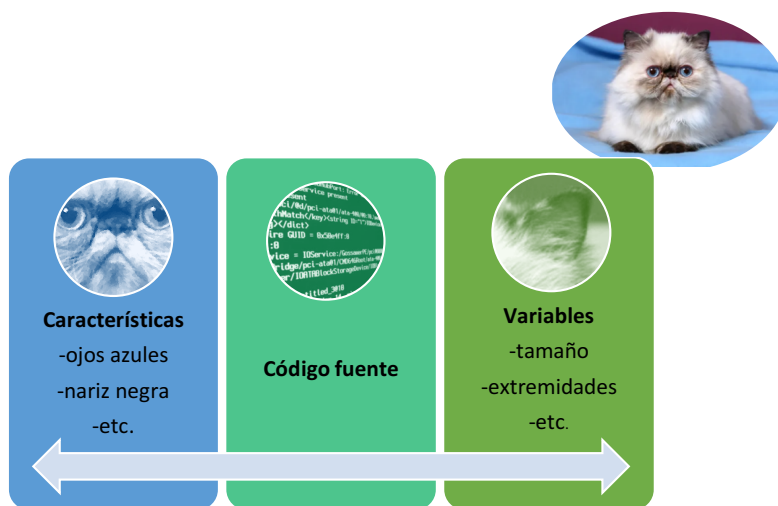


Fig. 2. Esquema aplicado de modelo funcional (ejemplo). Elaboración propia

Entendemos, por lo tanto, que, en principio, los tres elementos son entre ellos esenciales y complementarios o lo que es lo mismo, uno de ellos no prevalece sobre los demás, es decir, si de la ecuación elimináramos uno de los tres, el modelo podría no funcionar correctamente. Por estas razones, el solo conocimiento del código fuente no equivaldría al conocimiento íntegro del modelo funcional ya que, como ha sido aludido, sería necesario conocer también las características y las variables.

Esto, desde un punto de vista técnico —muy básico y seguramente no privado de errores y de necesarias y más puntuales explicaciones sistemáticas— reduce el interés para el código fuente. Ahora bien, desde la perspectiva jurídica, el escenario cambia y lo que nos enseña la técnica parece no casar con las exigencias jurídicas. En realidad, debemos hacer un esfuerzo para entender que el conocimiento inte-

gral del modelo, imprescindible para las ciencias técnicas, no es óbice para el derecho sobre todo cuando del conocimiento de una parte del modelo puede depender el ejercicio de determinados derechos. En este sentido, el derecho debe afrontar el desafío de establecer lo que debe ser transparente. Así, la transparencia del modelo funcional o de algunos de sus elementos parece convertirse en prioritaria cuando el interesado decide ejercer sus derechos. En todo esto hay, quizá, una base de partida relativamente general. Sangüesa señala que:

“La transparencia de datos y de algoritmos (transparencia algorítmica para abreviar) implica la capacidad de saber qué datos se utilizan, cómo se utilizan, quiénes los utilizan, para qué los utilizan y cómo se llega a partir de los datos a tomar las decisiones que afectan a la esfera vital de quien reclama esta transparencia. Si una persona ha sido rechazada en algún proceso (por ejemplo, no recibe una beca o un crédito), debería saber a partir de qué datos se ha tomado esa decisión y cómo se ha decidido excluirla, que es una cosa diferente. Igualmente, si un sistema de reconocimiento la ha clasificado como sospechosa de terrorista. Hoy en día, una esfera pública informada debería estar compuesta por agentes capaces de averiguar el subtexto del universo algorítmico en el que se desarrollan los ciudadanos como sujetos económicos y políticos”⁶.

En este mismo sentido, acertadamente, Cerrillo avisa de que la transparencia pública es un principio que debe gobernar el uso de los algoritmos que en ocasiones pueden ser opacos. Ahora, para evitar la opacidad es necesario que las administraciones públicas adopten medidas específicas que permitan no sólo el acceso a los algoritmos, sino también el conocimiento de su contenido y la comprensión de su funcionamiento y del fundamento sobre el que se asienta la decisión final. Esto hace que el autor mencionado abogue por el acceso al código fuente, al conocimiento de la documentación que contenga la descripción del algoritmo y de los datos que utiliza⁷.

⁶ R. Sangüesa, Inteligencia artificial y transparencia algorítmica: “*It’s complicated*” en *BiD: textos universitaris de biblioteconomia i documentació*, núm. 41 (deseembre). <<https://bid.ub.edu/41/sanguesa.htm>>. DOI: <https://dx.doi.org/10.1344/BiD2018.41.11> [Consulta: 23-05-2022].

⁷ Véase: A. Cerrillo i Martínez, “Com obrir les caixes negres de les administracions públiques? Transparència i rendició de comptes en l’ús dels algorismes” *Revista Catalana de Dret Públic*, núm. 58, junio 2019, p. 18.

Como podemos imaginar, la transparencia algorítmica es uno de los temas centrales de debate mediante el cual los sistemas algorítmicos pueden ser más proactivos. En este sentido, Busuioc señala que:

“Aunque es poco probable que la transparencia de los modelos resuelva por sí misma todos los problemas de información relacionados con el uso de los algoritmos de IA, la dependencia de los modelos patentados en el sector público engendra y agrava una fuerte dependencia de que los proveedores privados informen con veracidad sobre el funcionamiento (y el mal funcionamiento) de sus modelos, lo que resulta muy problemático dados los considerables costes financieros y de reputación que están en juego al configurar los incentivos de divulgación. Sin la transparencia de los modelos, los terceros independientes no podrán auditar de forma independiente el funcionamiento de los algoritmos (por ejemplo, probando el funcionamiento y las predicciones de los algoritmos con diferentes datos) y/o se quedarán sin pronosticar las características clave cuando intenten hacer ingeniería inversa del funcionamiento de los algoritmos, mientras que los organismos del sector público no podrán cumplir con sus obligaciones de divulgación administrativa hacia los ciudadanos afectados”⁸.

En nuestra opinión, debemos además ampliar el concepto de transparencia que en realidad no puede solo limitarse al conocimiento del código fuente (que en este caso es lo que aquí interesa), sino que la transparencia debe considerar una serie de elementos que entre ellos resultan ser complementarios. Recientemente, el *United Kingdom Central Digital and Data Office* publicó el denominado *Algorithmic Transparency Standard* que tiene el propósito de identificar una transparencia holística (o plena). De hecho, indica que la transparencia algorítmica debe:

- 1) Explicar el funcionamiento del modelo y el problema que esta trata de resolver.
- 2) Debe detallar los perfiles técnicos con especial indicación de quién es el responsable del proceso, el alcance de la herramienta (por ejemplo, la arquitectura del sistema), cómo la herramienta toma la decisión —considerando la participación humana—, los datos personales, la evaluación de impacto (especialmente

⁸ M. Busuioc, “Accountable Artificial Intelligence: Holding Algorithms to Account”, *Public Administration Review*, vol. 81, núm. 5, 2020, pp. 829–830. [DOI: 10.1111/puar.13293].

la ética de la IA), la descripción de los riesgos del modelo que podrían desembocar en sesgos, perjuicios o daños y, finalmente, menciona también los controles a los que se someterá el modelo, es decir una verdadera auditoría⁹.

Como puede entenderse, al concepto tradicional de transparencia, se le suma el de conocimiento y explicabilidad o mejor aún, la explicabilidad es un elemento de la transparencia. Sobre ello, indica Ortiz de Zárate Alcarazo:

“En este sentido, la transparencia sería la habilidad para hacer visible los componentes de un sistema de IA y sería condición necesaria, pero no suficiente, para que se cumpliera con el principio de explicabilidad. Para esto es necesario, como hemos mencionado, un ejercicio de interpretación posterior que haga comprensible para los seres humanos todo lo que tiene lugar dentro del sistema”¹⁰.

El derecho también tiene su parte de protagonismo ya que interviene en la gestión jurídica de lo que el algoritmo produce sobre todo cuando es necesario verificar la legitimidad de una decisión tomada automáticamente por una máquina o, lo que es más importante, que el algoritmo haya servido como herramienta consustancial para la toma de una determinada decisión. De hecho, Carloni señala que es importante verificar si la decisión:

“ [...] sea efectivamente adoptada en ausencia de una intervención humana en la fase de toma de decisiones, pero también al hecho de que la operación automatizada efectuada tenga un papel decisivo en la tramitación del procedimiento, y la corrección del “razonamiento” seguido por parte del funcionario instructor no sea susceptible de revisión. En el primer caso, el procedimiento y la decisión se realizan en ausencia (sustancial) del funcionario “humano”; en el segundo, el funcionario interviene, pero se halla en una posición de asimetría con respecto al ordenador y está, de hecho, “al servicio” de un procedimiento que madura y se lleva a cabo en un entorno sobre el que él no tiene control”¹¹.

⁹ Puede consultarse: <https://www.gov.uk/government/collections/algorithmic-transparency-standard> [consultado el 20 de junio de 2022].

¹⁰ L. Ortiz De Zárate Alcarazo, “Explicabilidad (de la inteligencia artificial)”, *Eunomía. Revista en Cultura de la Legalidad*, núm. 22, p. 334. DOI: <https://doi.org/10.20318/eunomia.2022.6819>.

¹¹ E. Carloni, “IA, algoritmos y Administración pública en Italia”, *Revista de Inter-net, Derecho y Política*. Dossier «El derecho para una inteligencia centrada en el

Por todo lo anterior, el derecho rompe con la visión holístico-técnica del modelo funcional —aunque no desconoce la importancia de esta visión integrada— favoreciendo el análisis de los elementos singulares que lo componen. Es así, en nuestra opinión, porque solo el interesado o su representante (o defensor), conoce la estrategia de sus actuaciones, de modo que solo él puede establecer las informaciones —y el alcance de ellas— que necesita recabar para su defensa y estas pueden estar representadas por los elementos aludidos —por separado— o en su conjunto.

En realidad, lo que acabamos de mencionar es, en parte lo que ha ocurrido en el conocido caso denominado bono social *vs.* Fundación Civio, caso que encuentra su punto álgido en la denegación de acceso al código fuente según, especialmente y sobre todo en la fase inicial de la cuestión, la aplicación del artículo 14.1 j de la Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno.

III. EL ACCESO AL CÓDIGO FUENTE ANTE LA LEY 19/2013, DE 9 DE DICIEMBRE (LTAIBG)

En relación con el acceso al código fuente conviene analizar lo ocurrido en el caso bono social *vs.* Fundación Civio¹². Concretamente, reseñamos lo siguiente: para poder obtener el bono social eléctrico, el Gobierno de España creó una aplicación informática denominada BOSCO. Dicho software fue cedido a las compañías eléctricas para que pudieran utilizarlo para verificar los requisitos de los solicitantes. En este escenario, la Fundación Ciudadana Civio, tras recibir quejas de distintos usuarios que vieron denegada la concesión del bono como consecuencia de la aplicación del sistema algorítmico, solicita al Ministerio para la Transición Ecológica acceso a determinada información del sistema BOSCO. El Ministerio para la Transición Ecológica da traslado de dicha solicitud a la Subdirección General

¹² ser humano y al servicio de las Instituciones», núm. 30, marzo, 2020, pp. 1-12. Señalamos que una explicación más extensa del tema que analizaremos se ha presentado como artículo de revista. En la actualidad dicho trabajo se encuentra en evaluación. Para este texto se ha procedido a efectuar una adaptación.

de Tecnologías de la Información y las Comunicaciones del Ministerio de Industria, Comercio y Turismo que a su vez solicitó informe y trasladó posteriormente la petición a la Subdirección General de la Administración Digital.

La Fundación Ciudadana Civio requirió las siguientes informaciones:

- a) La especificación técnica de dicha aplicación.
- b) Los resultados de las pruebas realizadas para comprobar que la aplicación implementada cumple la especificación funcional.
- c) El código fuente de la aplicación actualmente en producción.
- d) Cualquier otro entregable que permita conocer el funcionamiento de la aplicación.

Debido a la falta de respuesta, la Fundación Ciudadana Civio interpela al Consejo de Transparencia y Buen Gobierno que establece el derecho a conocer las especificaciones técnicas de la aplicación y los resultados de las pruebas, sin embargo, deniega el acceso al código fuente asumiendo que es de aplicación el artículo 14.1 j de la Ley 19/2013 en materia de propiedad intelectual. Recurrida judicialmente esta decisión, la Sentencia 143/2021 del Juzgado Central de lo Contencioso Administrativo reitera la denegación del acceso basándose, principalmente, en los informes técnicos que relacionan el conocimiento del código fuente con la posible vulnerabilidad por ciberataques del sistema. Hasta donde sabemos, la Fundación Ciudadana Civio ha decidido recurrir esta sentencia.

En este contexto, es interesante verificar los elementos que deberían prevalecer a la hora de admitir, denegar o limitar el acceso a la información y que quizá, en el caso aludido no han sido tenidos adecuadamente en cuenta¹³. El CTBG, en el punto V de la Resolución CI/002/2015 de 24 de junio de 2015, sobre los criterios interpretativos de los límites del derecho de acceso a la información, expone que:

“Los límites a que se refiere el artículo 14 de la LTAIBG, a diferencia de los relativos a la protección de los datos de carácter personal, no se

¹³ Sobre los límites del acceso a la información se recomienda la lectura de: L. Cotino, y A. Boix Palop (coords.), *Los límites al derecho de acceso a la información pública*, Tirant lo Blanch, Valencia, 2021.

aplican directamente, sino que de acuerdo con la literalidad del texto del número 1 del mismo, “podrán” ser aplicados. De esta manera, los límites no operan ni automáticamente a favor de la denegación ni absolutamente en relación con los contenidos. La invocación de motivos de interés público para limitar el acceso a la información deberá estar ligada con la protección concreta de un interés racional y legítimo. En este sentido su aplicación no será en ningún caso automática: antes, al contrario, deberá analizarse si la estimación de la petición de información supone un perjuicio (*test* del daño) concreto, definido y evaluable. Este, además no podrá afectar o ser relevante para un determinado ámbito material, porque de lo contrario se estaría excluyendo un bloque completo de información. Del mismo modo, es necesaria una aplicación justificada y proporcional atendiendo a la circunstancia del caso concreto y siempre que no exista un interés que justifique la publicidad o el acceso (*test* del interés público)”.

Por eso, es necesario establecer criterios que nos permitan evaluar si el conocimiento de determinada información podría dañar o perjudicar el bien jurídico que la norma en materia de transparencia busca proteger. El *test* de daño establece que el daño debe ser: a) concreto o lo que es lo mismo, real, b) definido, es decir, inconfundible e inequívoco y c) evaluable, que identifica la posibilidad de medirlo con precisión. Asimismo, deberá considerarse lo que es el interés público sobre el interés particular. Ahora bien, puesto que los límites introducidos por la LTAIBG no encuentran aplicación automática, asume particular relevancia la interpretación que hay que hacer de los mismos que deberá ser estricta, cuando no restrictiva¹⁴.

Por estas razones, en la materia que nos ocupa es posible realizar lo que podríamos llamar *test* de verificación. Una prueba que, con miras al caso bono social-Fundación ciudadana Civio, permitiría responder a determinadas preguntas. En la práctica:

- a) ¿El acceso al código fuente representa un daño real?
- b) ¿El acceso al código fuente determinaría un daño inconfundible y/o inequívoco?

¹⁴ Puede verse M. I. Martín Del Llano, *Consejo de Transparencia y buen Gobierno: estructura orgánica y eficacia de sus resoluciones* en S. Fernández Ramos, J. M. Pérez Monguió, y A. Galdámez Morales (coords.), *Nuevos enfoques sobre transparencia y derecho de acceso a la información pública*, Thomson Reuters-Aranzadi, 2022, pp. 68-69.

- c) ¿Es posible medir el daño que provocaría el acceso al código fuente?

Naturalmente y como ha sido referido anteriormente, también, es interesante verificar, mediante un ejercicio de ponderación, si en el caso analizado prevalece el interés público general o el interés particular. Todo parece apuntar a que las respuestas al *test* de verificación son negativas. Esta afirmación, en realidad, es la consecuencia de la lectura y análisis del informe presentado por el Subdirector General de Tecnologías de la información y de las Comunicaciones del Ministerio de Industria Comercio y Turismo (acontecimiento n. 76 del expediente judicial electrónico), que en materia de seguridad de la aplicación informática BOSCO, señala:

“Se considera que este tipo de actividades para reforzar la Seguridad por Diseño (y de esta forma, prescindir de la Seguridad por Oscuridad para poder liberar el código fuente mediante una licencia de fuentes abiertas), tiene también sentido de un punto de vista de un análisis coste-beneficio en el caso de aplicaciones con alto grado de reutilización por parte del sector público y privado. En el caso que nos ocupa, la aplicación ha sido diseñada con un propósito específico que difícilmente sería reutilizada por otras administraciones (la competencia energética no está transferida) y evidentemente en ningún caso por el sector privado. Por todo ello se considera que el equilibrio entre las medidas de Seguridad por Diseño y de Seguridad por Oscuridad que se han implementado en el desarrollo y construcción de la aplicación son los correctos y que, por razones de seguridad pública y seguridad nacional, el código fuente de la aplicación debe permanecer custodiado por la Administración y no debe procederse a su distribución”.

Asimismo, el citado informe y según señala la sentencia 143/2021 del Juzgado Central de lo Contencioso Administrativo, “la entrega del código fuente haría la aplicación sensible a ataques por vulnerabilidad “de día cero” o que están aún por descubrir en el momento en que se construye el producto software”.

La lectura del informe no arroja ningún tipo de seguridad fáctica. De hecho, el mismo documento no señala terminantemente la provocación de un daño causado como consecuencia del conocimiento del código fuente. Esto es así porque sencillamente no puede hacerlo ya que este posible suceso no se puede determinar *a priori*. En otras palabras y al menos en este punto, la denegación del acceso se basa sobre acontecimientos indefinidos y no concluyentes que en cambio

ponen de manifiesto la aplicación, *a contrario sensu*, de los criterios interpretativos del CTBG. En este mismo sentido, la aplicación del *test* de verificación, como mínimo, hace vislumbrar que es imposible asumir que el conocimiento del código fuente produzca un daño real e inequívoco (inconfundible).

Además, tampoco resultaría suficiente la admisión del posible mal uso de la aplicación BOSCO en directa relación con el derecho de protección de la propiedad intelectual. Precisamente en este sentido se pronunció la Comisión de Garantía del derecho de Acceso a la información Pública de Cataluña en la resolución de 23 de diciembre de 2015. Se trataba de un procedimiento de selección de docentes universitarios. Uno de los candidatos solicitó acceso al expediente presentado por otros participantes. Dicha solicitud de acceso se fundaba en el conocimiento necesario de los expedientes de los concurrentes que le permitieran al recurrente presentar el correspondiente recurso administrativo. La universidad, inicialmente, denegó el acceso invocando la presencia de derechos de propiedad intelectual. Más especialmente, la institución universitaria alegó que en el expediente el recurrente hubiera podido conocer trabajos originales cuya copia —por añadidura nuestra: conocimiento— no aportaría información útil mientras que sí suponía riesgos de apropiación o mal uso de los trabajos mencionados. En este caso, la comisión concluye que realmente no se trata de un daño efectivo, sino de una posibilidad meramente hipotética. La Comisión, en su dictamen, menciona otrosí la sentencia del Tribunal Supremo 3583 de 6 de junio de 2005 (Roj: STS 3583/2005-ECLI:ES:TS:2005:3583), que en un caso análogo señaló que la mera sospecha sobre la posibilidad de presumir el mal uso o la conducta antijurídica, no es motivo para limitar el derecho a obtener copia de los documentos públicos. Señala el juez: “[...] no cabe oponer, como hace el Letrado de las Cortes Generales, las consecuencias que se podrían producir en función de la utilización que el recurrente haga del conocimiento que va a obtener y de las copias que va a recibir. De ello será, ciertamente responsable el propio actor, pero no hay razón para presumir que va a conducirse de manera antijurídica (FJ 7^a)¹⁵”.

¹⁵ Sobre el acceso al código fuente pueden verse también, las sentencias dictadas en los casos Danfoss y Galina Meister del TJUE (Sentencia TJUE de 17 de octubre de 1989, asunto 109/88, Handels-og Kontorfunktion rernes Forbund i Danmark

Una vez más, en relación con el *test* de daño, la denegación del acceso carece de fundamento haciendo complicada la aceptación del escenario según el cual se deniega el conocimiento del código fuente para proteger al sistema de posibles, eventuales e indeterminables ciberataques o difusión de informaciones relativas al sistema de funcionamiento. Canals Ametller señala que “lo que es inteligente, es inseguro, pero no pasa nada, es decir, es necesario, simplemente, se trata de aumentar la seguridad digital”¹⁶. En definitiva, si existe un temor a posibles ataques cibernéticos debidos a la revelación del código fuente deberá ser la Administración correspondiente quien gestione el riesgo a sabiendas de que el riesgo cero no existe. Así, en relación con los límites de acceso, se establece un verdadero régimen estricto. La misma doctrina jurisprudencial del Tribunal de Justicia de la Unión Europea (TJUE), en el caso del Consejo de la Unión Europea vs. Sophie in’t Veld determina que: “dado que tales excepciones —las consagradas en el artículo 4 del Reglamento— invalidan el principio del mayor acceso posible del público a los documentos, deben interpretarse y aplicarse en sentido estricto”¹⁷.

En otras palabras, tenemos que desvincular la existencia de un algoritmo con la idea de perfección o de lo que Moreno Rebato identifica como “solidez técnica” de los sistemas de IA, apostando, más bien, por el enfoque de la previsión del riesgo minimizando los daños involuntarios, pero también considerando la necesidad de que el sistema resista a ataques malintencionados¹⁸. En definitiva, señala Guichot:

y Dansk Arbejdsgiverforening, en nombre de Danfoss y Sentencia TJUE de 19 de abril de 2012, asunto 415/10, Galina Meister y Speech Design Carrier Systems GmbH.

¹⁶ Ponencia de D. Canals Ametller, «*Ciberseguridad y Gobiernos Locales*». Jornada *Inteligencia artificial y Gobiernos Locales*. Diputación de Sevilla, Fundación Democracia y Gobierno Local, Sevilla 23 de marzo de 2022.

¹⁷ Véase: E. Pérez Conchillo, P. Cruz Mantilla de los Ríos, A. González Moro, A. Galdámez Morales y J. Sierra Rodríguez, “Panorámica jurisprudencial sobre el derecho de acceso a la información pública en el ámbito europeo”, *Revista de Derecho Político* (UNED), núm. 113 enero-abril, 2022, p. 212.

¹⁸ En este sentido Véase M. Moreno Rebato, *Inteligencia artificial (umbrales éticos, derecho y Administraciones públicas)*, Thomson Reuters-Aranzadi, Cetinia, Cizur Menor (Navarra), p. 40.

“Aparece aquí como clave de bóveda el principio de ponderación, que es la espita que permite en el caso concreto valorar la importancia que para la sociedad tiene conocer la información en cuestión, incluso si con ello se causa un perjuicio para uno de sus bienes”¹⁹.

Esto no es todo. La falta de referencias legislativas consolidadas y aplicables a la materia —más allá del listado del artículo 14.1 de la LTAIBG—, permitiría a las administraciones una excesiva autonomía que podría limitar lo que es el necesario estudio de los intereses contrapuestos, es decir, acceso y denegación de este. En este sentido, señala, Milazzo, la discrecionalidad de la Administración podría llegar a ser verdaderamente amplia²⁰. En realidad, es lo que ocurre con el caso que nos ocupa: la denegación de acceso al código fuente parece ser discrecional y aunque es precisamente la Administración pública la que tiene reconocida la discrecionalidad administrativa, esta, en todo caso, deberá estar justificada por los fines que persigue y que la fundamentan. En nuestra opinión, referirse genéricamente al listado del artículo 14.1 j de la Ley de Transparencia sin mantener un fundamento basado en lo que anteriormente definimos como el *test* de verificación, no equivale a fundamentar la decisión de denegar el acceso. Por añadidura, nos sumamos a la idea de Huergo Lora quien sabiamente apunta que, precisamente los algoritmos limitan la discrecionalidad administrativa, pues, debemos señalar que estos deberán ser necesariamente transparentes y permitir su conocimiento, hasta del andamiaje algorítmico²¹.

¹⁹ E. Guichot Reina, “Los límites de la transparencia y el derecho de acceso a la información” en I. Martín Delgado (dir.), *Transparencia y acceso a la información pública: de la teoría a la práctica*, Iustel, 2019, p. 59.

²⁰ P. Milazzo, “L’istituto simbolo della trasparenza amministrativa in Italia: il diritto di accesso alle informazioni della Amministrazione. Tra evoluzione normativa tumultuosa, dubbi interpretativi ed incertezza di tutela” en P. Pérez Tremps, y M. Revenga Sánchez (coords.), *Transparencia, acceso a información pública y lucha contra la corrupción. Tres experiencias a examen: Brasil, Italia, España*. Tirant lo Blanch, p. 110.

²¹ Puede verse: A. Huergo Lora, “Administraciones públicas e inteligencia artificial: ¿más o menos discrecionalidad?”, *El Cronista del Estado social y democrático de derecho*, núm. 96-97, Iustel, 2021, p. 93.

IV. EL SESGO ALGORÍTMICO COMO BASE DE PARTIDA

En un sistema basado en algoritmos y en IA, debemos entender el sesgo como un elemento más de la ecuación de funcionamiento. En el ámbito que nos corresponde, solemos asociar el sesgo con la discriminación negativa que puede producir el modelo funcional. Desde luego este es un escenario que puede ocurrir, sin embargo, en realidad, el sesgo es un error sistemático en el que puede incurrir el modelo funcional. En otras palabras, debemos considerar que la máquina que consideramos inteligente e insuperable también comete errores que precisamente la inteligencia humana —es decir lo contrario de la inteligencia artificial— está llamada a corregir. Está claro que, en el sentido explicado, aludimos a errores y sesgos involuntarios e impredecibles de quien programa la máquina. Un contexto distinto lo encontramos cuando el sesgo se produce arbitrariamente transformándose en una discriminación deliberada de quien programa el sistema²². En este sentido, destacan por lo menos tres patrones que vale la pena mencionar. El primero: según el estado actual del conocimiento técnico y tecnológico no existen máquinas perfectas. En segundo lugar: los sistemas algorítmicos deben pasar por auditorías periódicas que procedan a verificar el correcto funcionamiento y desde luego la ética del sistema. Finalmente, en tercer lugar, la existencia de un sistema basado en algoritmos y/o IA no puede prescindir de la aportación humana especialmente cuando fuese necesaria la corrección del error²³.

Ahora bien, naturalmente, no podemos conformarnos con la imperfección de los sistemas algorítmicos, todo lo contrario. Conocido este límite, es necesario vigilar para que se puedan detectar y corregir los errores que finalmente pueden desembocar en una vulneración de derechos y garantías. Una vez más, las fórmulas jurídicas representan

²² En un sentido parecido se pronuncia P. Zuddas, “Decisión algorítmica y principio de igualdad”, *Revista General de Derecho Administrativo*, núm. 60, 2022, pp. 4-5.

²³ En materia de reserva de humanidad, véase: J. Ponce Solé, “Inteligencia artificial, derecho administrativo y reserva de humanidad: algoritmos y procedimiento administrativo debido tecnológico”, *Revista General de Derecho Administrativo*, núm. 50, 2019, p.12.

una de las herramientas que pueden colaborar con este proceso de control. De hecho, en la actualidad, existen ciertos principios sobre los que asentar la construcción de sistemas proactivos y conocibles. Es necesario comenzar desde las bases entendiendo que los ciudadanos debemos tener, como mínimo, el derecho a lo que podríamos llamar el *conocimiento prudencial*, es decir, el hecho de que un sistema automatizado está operando con y sobre nosotros y, por ende, la comprensión básica de su funcionamiento. Como decíamos, algunas normas colaboran para que esto ocurra. Es el caso de los artículos 13.2 f) y 15.1 h) Reglamento (UE) 2016/679 del parlamento europeo y del consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales (RGPD) que señalan el derecho del interesado a que se le facilite información cuando sus datos personales se obtengan de forma automatizada (art. 13) y, asimismo, el derecho de acceso cuando los datos estén siendo tratados de forma automatizada (art. 15). En definitiva, el RGPD quiere garantizar el derecho al respeto de la privacidad del interesado en todo su alcance. De esta forma se ha pronunciado el Tribunal de Distrito de La Haya en la conocida como “Sentencia SyRI”. El Tribunal señalaba que este software —un modelo de análisis utilizado por el Gobierno para rastrear posibles fraudes al Estado— no respetaba la privacidad del ciudadano, vulnerando el artículo 8 de la Convención Europea de Derechos Humanos relativo al derecho a la vida privada²⁴.

También vale la pena citar el artículo 22 del RGPD que introduce la fórmula: Decisiones individuales automatizadas, incluida la elaboración de perfiles. Especialmente, el apartado primero de esta norma señala: “Todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar”. Una vez más la automatización en el *centro de la diana* que en definitiva intensifica las exigencias de transparencia²⁵.

²⁴ El Tribunal además aborda la cuestión del sesgo algorítmico. De hecho, señala que «SyRI puede tener efectos no deseados, como estigmatizar y discriminar a la ciudadanía, por la ingente cantidad de información que recoge».

²⁵ En este sentido véase: M. Medina Guerrero, “El derecho a conocer los algoritmos utilizados en la toma de decisiones. aproximación desde la perspectiva

También en el ámbito europeo, la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial parece apostar por una norma proactiva en clara garantía de los derechos de los ciudadanos europeos. Esta voluntad se nota especialmente con la introducción de la clasificación de los sistemas de IA, estableciendo prácticas de IA prohibidas, de alto riesgo y de riesgo limitado. Especialmente interesantes son los dos primeros sistemas, que podríamos entender como un semáforo que indica lo que se puede o no se puede hacer. El artículo 5 de la propuesta establece tajantemente:

“Estarán prohibidas las siguientes prácticas de inteligencia artificial:

a) La introducción en el mercado, la puesta en servicio o la utilización de un sistema de IA que se sirva de técnicas subliminales que trasciendan la conciencia de una persona para alterar de manera sustancial su comportamiento de un modo que provoque o sea probable que provoque perjuicios físicos o psicológicos a esa persona o a otra.

b) La introducción en el mercado, la puesta en servicio o la utilización de un sistema de IA que aproveche alguna de las vulnerabilidades de un grupo específico de personas debido a su edad o discapacidad física o mental para alterar de manera sustancial el comportamiento de una persona que pertenezca a dicho grupo de un modo que provoque o sea probable que provoque perjuicios físicos o psicológicos a esa persona o a otra.

c) La introducción en el mercado, la puesta en servicio o la utilización de sistemas de IA por parte de las autoridades públicas o en su representación con el fin de evaluar o clasificar la fiabilidad de personas físicas durante un período determinado de tiempo atendiendo a su conducta social o a características personales o de su personalidad conocidas o predichas, de forma que la clasificación social resultante provoque una o varias de las situaciones siguientes: 1) un trato perjudicial o desfavorable hacia determinadas personas físicas o colectivos enteros en contextos sociales que no guarden relación con los contextos donde se generaron o recabaron los datos originalmente; 2) un trato perjudicial o desfavorable hacia determinadas personas físicas o colectivos enteros que es injustificado o desproporcionado con respecto a su comportamiento social o la gravedad de este.

d) El uso de sistemas de identificación biométrica remota “en tiempo real” en espacios de acceso público con fines de aplicación de la ley, salvo y en la medida en que dicho uso sea estrictamente necesario para alcanzar uno o varios de los objetivos siguientes: 1) la búsqueda selectiva de posibles víctimas concretas de un delito, incluidos menores desaparecidos; 2) la prevención de una amenaza específica, importante e inminente para la vida o la seguridad física de las personas físicas o de un atentado terrorista; 3) la detección, la localización, la identificación o el enjuiciamiento de la persona que ha cometido o se sospecha que ha cometido alguno de los delitos mencionados en el artículo 2, apartado 2, de la Decisión Marco 2002/584/JAI del Consejo [...]”.

Pero también:

“El uso de sistemas de identificación biométrica remota “en tiempo real” en espacios de acceso público con fines de aplicación de la ley para conseguir cualquiera de los objetivos mencionados en el apartado 1, letra d), tendrá en cuenta los siguientes aspectos:

a) la naturaleza de la situación que dé lugar al posible uso, y en particular la gravedad, probabilidad y magnitud del perjuicio que se produciría de no utilizarse el sistema;

b) las consecuencias que utilizar el sistema tendría para los derechos y las libertades de las personas implicadas, y en particular la gravedad, probabilidad y magnitud de dichas consecuencias”.

En materia de sistemas de alto riesgo, la propuesta de reglamento es otrosí clara y establece que:

“Un sistema de IA se considerará de alto riesgo cuando reúna las dos condiciones que se indican a continuación, con independencia de si se ha introducido en el mercado o se ha puesto en servicio sin estar integrado en los productos que se mencionan en las letras a) y b). [...] a) el sistema de IA está destinado a ser utilizado como componente de seguridad de uno de los productos contemplados en la legislación de armonización de la Unión que se indica en el anexo II, o es en sí mismo uno de dichos productos; b) conforme a la legislación de armonización de la Unión que se indica en el anexo II, el producto del que el sistema de IA es componente de seguridad, o el propio sistema de IA como producto, debe someterse a una evaluación de la conformidad realizada por un organismo independiente para su introducción en el mercado o puesta en servicio”.

Se trata en definitiva de apostar por una IA basada en la persona, es decir, el reglamento prioriza las personas sobre otros intereses (comerciales, económicos etc.), cumpliendo además con el respeto

de la Carta de los Derechos Fundamentales de la Unión Europea y el Derecho derivado de la Unión en materia de protección de datos, protección de los consumidores, no discriminación e igualdad de género²⁶.

Todo esto debe entenderse como la génesis de la lucha contra el sesgo algorítmico. Estas normas dibujan una verdadera línea roja que ningún modelo funcional puede superar. Es de entender, por tanto, que el catálogo de derechos fundamentales es inquebrantable, de manera que cualquier sistema que aluda a una posible vulneración de estos derechos estará abogado a la desaparición —o directamente a la imposibilidad de ser comercializado—, por lo menos en el entorno europeo.

Quisimos dedicar cierta amplitud al sesgo algorítmico porque, en nuestra opinión, guarda una estrecha relación con el conocimiento del modelo funcional y por ende de los elementos que lo componen, entre otros el código fuente.

Un algoritmo sesgado toma decisiones sesgadas o, lo que es lo mismo, podemos preguntarnos si el sistema habría tomado la misma decisión sin el sesgo inicial. Esta afirmación, que en principio parece ser redundante, en realidad despierta cierto interés puesto que los usuarios esperamos que precisamente gracias a la inteligencia de la máquina, esta sea menos discrecional y corrija los errores. No es exactamente así y por eso, si la rectificación del error no se produce de oficio, la única alternativa es que el usuario tenga acceso a los elementos del modelo funcional y pueda demostrar el error y solicitar su corrección. En este sentido, precisamente en el caso *bono social-Fundación ciudadana Civio*, el conocimiento del código fuente junto a las características, a las variables, y a las demás funciones del algoritmo, hubiera podido ser útil para la detección del posible error que desembocó en una decisión supuestamente sesgada para algunos usuarios.

²⁶ De aquí deriva cierta polémica según la cual la propuesta de reglamento alejaría a los Estados miembros de la UE de la competitividad en la materia con respecto a los países asiáticos y Estados Unidos.

V. A MODO DE CONCLUSIÓN

Debemos partir de una afirmación sencilla, pero de cierta importancia: el sector público y el funcionamiento de la Administración pública tienen especificidades tan exclusivas que los convierte en un sistema único e *irreplacable*. Una de estas características, generalísima, es que, a diferencia de cualquier otro sector, el sector público debe ejercer su actividad y su función por y para el interés público general. Es este el escenario en el que debemos considerar las cuestiones a las que aludimos a lo largo de este trabajo. Lo que acabamos de mencionar debe ser el *leitmotiv* que guía el debate alrededor del acceso al código fuente, la transparencia algorítmica o los modelos algorítmicos que se utilizan en las Administraciones públicas. Ahora bien, el concepto de interés público es indeterminado y al mismo tiempo constituye una de las formas de limitar la potestad —y la discrecionalidad— administrativa. Nos interesa especialmente este último escenario. Es decir, el auténtico debate consiste en verificar si el acceso al código fuente —y si ampliamos la perspectiva: el conocimiento y la *explicabilidad* del sistema algorítmico— es una herramienta para limitar el poder discrecional de las Administraciones públicas especialmente cuando éstas tomen decisiones opacas para el ciudadano. En nuestra opinión, las acciones que el interesado puede ejercer ante la Administración pública tras el conocimiento de los datos mencionados sí representan un posible límite a la discrecionalidad administrativa. Esto no es necesariamente negativo ni tampoco rompe con la visión tradicional de una Administración pública casi *omnipotente*. El conocimiento de las herramientas disruptivas que se utilizan en el seno de los organismos públicos debe ser entendido como el origen del ejercicio del derecho de rectificación.

Por todo esto, la Administración pública debe garantizar el acceso a todas aquellas informaciones que precisamente colaboran en que se puedan modificar las decisiones administrativas. Esta afirmación encuentra, como mínimo, una sencilla crítica: el derecho administrativo establece ya una serie de recursos mediante los cuales, precisamente, el interesado solicita la rectificación aludida. Sin duda es así, sin embargo, uno de los factores fundamentales de cualquier recurso administrativo es su fundamentación y para que esto ocurra es necesario que el recurrente tenga a disposición todos los datos que crea oportu-

nos para su defensa. Así, en el caso que nos ha ocupado, la denegación de acceso al código fuente implica que el interesado no disponga de lo que cree necesario para su defensa produciéndose, por ende, una verdadera indefensión que en última instancia podría desembocar en la vulneración de normas constitucionales como el artículo 24 de la Constitución Española.

Podemos ir más allá. En realidad, el verdadero reto no es que la Administración pública dé a conocer solo código fuente, sino que ponga a disposición de quien lo solicite todo el modelo funcional (características, código fuente y variables). Solo de esta manera se habrá logrado una transparencia plena. De hecho, todo parece apuntar a que a mayor apertura del código fuente —y del modelo funcional— existiría una mayor seguridad del sistema. Este escenario parece razonable porque precisamente a través del conocimiento se podrían rectificar errores informáticos, sesgos, etc. En otras palabras, lo desconocido no equivale a más seguro.

Ahora bien, ¿cómo lograr esta transparencia plena? Sin desconocer las importantes contribuciones del sector privado en materia disruptiva de las que se sigue beneficiando el sector público, es necesario apostar por la emancipación de la Administración pública. La transición digital debe formar determinadas categorías de empleados públicos para que produzcan sistemas disruptivos adaptados a las necesidades de los organismos en los que ejercen su labor.

No hay que descartar la utilización de los códigos fuente abiertos, que podrían representar una alternativa válida a las fórmulas totalmente cerradas.

Por último y en todo caso, la Administración pública debe desvincular el conocimiento de los sistemas disruptivos del miedo a su publicación. Naturalmente, es necesario un gran esfuerzo dirigido a aumentar constantemente la ciberseguridad, escenario éste que paulatinamente propiciaría una mayor propensión a rendir cuenta de las características técnicas de los sistemas utilizados.

En definitiva, se trata de lograr una transparencia plena basada en la persona y sus garantías sin desconocer los límites que en materia de acceso plantea, por ejemplo, la LTAIBG. Límites que como ya ha sido

explicado, deben filtrarse según el *test* de verificación, precisamente para garantizar el interés público general.

Acceso al código fuente y a los algoritmos de las Administraciones inteligentes.

Lecciones a partir de experiencias comparadas

MARÍA ESTRELLA GUTIÉRREZ DAVID¹
Profesora de Derecho Constitucional
Universidad Complutense de Madrid

I. INTRODUCCIÓN

En el Derecho comparado —al igual que en el Derecho español—, la calificación jurídica de los programas informáticos utilizados por las Administraciones públicas —junto con el código fuente que los sustenta—, y en su caso, los algoritmos subyacentes —sean o no de IA— viene siendo harto problemática. En unos casos se les califica como actos administrativos², en otros como reglamentos³, e incluso, como es el caso de la doctrina alemana, se les ha considerado como instrucciones (“*Verwaltungsvorschrift*”)⁴.

En los últimos tiempos, el debate se ha intensificado si cabe aún más precisamente con ocasión del ejercicio del derecho de acceso al código fuente y a los algoritmos utilizados por las Administraciones

¹ Se ha realizado este trabajo en el marco del Proyecto de Investigación «El impacto de la inteligencia artificial en los servicios públicos: Un análisis jurídico de su alcance y consecuencias en la asistencia sanitaria (PGC2018-098243-B-I00), del Ministerio de Ciencia, Innovación y Universidades, dirigido por José Vida Fernández

² J.B. Auby, “Algorithmes et Smart Cities: Données Juridiques”, *Revue Générale du Droit*, num. 29878, 2018, p. 21.

³ 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”. *Revista de Derecho Público: Teoría y Método*, vol. 1, 2020, p. 75. DOI: 10.37417/RPD/vol_1_2020_33.

⁴ Vid. E. Buoso, “Fully Automated Administrative Acts in the German Legal System”, *European Review of Digital Administration & Law*, vol. 1, núm. 2, 2020, p. 121.

en el contexto de la llamada “Administración 4.0”, caracterizada por la automatización, la interconexión y el uso de programas informáticos que incorporan IA, capaces no sólo de “realizar operaciones de cálculo complejas en poco tiempo, sino dotados de la capacidad de aprender de sus errores y, en cierta medida, de emular el funcionamiento de la mente humana”⁵.

Desde el punto de vista de los regímenes de transparencia y de acceso a la información pública, la cuestión se ha centrado en dilucidar si el código fuente y, en su caso, los algoritmos, pueden ser considerados como “información pública”; y, en caso afirmativo, si dicho acceso podría verse afectado por algún límite legal, normalmente, por motivos de seguridad pública⁶ (lo que incluye la seguridad de los sistemas de información de la propia Administración⁷); la prevención, investigación y sanción de los ilícitos penales, administrativos o disciplinarios (incluyendo el riesgo de elusión de la Ley)⁸; la protección de la propiedad intelectual y los intereses comerciales⁹.

Se viene argumentando, no obstante, que el derecho de acceso al código fuente y, en su caso, al algoritmo subyacente, no es el instrumento adecuado para garantizar una adecuada transparencia de las decisiones automatizadas. Sin embargo, un análisis de la casuística comparada relativa al derecho de acceso evidencia cómo las resoluciones estimatorias a la información técnica asociada a esta clase de decisiones —no sólo el código fuente o el algoritmo, sino también las especificaciones técnicas, la caracterización de los *dataset* utilizados como inputs de los modelos predictivos o las auditorías de terceros— permiten el escrutinio ciudadano y, en particular, la detección de algunas de las patologías —errores en el diseño y en la validación de los modelos algorítmicos, los sesgos discriminatorios— que aquejan a la actividad automatizada y algoritmizada de las organizaciones pú-

⁵ D. Ponte y G. Pernice, *L'intelligenza artificiale e l'algoritmo a contatto col diritto amministrativo: rischi e speranze*, Consiglio di Stato, 2021.

⁶ CADA. Opiniones 20163619, de 20 de octubre de 2016; y, 20200496, de 12 de marzo de 2020.

⁷ Sheridan v. U.S. Office of Pers. Mgmt., 278 F. Supp. 3d 11 (D.D.C. Sept. 29 2017) (Jackson, J.)

⁸ *Ibidem*.

⁹ ICO. FS50630372, de 18 de julio de 2019.

blicas, así como el conocimiento público de los impactos (adversos), individuales y colectivos, de las mismas.

II. LA DISCUTIDA NATURALEZA JURÍDICA DEL CÓDIGO FUENTE Y DE LOS ALGORITMOS PÚBLICOS. SU CALIFICACIÓN COMO “INFORMACIÓN PÚBLICA”

En EEUU, la distinción entre “informática instrumental” e “informática decisional”¹⁰, ha sido uno de los argumentos relevantes utilizados por Administraciones públicas y Tribunales para la calificación del *software* y del código fuente que lo sustenta como “información pública”. La distinción es relevante porque en el caso de la informática instrumental, el código fuente del *software* instrumental empleado por la Administración no podría considerarse información pública, al no satisfacer los fines de la legislación de transparencia —el escrutinio del proceso de toma de decisiones de instituciones y responsables públicos. No ocurriría así lo mismo con la informática decisional.

En este sentido, la calificación de un programa informático “documento público” (“agency records”) a los efectos de la *Freedom Information Act* (FOIA-US) se ha centrado en determinar, a partir de un análisis casuístico, la “naturaleza y funcionalidad del software en cuestión” y, más concretamente, en qué medida el ejercicio del derecho de acceso resulta esclarecedor para el escrutinio público de la “estructura, funcionamiento y proceso de toma de decisiones de la agencia”¹¹.

¹⁰ En el caso español, el derogado Real Decreto 263/1996, de 16 de febrero, por el que se regula la utilización de técnicas electrónicas, informáticas y telemáticas por la Administración General del Estado, contemplaba esta distinción entre informática instrumental y decisional, al diferenciar claramente entre el uso de programas y aplicaciones para el ejercicio de potestades administrativas de carácter decisorio (art. 5.1) o de carácter meramente instrumental, es decir, los programas y aplicaciones estandarizados, de uso corriente, y cuya virtualidad se limita a facilitar el trabajo administrativo sin que su aplicación afecte directamente a la toma de decisiones por los órganos o las entidades competentes (art. 5.2 y Exposición de motivos).

¹¹ Department of Justice, *Guide to the Freedom of Information Act. Procedural Requirements*, 2021, pp. 11-12.

Así, por ejemplo, el Tribunal del Distrito Norte de California analizó una resolución desestimatoria de acceso al amparo de la FOIA con relación a “CLERVER”, un *software* de videoconferencia desarrollado por un contratista del Departamento de Energía (DOE) y sujeto a una licencia no exclusiva de uso. En su sentencia, el Tribunal concluyó que CLERVER no podría considerarse documento público incluso si el DOE fuese titular originario o controlase [el *software*], porque no facilita información sobre el funcionamiento, estructura o procesos de toma de decisiones de la Administración”¹². En cambio, en otro caso se concluye que el *software* era un documento público porque estaba “adaptado de forma única a su base de datos subyacente”, de manera que “el diseño del *software* y la capacidad de manipular los datos reflejan el estudio [de la agencia]”, por lo que permite la preservación de la información y “perpetúa el conocimiento” sobre las actividades de la Administración¹³.

En Italia, no existe acuerdo ni en la jurisprudencia ni en la doctrina sobre cuál es la naturaleza jurídica de los programas informáticos—incluidos los algoritmos subyacentes— utilizados por la Administración.

El Tribunal Administrativo Regional Lazio-Roma (TAR Lazio) ha calificado como “acto administrativo informático”, en el sentido previsto en el art. 22.1.d) de la Ley 241 de 1990¹⁴, el algoritmo implementado por un *software* utilizado por el Ministero dell’Istruzione, dell’Università e della Ricerca (MIUR), para el tratamiento automatizado de solicitudes de movilidad interprovincial del personal docente a través del Portal POLIS (“Presentazione On Line-

¹² *Gilmore v. US Dept. of Energy*, 4 F. Supp. 2d 922 (N.D. Cal. 1998).

¹³ *Cleary, Gottlieb, Steen & Hamilton v. HHS*, 844 F. Supp. 770, 781-82 (D.D.C. 1993).

¹⁴ LEGGE 7 agosto 1990, n. 241. Nuove norme in materia di procedimento amministrativo e di diritto di accesso ai documenti amministrativi. El precepto en cuestión define el «documento administrativo» como “cualquier representación gráfica, fotocinematográfica, electromagnética o de cualquier otra especie del contenido de actos, internos o no, relativos a un procedimiento específico, detenida por una Administración pública y concerniente a actividades de interés público, independientemente de la naturaleza publicística o privada de su disciplina sustancial”.

delle Istanze”) durante el curso 2016/2017, de acuerdo con la Orden Ministerial 241/2016¹⁵:

“– el algoritmo termina [...] por *sustentar el procedimiento en sí mismo*, ya que la identificación, en la práctica, de la plaza específica del profesor individual en el ámbito de la movilidad es identificada exclusivamente por el algoritmo antes mencionado.

– Los actos endoprocedimentales de adquisición de los datos necesarios para los fines de la correspondiente instrucción, así como el acto que pone fin al procedimiento, han convergido, en consecuencia, y *se agotan en el mero funcionamiento del algoritmo* en cuestión con la consecuencia adicional de que puede y, de hecho, debe considerarse la *asimilación del algoritmo en cuestión al acto administrativo* o mejor, [...] el reconocimiento de la atribución directa del software que gestiona el algoritmo a la categoría de los denominados *actos administrativos informáticos* a que se refiere la letra d) del art. 22 de la ley n. 241 de 1990 [cursiva nuestra].

– [...] el algoritmo es una *expresión directa de la actividad realizada por la Administración pública* que es, sin duda, actividad de interés público en cuanto interesa a la organización del servicio público representado por la educación pública y, de hecho, el algoritmo antes mencionado se incorporó al procedimiento como elemento decisivo y el mismo está detentado [...] por la misma Administración ministerial que lo [licitó] y, por tanto, lo ha utilizado para sus propios fines” [cursiva nuestra].

Aunque el algoritmo no concretiza en sí mismo el acto administrativo, para el TAR Lazio es estrictamente funcional al contenido dispositivo del acto mediante el cual se ordena la movilidad del personal afectado y, por tanto, es sustancialmente similar a este último. El algoritmo está, además, en poder de la misma Administración que lo utiliza con la finalidad de gestionar una actividad que tiene relevancia pública y, por tanto, su naturaleza y origen en el Derecho privado, al ser desarrollado por un contratista del Ministerio, no es determinante para su calificación jurídica.

Por su parte, el Consejo de Estado italiano no parece tener una posición clara sobre la cuestión. Si bien en un primer pronunciamiento coincide con la opinión del TAR Lazio al calificar al algoritmo del MIUR de “acto administrativo informático”¹⁶; en otra decisión pos-

¹⁵ T.A.R. Lazio–Roma, Sec. III Bis, Sentencia 22 de marzo 2017, n. 3769/2017.

¹⁶ Consiglio di Stato, Sec. VI, Sentencia 8 abril 2019, n. 2270, pár. 8.2: “De hecho, la norma técnica que rige cada algoritmo sigue siendo una norma administrativa

terior parece decantarse por considerar el algoritmo en cuestión como una herramienta al servicio de la acción administrativa¹⁷.

La doctrina ha tomado posición con relación a esta jurisprudencia contradictoria. Así, un sector viene calificando los programas informáticos utilizados por la Administración como “instrumentos de la acción administrativa”, en tanto que simples herramientas técnicas en las cuales los desarrolladores —normalmente, contratistas de la Administración— se limitan a ejecutar las instrucciones del poder adjudicador. Mientras que otro sector doctrinal considera los programas

general, construida por el hombre y no por la máquina, para ser luego (sólo) aplicada por ésta, aunque lo haga de forma exclusiva. Por lo tanto, esta regla algorítmica:

- posee pleno valor jurídico y administrativo, aunque se decline en forma matemática, y como tal, como hemos dicho, debe someterse a los principios generales de la actividad administrativa, como son los de publicidad y transparencia (Art. 1 l. 241/90), razonabilidad, proporcionalidad, etc.
- no puede dar lugar a aplicaciones discrecionales (de las que el ordenador está desprovisto), sino que debe prever razonablemente una solución definida para todos los casos posibles, incluso los más improbables (y esto la hace algo diferente de muchas normas administrativas generales) [...]
- ve siempre la necesidad de que la Administración desempeñe un papel *ex ante* de mediación y composición de intereses, también mediante la prueba, actualización y perfeccionamiento constantes del algoritmo (especialmente en el caso del aprendizaje progresivo y de *deep learning*).
- debe contemplar la posibilidad de que —como se ha afirmado autoritativamente— sea el juez quien «deba realizar, por primera vez a nivel ‘humano’, valoraciones y apreciaciones realizadas directamente por medios automáticos», con la consecuencia de que la decisión robotizada «exige al juez valorar la corrección del proceso automatizado en todos sus componentes».

Por lo tanto, en última instancia, el algoritmo, es decir, el programa informático, debe considerarse a todos los efectos como un «acto administrativo informatizado» [cursiva nuestra].

Nótese, sin embargo, que la argumentación desarrollada por el Consejo resulta ciertamente confusa porque inicialmente identifica al algoritmo con una “norma técnica” y, en esencia, “norma administrativa general”, para luego concluir que se trata de un “acto administrativo informático”.

¹⁷ Consiglio di Stato, Sec. VI, Sentencia 13 de diciembre 2019, n. 8472/2019, pár. 10: “[...] el uso del algoritmo debe enmarcarse correctamente en términos de *un módulo organizativo, una herramienta procedimental y de instrucción*, sujeta a las verificaciones propias de cualquier procedimiento administrativo, que sigue siendo el *modus operandi* de la decisión de autoridad, que debe llevarse a cabo sobre la base de la legislación que atribuye la potestad y los fines atribuidos por ella al organismo público, titular de la potestad [cursiva nuestra]”.

informáticos como verdaderos “actos administrativos”, en la medida en que expresarían una voluntad de la Administración condicionada a la ocurrencia del supuesto de hecho previamente identificado y definido en el programa en cuestión¹⁸.

Orofino considera que, al estar constituido por una serie de signos y expresiones que no se manifiestan según las reglas normales de la comunicación lingüística, el programa informático no puede considerarse un acto administrativo, sino que, en el mejor de los casos, puede constituir el objeto de una voluntad administrativa: la voluntad de hacer propia la decisión tomada por la máquina. Además, si los actos administrativos son declaraciones de voluntad dictadas por una autoridad, la declaración, para ser tal, debe ser comunicada de manera que permita a los destinatarios comprender su significado. Por lo tanto, el software no puede considerarse un acto administrativo, sino más bien un “instrumento de la acción administrativa”. Ahora bien, con independencia de la naturaleza jurídica del *software*, éste debe considerarse como un documento público accesible en el sentido establecido por el art. 22.1.d) de la Ley núm. 241 de 1990¹⁹.

Para Cavallaro y Smorto, el planteamiento realizado por el TAR Lazio sobre el rol (técnico) del algoritmo en la asignación de la plaza específica, por su propia caracterización, evoca más bien a la propia definición del procedimiento administrativo. Ahora bien, lo relevante de esta sentencia es que pone de relieve el verdadero alcance innovador de la decisión automatizada: “[...] puede ocurrir que el algoritmo, entendido como regla técnica, asuma un papel que va más allá del mero supuesto en el que se basa la decisión, y que incluso llegue a constituir un sistema de formación de la propia voluntad procedimental”²⁰.

¹⁸ Por todos, y resumiendo la doctrina italiana sobre la cuestión, véase G. Manco-su, “Les algorithmes publics déterministes au prisme du cas italien de la mobilité des enseignants”, *Rivista Italiana di Informatica e Diritto*, Fasciolo 1-2019, p. 77.

¹⁹ A.G., Orofino, “The implementation of the Transparency Principle in the Development of Electronic Administration”, *European Review of Digital Administrative & Law*, Volume 1, Issue 1-2, June-December, 2020, pp. 125-126.

²⁰ M. C. Cavallaro y G. Smorto, “Decisione pubblica e responsabilità dell’amministrazione nella società dell’algoritmo”, *Federalismi. Rivista di Diritto Pubblico Italiano, Comparato, Europeo*, núm. 16, 2019, p. 16.

Frente al planteamiento italiano, la Comisión francesa de Acceso a los Documentos Administrativos (CADA) ha abordado de manera diferente la cuestión. En un primer momento, la Comisión de Acceso reconoce el carácter de documentos administrativos del código fuente de los algoritmos utilizados por la Administración. Doctrina que acaba siendo codificada por el legislador francés. Y, más recientemente, se ha pronunciado sobre la naturaleza jurídica de los algoritmos públicos.

En efecto, la llamada *Ley Lemaire* de 2016²¹ vino a codificar lo que ya ha habido reconocido la doctrina del CADA, previamente, a saber, el carácter de “documento administrativo” del código fuente y de los algoritmos implementados en los programas informáticos utilizados por la Administración²². Así, el art. L300-2 Código de Relaciones entre el Público y la Administración (CRPA) califica el código fuente utilizado por una Administración como “documento administrativo”. Y, a su vez, el art. L311-3-1 del mismo texto legal dispone que: “[...] las decisiones individuales adoptadas sobre la base de un tratamiento algorítmico incluirán una mención explícita al informar a la parte interesada. Las reglas que definen este tratamiento y las características principales de su implementación serán comunicadas por la Administración a la parte interesada si así lo solicita.”

²¹ LOI n° 2016-1321 du 7 octobre 2016 pour une République Numérique. JORF n° 0235 du 8 octobre 2016.

²² CADA. Opinión núm. 20142953, de 16 de octubre de 2014, donde se estimó el acceso a un programa informático desarrollado por una empresa privada, relacionado con un contrato público para la construcción del Musée des Confluences de Lyon, con eliminación de aquellas partes que pudieran verse afectadas por el secreto en materia industrial o comercial; Opinión núm. 20144578, de 8 de enero de 2015, se estimó el derecho de acceso al código fuente de un *software* elaborado por la Dirección General de Finanzas que simulaba el cálculo del impuesto sobre la renta, para poder reutilizarlo en un trabajo de investigación universitario, pues se consideró que los archivos informáticos que constituían el código fuente solicitado, producido por la Administración como parte de su misión de servicio público, tenían el carácter de documentos administrativo; Opinión 20161990, del 23 de junio de 2016, se consideró que el algoritmo desarrollado por el Ministerio de Educación francés, conocido como «APB» (*Admission Post-Bac*) para la tramitación de las solicitudes de admisión a los grados universitarios tenía también la consideración documento administrativo en el sentido del CRPA, y por tanto, era accesible.

Este último precepto ha sido objeto de desarrollo reglamentario por el Consejo de Estado francés²³. Concretamente, el art. R311-3-1-1 establece que la decisión administrativa individual deberá contener una “mención explícita” de la “finalidad perseguida por el tratamiento algorítmico”. Ello implica el “derecho a obtener la comunicación de las reglas que definen el tratamiento y las principales características de su aplicación, así como las modalidades de ejercicio de este derecho a la comunicación y de revisión, si corresponde, ante la Comisión de Acceso a los Documentos Administrativos”. A su vez, el art. R311-3-1-2 concreta la información específica que debe facilitarse al interesado que haya sido objeto de una decisión individual fundamentada en un tratamiento algorítmico en caso de que el interesado así lo reclame. Esta información deberá facilitarse de una “forma inteligible” y deberá incluir: “1º El grado y el modo de contribución del tratamiento algorítmico a la toma de decisión; 2º Los datos tratados y sus fuentes; 3º Los parámetros de tratamiento, y si procede, su ponderación, aplicados a la situación de interesado; 4º Las operaciones efectuadas por el tratamiento.”²⁴

²³ *Decreto N° 2017-330, de 14 de marzo de 2017, relativo a los derechos de las personas que sean objeto de decisiones individuales adoptadas sobre el fundamento de un tratamiento algorítmico.*

²⁴ Del marco jurídico previsto por el legislador francés deben destacarse dos notas importantes. En primer lugar, a la hora de reconocer este derecho del interesado, el legislador francés no ha discriminado entre tipos de algoritmos, en la medida en que los algoritmos que no incorporan técnicas de IA también pueden tener impactos adversos sociales e individuales relevantes, como evidencian los algoritmos del MIUR y de Ofqual, que se comentarán más adelante. En segundo lugar, el derecho de acceso a la información relativa a los tratamientos algorítmicos que sirvan de fundamento de las decisiones administrativas individuales puede ser ejercido tanto por las personas físicas como por las personas jurídicas, más allá del planteamiento de Reglamento General de Protección de Datos con relación a las limitaciones que plantea el derecho a una explicación sobre la lógica algorítmica, dado que este derecho sólo se aplicaría a las personas físicas. De hecho, la *Memoria de Impacto Normativo del Proyecto de Ley para una República Digital* enfatiza que este derecho acceso a la información relativa a los tratamientos algorítmicos completa y refuerza el marco jurídico aplicable a las personas físicas en materia de protección de datos. Cfr. LEGIFRANCE (2015). *Projet de Loi pour une République numérique. Etude d'Impact*. 9 de diciembre, pp. 10-12.

En coherencia con todo lo anterior, la CADA ha venido calificando como “documentos administrativos”, no sólo el código fuente²⁵ o los algoritmos implementados por una Administración²⁶, sino también la documentación técnica relativa al código fuente, como puede ser el documento de especificación de requisitos de *software*²⁷. Así, por ejemplo, con relación al código fuente de la plataforma Parcoursup, la Autoridad francesa ha estimado el derecho de acceso a las especificaciones del *software*, “presentadas sintéticamente”²⁸. Asimismo, el código fuente del programa utilizado por el Fondo Nacional de Subsidio Familiar (CNAF) para el cálculo completo de ayudas económicas de carácter familiar o social ha sido calificado también por la CDA como “documento administrativo”, junto con los archivos SQL de dicho código fuente y las especificaciones funcionales utilizadas para calcular las ayudas a la vivienda, la renta solidaria activa, las asignaciones familiares y la prima de actividad²⁹.

Finalmente, en una reciente resolución, la CADA estima el derecho de acceso al código fuente completo de la plataforma informática Parcoursup, habilitada para la gestión automatizada del procedimiento nacional de preinscripción en el primer año de enseñanza universitaria pública. A propósito de la misma, subraya: “Un algoritmo público es un *procedimiento administrativo, total o parcialmente infor-*

²⁵ Con relación a resoluciones estimatorias del derecho de acceso al código fuente, véanse, CADA. Opiniones 20191797, de 16 de enero de 2020; 20182093, de 6 de septiembre de 2018; 20182120, de 6 de septiembre de 2018; 20182455, de 6 de septiembre de 2018; 20182682, de 6 de septiembre de 2018; 20180276, de 19 de abril de 2018; 20161990, de 23 de junio de 2016; 20161989, de 23 de junio de 2016. En sentido desestimatorio, véanse también las Opiniones núms. 20200496, de 12 de marzo de 2020; 20181891, de 18 de julio de 2019; 20184400, de 10 de enero de 2019; 20180376, 31 de mayo de 2018.

²⁶ En sentido estimatorio, CADA. Opiniones núms. 20182093, de 6 de septiembre de 2018; 20182120, de 6 de septiembre de 2018; 20182455, de 6 de septiembre de 2018; 20173235, de 30 de noviembre de 2017; 20163835, de 6 de octubre de 2016; 20161990, de 23 de junio de 2016; 20161989, de 23 de junio de 2016; en sentido desestimatorio, 20201743, de 10 de septiembre de 2020; 20184400, de 10 de enero de 2019.

²⁷ CADA. Opiniones núms. 20184400, de 10 de enero de 2019; 20182093, de 06 de septiembre de 2018; 20182120 y 20182455, de 6 de septiembre de 2018, respectivamente.

²⁸ CADA. Opinión núm. 20182093, de 06 de septiembre de 2018.

²⁹ CADA. Opinión núm. 20181891, de 18 de julio de 2019.

matizado, que interviene en un proceso de toma de decisiones para los ciudadanos. El código fuente es la traducción informática de ese algoritmo. Explica el método de toma de decisiones administrativas, permite controlar la interpretación y aplicación de la regla de derecho implementada por las autoridades públicas [*cursiva nuestra*]

Por su parte, la Autoridad británica también ha resuelto a favor de divulgar las especificaciones técnicas documentadas sobre pantallas de usuario y “procesos batch o por lotes”, así como los metadatos contenidos en un diccionario de datos que describía las características de los datos y las relaciones entre éstos de un software utilizado por la Students Loans Company (SLC)³⁰ para monitorizar los datos de recobro de los préstamos, pues el acceso a dicho documento permitiría al solicitante “entender exactamente qué consultas se automatizan, cuál es el sistema para obtener el volcado de la base de datos, etc.”³¹.

Ahora bien, el ICO ha venido siendo reticente a la hora de estimar el acceso al código fuente de las aplicaciones utilizadas por la Administración al considerar prevalente la salvaguarda de los intereses comerciales y los derechos de propiedad intelectual de la propia Administración o de terceros, y ello a pesar de reconocer el interés público que plantea el uso de ciertas aplicaciones con impactos sociales e individuales³².

³⁰ La «Student Loans Company» es una empresa pública cuya función es conceder préstamos y becas a los estudiantes universitarios y de formación profesional y cobrar los reembolsos de dichos préstamos. Es un organismo participado por el Departamento de Educación del Gobierno del Reino Unido, el Gobierno de Escocia, el Gobierno de Gales, el Departamento de Economía de Irlanda del Norte, la Autoridad Tributaria y Aduanera y el Servicio de Admisión Universitario y de Formación Profesional. Vid. <https://www.gov.uk/government/organisations/student-loans-company/about>.

³¹ ICO. FS50323800, de 9 de diciembre de 2010.

³² ICO. FS50371026, de 9 de enero 2012, párr. 27, 32-33; FS50459127, de 4 de marzo 2013, párr. 27, 30-31. En ambos casos, se deniega el acceso a un *software* utilizado por el Departamento de Trabajo y Pensiones, LIME, para valorar la incapacidad laboral de los trabajadores, a pesar de reconocer las dudas existentes en la opinión pública sobre el funcionamiento de este programa, que “tendrá un impacto en la vida de muchas personas”. “La cuestión de las evaluaciones médicas —afirma el ICO— ha suscitado una gran atención por parte de los medios de comunicación, y el hecho de que una mejor comprensión de la forma en que

III. DETECCIÓN Y ESCRUTINIO PÚBLICO DE PATOLOGÍAS DE LA AUTOMATIZACIÓN Y ALGORITMIZACIÓN DE LA ACTIVIDAD ADMINISTRATIVA. EL INTERÉS PÚBLICO EN EL ACCESO

Desde una perspectiva iusfundamental, resulta ya un lugar común referirse a los riesgos que, para los derechos y libertades de los ciudadanos, comportan la creciente automatización y algoritmización de la actividad de las organizaciones, incluidas las Administraciones públicas: ampliación de la discriminación y de los sesgos sociales, la pérdida de privacidad, los impactos individuales y sociales asociados al reconocimiento facial y la evaluación del riesgo criminal, o la desinformación entre otros³³.

La cuestión resulta trascendental, porque en palabras de Fernández Salmerón, la llamada “opacidad algorítmica”, resultante de las complejidades técnicas que se derivan de la automatización y de la implementación de algoritmos de IA en la actividad administrativa, plantea un serio problema para el control judicial y un riesgo de “abandono de los principios esenciales que rigen en la Administración Pública”³⁴. No en balde, si en algo ha insistido el Consejo de Estado Italiano es que la utilización de un procedimiento administrativo automatizado “no puede ser motivo para eludir los principios que conforman nuestro ordenamiento jurídico y que rigen el desarrollo de la actividad administrativa”³⁵.

En este sentido, un sector de la doctrina considera que el derecho de acceso al código fuente y al algoritmo subyacente pueden contri-

se toman las decisiones conducirá a un debate mejor informado y a una mayor confianza en el proceso”.

³³ M. Brundage, S. Avin, J. Wang *et al.* *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*, version 2, 20 de abril 2020. <https://arxiv.org/abs/2004.07213>.

³⁴ M. Fernández Salmerón, “The Risk of a Government as a Big Brother”, J. Vida Fernández, dir., *Conference on Digital transformation of Government: towards a digital Leviathan?*, Universidad Carlos III e Indiana Law Review, 23-24 de junio 2022.

³⁵ Consiglio di Stato, Sec. VI, Sentencias 8 abril 2019, n. 2270, pár. 8.2; y 13 de diciembre 2019, n. 8472/2019, pár. 10.

buir a fomentar la transparencia algorítmica³⁶, en la medida en que ello implicaría “un cierto grado de escrutinio público” de estos sistemas que automatizan la actividad administrativa³⁷.

En cambio, otro sector considera insuficiente el derecho de acceso³⁸, en la medida en que el conocimiento del código fuente no garantizaría en sí mismo la transparencia por la incapacidad del ciudadano —no experto— en comprender el lenguaje de las máquinas, y mayormente, si se trata de sistemas basados en IA³⁹. Es más, se entiende que la legislación de transparencia y acceso puede ser un instrumento excesivamente restrictivo para garantizar la efectiva transparencia de los sistemas automatizados⁴⁰.

Sin embargo, algunos de los casos analizados en materia de derecho de acceso, evidencian que, en mayor o menor medida, el acceso al código, al modelo algorítmico implementado, así como la documentación e información asociada (especificaciones técnicas, descripción y caracterización del *dataset*, métricas de validación, o auditorías de terceros) pueden contribuir a la comprensión y escrutinio público del funcionamiento y de los efectos individuales o colectivos de la automatización y algoritmización de la actividad administrativa.

En particular, el acceso a la información técnica ha sido clave para la mejor comprensión de algunas de las patologías que aquejan a estos sistemas y ha abierto la puerta a la revisión judicial de las decisiones administrativas automatizadas, como ha sido el caso del algoritmo de movilidad docente del MIUR. Los resultados del algoritmo hablan por sí solos: miles de profesores trasladados a cientos de

³⁶ N. Diakopoulos, “Algorithmic accountability reporting: on the investigation of black boxes”, *Columbia Journalism School*, 2014.

³⁷ J. New y D. Castro. How Policymakers Can Foster Algorithmic Accountability. Center for Data Innovation, 2018, p. 8. <https://x.gd/WfPDw>.

³⁸ A. Boix Palop, “Los algoritmos son reglamentados...”, cit., p. 242.

³⁹ J. A. Kroll, J. Huey et al. Accountable Algorithms, *University of Pennsylvania Law Review*, vol. 165, núm. 633, 2017, p. 638; Orofino, “The implementation of the transparency principle”, cit., p. 130, quien insiste en que, a pesar de la incomprensibilidad técnica para el ciudadano común, el acceso al código debe ser garantizado en todo caso.

⁴⁰ J. Valero Torrijos, “Las garantías jurídicas de la inteligencia artificial en la actividad administrativa desde la perspectiva de la buena administración”, *Revista Catalana de Dret Públic*, vol. 58, 2019, p. 89.

kilómetros de sus hogares, frente a otros con menor puntuación que fueron asignados a un puesto en la misma provincia en la que vivían. Profesores de Apulia y Calabria tuvieron que trasladarse a la provincia de Milán, cuando deberían haber sido asignados a sus regiones. Como resultado, hubo numerosas demandas y recursos⁴¹.

Como señaló el Consejo de Estado italiano al resolver los recursos contra las asignaciones de plazas a los docentes, realizadas por el algoritmo⁴²:

“[...] la norma algorítmica no sólo debe ser cognoscible en sí misma, sino también estar sujeta al pleno conocimiento, y a la plena revisión, del juez administrativo. En efecto, la citada exigencia responde a la necesidad irrenunciable de poder comprobar cómo se ha ejercido concretamente la potestad, planteándose en última instancia como una *declinación directa del derecho de defensa del ciudadano, al que no se le puede impedir conocer las modalidades (aunque sean automatizadas) con las que se ha adoptado concretamente una decisión destinada a afectar a su esfera jurídica*.

Sólo así es posible llevar a cabo, incluso en los tribunales, una valoración completa de la legitimidad de la decisión [...]

En este sentido, la decisión administrativa automatizada requiere que el juez evalúe primero la *corrección del proceso informático en todos sus componentes: desde su construcción, pasando por la introducción de datos, hasta su validez, hasta su gestión*. De ahí, como se ha mencionado anteriormente, *la necesidad de garantizar que ese proceso, a nivel administrativo, se desarrolle de forma transparente, mediante la posibilidad de conocer los datos introducidos y el propio algoritmo*.

En segundo lugar, consecuente con lo primero, el juez debe poder es-cudriñar la propia lógica y razonabilidad de la decisión administrativa robotizada, es decir, de la “*regla*” que rige el algoritmo”.

De hecho, la imposibilidad de comprender cómo adopta las decisiones el algoritmo, máxime cuando tales decisiones resultan arbitrarias y con efectos lesivos en el ciudadano, deben llevar a la sanción de nulidad de tales decisiones⁴³.

⁴¹ B. Aragona. *Algorithm Audit: Why, What, and How?*, Rotledge Focus, Abingdon, New York, 2022, p. xi.

⁴² Consiglio di Stato, Sec. VI, Sentencia de 8 abril 2019, n. 2270, pár. 8.

⁴³ *Idem*, pár. 9. “[...] procede estimar el recurso, ya que en este caso se han vulnerado los principios de imparcialidad, publicidad y transparencia, al no poder entenderse la defraudación de las legítimas expectativas de las personas situadas

IV. EL PROBLEMA DE LOS ERRORES EN EL DISEÑO DE LOS ALGORITMOS DEL MIUR Y DE OFQUAL

Entre las críticas más habituales a la automatización o al uso de algoritmos predictivos —sean o no de IA— en el ámbito de la actividad administrativa, se considera que estos sistemas pueden generar efectos *extra legem y contra legem*, como consecuencia de una traducción incorrecta —querida o no— del contenido de la norma jurídica en lenguaje de código.

Sin perjuicio de su eficacia, la formalización y la transposición de las normas jurídicas a un lenguaje informático puede afectar no sólo a la naturaleza, sino también al alcance y la significación de estas normas. La realidad es que el desarrollador del modelo, normalmente, el experto informático —*in house* o contratista de la Administración—, al traducir la norma en código, puede incurrir en errores de programación o incorporar inevitablemente su propia interpretación subjetiva de la norma, diseñando, así, voluntaria o involuntariamente, un modelo que modifica el contenido y alcance de la misma⁴⁴.

El algoritmo del MIUR es un claro ejemplo de esto que se dice. Al reconocer el derecho de acceso del sindicato reclamante al código fuente, y al algoritmo subyacente, la Sentencia del TAR Lazio-Roma nos da la clave de la importancia de la cuestión. Sea comprensible o no el lenguaje de programación para el ciudadano medio o para el mismo juez, el derecho de acceso permite la verificación de la correc-

en una determinada posición de la lista de clasificación. En efecto, la imposibilidad de comprender la forma en que, mediante el citado algoritmo, se asignaron los puestos disponibles constituye en sí misma un vicio capaz de viciar el procedimiento. No sólo eso, sino que los resultados del procedimiento sí parecen estar caracterizados por la ilógica e irracionalidad que denuncian los recurrentes, ya que se han producido situaciones paradójicas en las que a profesores con muchos años de servicio se les han asignado áreas territoriales que nunca habían solicitado y que están situadas a cientos de kilómetros de su ciudad de residencia, mientras que otros profesores, con menos titulación y antigüedad, han obtenido las mismas plazas que solicitaron”.

⁴⁴ D. Citron, “Technological Due Process”, *Wash. UL Rev.*, núm. 85, pp. 1254-1255; D. Bourcier y P. De Filippi, “Les algorithmes sont-ils devenus le langage ordinaire de l’administration?”, G. Koubi, L. Cluzel-Metayer y W. Tamzini, *Lectures critiques du Code des relations Public et Administration*, LGDJ, hal-01850928f, 2018, pp. 200 y 207.

ción de la decisión administrativa automatizada si no directamente por el ciudadano medio, sí a través de una opinión técnica experta:

“Que el *software* se compile utilizando lenguajes de programación habitualmente incomprensibles no solo para el funcionario que lo utiliza con el fin de redactar la decisión final del procedimiento administrativo sino también para el destinatario particular del propio acto administrativo no es un argumento determinante para negar su naturaleza de información pública, dado que, por un lado, la citada circunstancia es consecuencia de la elección discrecional de la Administración de utilizar una herramienta innovadora, como es la programación informática, para la gestión de un procedimiento de su propia responsabilidad y competencia; y, por otro, *el destinatario particular del acto administrativo, en particular, puede, en todo caso, hacer uso legítimo de la actividad profesional de un informático competente con el fin de comprenderlo y comprobar su corrección*” [cursiva nuestra].

Precisamente, una de las consecuencias inmediatas de la Sentencia estimatoria del TAR Lazio-Roma en este asunto es que, al reconocer al sindicato reclamante el derecho de acceso al código fuente, posibilitó en términos materiales el análisis pericial del código reclamado y cuestionado. El informe técnico pericial realizado posteriormente al código por expertos informáticos de las Universidades de Tor Vergata y La Sapienza de Roma evidencia la trascendencia de esta cuestión y cómo el derecho de acceso garantiza el fin último de las leyes de transparencia, a saber, hacer efectivo el derecho de los ciudadanos a “conocer cómo se toman las decisiones que les afectan, cómo se manejan los fondos públicos o bajo qué criterios actúan nuestras instituciones”⁴⁵, contribuye a la debida rendición de cuentas de las instituciones públicas, de sus representantes y gestores⁴⁶.

En efecto, el informe técnico pericial posterior reveló que el algoritmo en cuestión había sido desarrollado por programadores externos a la Administración distintos utilizando dos lenguajes de programación diferentes. Para la fase A del procedimiento de adjudicación de plazas, se había desarrollado en “lenguaje COBOL”, un lenguaje

⁴⁵ Vid. Exposición de motivos de la Ley 19/2013, de 9 de diciembre, de Transparencia, Acceso a la Información Pública y Buen Gobierno.

⁴⁶ Sentencia del Tribunal de Primera Instancia (Sala Cuarta), Kuijer/Consejo (T-211/00), de 7 de febrero de 2002, párrafo 52.

de programación obsoleto y reemplazado por otros con mejor rendimiento en términos de sintaxis lógico-aritmética. Para las fases B, C y D, se había empleado “lenguaje C”. Esta práctica se consideró no recomendable, teniendo en cuenta que las fases en cuestión debían interactuar y compartir datos y resultados entre sí. El Informe subrayaba, además, que el precio pagado a los contratistas por el desarrollo del código, unos 440 euros —muy por encima de los precios de mercado⁴⁷, ni siquiera se justificaba por necesidades de racionalización y eficiencia del gasto público.

Por si lo anterior no fuera poco, la pericial también determinó la existencia de omisiones relevantes en la información aportada por el Ministerio de Educación que impedían el adecuado escrutinio del *software* cuestionado. En particular, se observó que en las fases B, C y D del algoritmo las funciones incluidas hacían uso de una base de datos a través de la cual, presumiblemente, se gestionaba la información de entrada y salida del programa, así como los posibles resultados y valores temporales. Sin embargo, la documentación relativa a la estructura y formato de la base de datos no fue facilitada por la Administración a los peritos, con lo cual no fue posible una correcta y completa verificación del programa en cuestión. En este sentido, el informe pericial consideró que las posibles anomalías del algoritmo desarrollado podían derivar no solo de errores en el diseño del código fuente, sino también de una incorrecta preparación y gestión de los datos de entrada que eran objeto de procesamiento, lo que irremediablemente afectaba a los resultados obtenidos. Por último, se había restringido, mediante el bloqueo de determinados archivos, la posibilidad de realizar una verificación directa y automática, a través de software específico para ello y de compiladores que agilizarían tal verificación de la corrección lógica y sintáctica del programa. Se consideró que este bloqueo se había introducido con el fin de restringir la verificación de las funciones que realizaba el algoritmo, de manera que ésta sólo podría realizarse mediante el “copiado manual” de las

⁴⁷ Gilda degli Insegnanti Oristano, *Perizia tecnica sul contestato algoritmo. Interviste video de La Tecnica della Scuola. Tecnica della Scuola intervista il coordinatore nazionale, Rino Di Meglio, l'ing. Alessandro Salvucci e l'avv. Michele Bonetti*. 15 de junio de 2017. <https://x.gd/4RlMk>.

líneas de código. Y ya tan solo para la fase A, ascendía a unas 29600 líneas. Las conclusiones del Informe no podían ser más demoledoras:

“Es evidente que la falta de tales aclaraciones, así como la falta de las líneas reclamadas del código, la base de datos, las líneas que usa el *software* para leer y escribir los datos (no tanto en el contenido como en la forma) así como las especificaciones técnicas, configuran una *conducta poco transparente*, a pesar de la orden de exhibición de datos y documentos por parte del TAR al Ministerio.

Estas omisiones invalidan irreversiblemente la posibilidad de una supervisión completa sobre el uso concreto del algoritmo y, por tanto, cómo se ha determinado la movilidad de los docentes por todo el país.”⁴⁸

En este contexto, no está de más recordar que la jurisprudencia del Tribunal Europeo de Derechos Humanos ha advertido que la “obstinada renuencia” o la actitud “dilatoria” de los órganos administrativos a la hora de dar acceso a la información pública, en incumplimiento de resoluciones de las Autoridades de control o de los Tribunales que han estimado el derecho de acceso del solicitante, debe ser considerada como una “restricción arbitraria” contraria al principio de legalidad y una injerencia arbitraria del art.10 del Convenio Europeo de Derechos Humanos⁴⁹.

El algoritmo de Ofqual evidencia nuevamente cómo un modelo algorítmico diseñado defectuosamente puede tener impactos sociales adversos, una vez más en el ámbito educativo. En este caso, no se discute sobre el derecho de acceso al código fuente o al algoritmo utilizado por la Administración, pues la documentación técnica y la explicación del modelo algorítmico fueron publicados. Pero sí que se pone de manifiesto el interés público en el acceso a determinados resultados desagregados del algoritmo que no habían sido hechos públicos previamente, en la medida en que contribuyan a posibilitar o completar la interpretabilidad de los resultados de los modelos predictivos y, por tanto, un escrutinio amplio público de las decisiones públicas basadas o adoptadas en estos modelos.

⁴⁸ Cfr. A. Salvucci, M. Giogi, E. Barchiesi y M. Scafidi, *Perizia tecnica preliminare sull'analisi dell'algoritmo che gestisce Il software della mobilità docenti per l'a.s. 2016/2017*. 4 de junio de 2017, pp. 12-17.

⁴⁹ Kenedi v. Hungria (2009), §45; Youth Initiative for Human Rights v Serbia (2013), §24-26.

Con motivo de la pandemia de la COVID-19, en el Reino Unido, los exámenes oficiales correspondientes al nivel “A-Level”, que dan acceso a los estudios universitarios, fueron suspendidos y sustituidos por un sistema basado en un algoritmo desarrollado por Ofqual, el organismo regulador de los exámenes oficiales y calificaciones, que predecía la calificación que habrían obtenido los alumnos si éstos se hubiesen realizado presencialmente.

Tras la publicación de las calificaciones, se produjeron numerosas quejas por todo el país sobre el proceso y los resultados. Casi el 40% de las 700.000 calificaciones remitidas por los centros (CAG) habían sido revisadas a la baja en, al menos, un nivel; y, en un 3,5% dicha reducción se concretó dos o más niveles⁵⁰. El propio Comisionado de acceso a la información se hizo eco de las desviaciones del algoritmo respecto de las estimaciones realizadas por los centros, así como de las fuertes críticas de la opinión pública y de los principales medios de comunicación al procedimiento elegido y a los resultados⁵¹.

Una de las críticas más recurrentes fue que las calificaciones a la baja realizadas por el algoritmo habían afectado especialmente a los colegios públicos, normalmente, con mayor número de estudiantes y peor financiados que los colegios privados⁵². En este sentido, los centros privados vieron cómo la proporción de las mejores calificaciones de nivel A imputadas aumentaron en más del doble que la de los colegios públicos. En el caso de los centros independientes, con un número de estudiantes reducido, los resultados de las calificaciones en el nivel A o A* crecieron 4,7 puntos porcentuales, pasando del 43,9% en 2019 al 48,6% por ciento en 2020, mientras que en los colegios públicos la cifra varió sólo en dos puntos, del 19,8% en 2019 al 21,8 por ciento en el 2020⁵³.

⁵⁰ R. Adams, S. Weale y C. Barr. “A-level results: almost 40% of teacher assessments in England downgraded”. *The Guardian*, 13 de agosto 2020. <https://x.gd/amSfY>.

⁵¹ ICO. IC-70514-H7K5, de 5 de agosto 2021, par. 54.

⁵² J. Faraway. *An Alternative to the OFQUAL algorithm*. 28 de agosto 2020. <https://x.gd/DsDd0>. BBC. *A-levels and GCSEs: How did the exam algorithm work?*, 20 de agosto 2020. <https://x.gd/GIsID>.

⁵³ D. Hughes, “What is the A-level algorithm? How the Ofqual’s grade calculation worked – and its effect on 2020 results explained”. *inews*. 17 de agosto 2020. <https://x.gd/anMGq>.

Pero ¿cómo funcionaba realmente el controvertido algoritmo? Para determinar las calificaciones (predicciones), el organismo regulador competente, Ofqual, diseñó un algoritmo cuya finalidad no era predecir directamente las calificaciones A-level individuales de los alumnos, sino el porcentaje de alumnos que en un centro concreto “j” deberían recibir una determinada calificación “k” dentro de los rangos posibles A*, A, B, C, D, E y U. El algoritmo en cuestión era más bien de tipo heurístico⁵⁴ y, a pesar de algunas opiniones, no incorporaba en realidad estrategias de IA⁵⁵.

Para llevar a cabo este proceso, Ofqual pidió a los centros que remitieran, para cada estudiante y para cada asignatura en la que estuvieran matriculados, dos tipos de información: por un lado, la calificación estimada por los profesores, denominada “Centre Assessment Grade” (CAG); y, por otro, un ranking de cada estudiante en cada rango de calificación, ordenado de mejor a peor, comparado con el resto de los estudiantes del mismo centro con el mismo CAG. Lo que ocurrió, sin embargo, fue que Ofqual sólo aplicó el CAG en casos concretos⁵⁶. En cambio, en otros centros, al menos, teóricamente, las predicciones de las calificaciones de los estudiantes fueron una combinación de los CAG y del algoritmo.

En particular, el CAG fue la única o principal información de referencia para la evaluación de los candidatos privados y en los centros con un número pequeño de matriculaciones en las asignaturas en el curso 2020 y en la serie histórica considerada, ya que a menor número de estudiantes la evidencia estadística se debilita. En cambio, en el

⁵⁴ S. Bennett, *On A Levels, Ofqual and Algorithms*, 20 de agosto 2020. <https://x.gd/G9AmA>.

⁵⁵ T. Paulden, “A cutting re-mark”. *Significance*, Vol. 17, núm. 5, 2020, pp. 4-5. <https://doi.org/10.1111/1740-9713.01436>. Algunos autores, sin embargo, han identificado el algoritmo de Ofqual con técnicas de aprendizaje automático. Véanse, por ejemplo, Y. Hetch. “UK’s Failed Attempt to Grade Students by an Algorithm. Why engineering alone isn’t enough to fix broken social systems”. *Towards AI*. <https://x.gd/g4PbX>; S. Akgun, C. Greenhow. “Artificial intelligence in education: Addressing ethical challenges in K-12 settings”, *AI Ethics*, Springer, 2021, p. 6. <https://doi.org/10.1007/s43681-021-00096-7>.

⁵⁶ Ofqual. *Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report*, pp. 97-102. <https://x.gd/k4h99>

caso de los centros con mayor número de estudiantes matriculados, el proceso de estandarización aplicado por el algoritmo se basaba en un enfoque que otorgaba “un mayor peso a la evidencia estadística histórica del rendimiento de los centros (teniendo en cuenta el rendimiento previo de los estudiantes) que a los CAG remitidos”⁵⁷. Para la realización de sus predicciones, el algoritmo tenía en cuenta las siguientes fuentes de información.

La distribución histórica de calificaciones en los centros educativos para cada asignatura en los tres últimos cursos académicos precedentes (2017-2019). Esta distribución histórica se sometió a un doble proceso de ajuste. Por un lado, se añadió una corrección por las diferencias entre los resultados académicos de la cohorte en el año 2020 y los resultados previos de las cohortes de 2017-2019; y, por otro, se tuvo en cuenta la proporción de estudiantes que, en un determinado centro, podría ser emparejada con la cohorte de estudiantes históricos en función de sus resultados académicos anteriores.

Los rankings de estudiantes elaborados por cada centro fueron moderados a nivel nacional mediante la imputación de una calificación “proxy”. Para ello, la distribución de calificaciones generada algorítmicamente se sometió a un proceso de normalización transformando las calificaciones ordinales (A*, A, B, etc.) en una puntuación pseudo-numérica asignada a partir del orden establecido por el ranking⁵⁸. El resultado fueron unas puntuaciones de corte normalizadas (“notional cut-scores”) que delimitaban los límites de las calificaciones a nivel nacional. Estas puntuaciones de corte buscaban que no se produjera una inflación de las calificaciones de un año para otro. Las calificaciones finales de cada centro se imputaron distribuyendo a los estudiantes de manera uniforme en el intervalo de calificaciones disponibles para cada calificación pseudo-numérica.

Según los expertos, en la práctica, el CAG fue sistemáticamente ignorado para los centros grandes, lo cual introdujo las distorsiones reflejadas en los discutidos resultados⁵⁹. Así, por ejemplo, si la estima-

⁵⁷ *Idem*, pp. 11, 92-93, 95-97. <https://x.gd/k4h99>.

⁵⁸ Por ejemplo, si hubiera 3 alumnos clasificados como 1, 2 y 3 con una calificación de A en un determinado centro, las puntuaciones podrían ser las siguientes: alumno 1: 600 (‘alto’ A); alumno 2: 550 (‘media’ A); alumno 3: 500 (‘bajo’ A).

⁵⁹ S. Bennett. *On A Levels...*, cit.

ción del CAG para un alumno de un centro concreto era un nivel A* en matemáticas, el algoritmo habría reducido dicha estimación a un nivel A, e incluso un nivel B o C, en función de la cohorte histórica del centro y de la puntuación pseudo-numérica de corte imputada a partir del ranking. Tras varios días de creciente presión por parte de la opinión pública y de los medios de comunicación, Ofqual anuló las “calificaciones sintéticas” y las sustituyó por las calificaciones originalmente asignadas por los profesores. El “algoritmo mutante”—como fue calificado por Boris Johnson— fue retirado no tanto por razones estadísticas, sino políticas⁶⁰.

El contexto precedente descrito sirve para entender mejor el sentido de la reclamación que tuvo que resolver el ICO frente a una resolución desestimatoria de acceso que tenía por objeto precisamente los resultados desagregados del algoritmo, y en particular, el nombre de cada centro, el % de calificaciones que habían sido mejoradas en 2 niveles para cada centro respecto de su CAG; el % mejorado en 1 nivel; el % con el mismo nivel; % reducido en -1 nivel; % reducido en -2 niveles; % reducido en -3 niveles⁶¹.

Ofqual había considerado aplicable la sección 36(2)(c) de la FOIA para justificar la desestimación de la solicitud en cuestión, alegando que la divulgación de dicha información “perjudicaría o probablemente podría perjudicar la dirección efectiva de los asuntos públicos”. El organismo regulador argumentó que la decisión de no publicar los datos educativos a nivel de centro para 2020 se adoptó para garantizar que el proceso de calificación fuera lo más sólido posible y que los profesores elaborasen los CAG sin temor a ser juzgados. En particular, consideraba que los centros tenían una expectativa legítima de que el rendimiento a nivel de centro no se haría público. En este sentido, la relación de Ofqual con las partes interesadas, y particularmente con los profesores, los centros y sus representantes, se vería perjudicada por la divulgación de la información en cuestión, lo que afectaría a la capacidad del organismo regulador para desempeñar sus funciones. Se entendía también que la divulgación también

⁶⁰ T. Harkness, “How Ofqual Failed the Algorithm Test”, *UNHERD*, 18 de agosto de 2020. <https://x.gd/Cy3pe>.

⁶¹ ICO. IC-70514-H7K5, de 5 de agosto 2021, par. 17.

podría socavar la política educativa del Gobierno. Finalmente, Ofqual consideró que la divulgación podría llevar a la comparación y el escrutinio públicos de centros concretos sobre la base de la variación de los CAG y los ajustes realizados por el algoritmo, y dar lugar a una percepción distorsionada o a una comparación injusta entre centros menos fiables, más exigentes o más indulgentes que otros. Ello conduciría a un desvío de recursos para la gestión de la publicidad adversa que la divulgación podría causar, especialmente en los centros que se encontrasen en los extremos de la variación⁶².

A pesar de que el Comisionado estimó que la sección 36 de la FOIA había sido correctamente aplicada por Ofqual, sin embargo, consideró que el regulador había omitido la ponderación del test del interés público. A este respecto, el análisis realizado por el ICO se inclinó hacia la prevalencia del interés público en la difusión de la información. En primer lugar, el Comisionado entendía que la divulgación de la información solicitada permitía responder a las preocupaciones de la opinión pública sobre el algoritmo y permitía una mejor rendición de cuentas por parte de los centros respecto de cualquier discrepancia o mala aplicación en los CAGs asignados. A su vez, esto posibilitaría a la cohorte de estudiantes de A-level del curso 2020 verificar si su GAC se había asignado de forma injusta o discriminatoria. Por ejemplo, cualquier estudiante que considerase que su CAG era demasiado bajo podía tener mayores incentivos para impugnar esta estimación en aquellos centros en los que un número significativo de calificaciones fue revisado al alza en 2020. En segundo lugar, el Comisionado reconocía que la divulgación de la información podía generar la desviación de los recursos de los centros, pero consideró proporcionales tales desviaciones, teniendo en cuenta los procedimientos existentes, las incoherencias que la información revelada podía poner de manifiesto y el interés público continuado de esta información. En última instancia, la difusión de la información permitiría obtener una fotografía más amplia de un proceso que había dado lugar a incoherencias muy significativas y demostraría cuán justificada y generalizada era la preocupación por el diseño y funcionamiento del algoritmo. Los posibles efectos adversos que el escrutinio público y de los medios

⁶² *Idem*, parr. 19, 32, 34-36, 42 y 44.

podía tener sobre los centros quedaban superados por el hecho que la divulgación podía animar a los estudiantes, o a sus padres, a presentar reclamaciones como las que ya estaban en marcha⁶³.

¿Qué ocurrió entonces? Sencillamente que el diseño del algoritmo dio lugar a “sesgos técnicos que no se identificaron ni se corrigieron de forma proactiva”⁶⁴. Y ello a pesar de que la Royal Statistical Society había ofrecido inicialmente su ayuda, si bien tuvo que declinarla finalmente por el acuerdo de confidencialidad restrictivo que Ofqual pretendía imponer⁶⁵.

Algunos de las deficiencias del algoritmo fueron identificadas por los expertos tras analizar el extenso informe publicado por Ofqual donde se daban detalles del modelo. En primer lugar, el algoritmo no había sido aplicado de manera uniforme en los centros. Para aquellos con 5 o menos estudiantes, sólo se consideró la calificación asignada por los centros sin aplicar ningún proceso de estandarización; para los centros pequeños, Ofqual había utilizado una combinación entre el CAG y una versión reducida del algoritmo. Esto introdujo inmediatamente un primer punto de desigualdad en el sistema de calificación, ya que los centros con menor número de estudiantes tienen más probabilidades de ser privados, y las calificaciones previstas por los centros, con carácter general, eran más altas que las generadas por el algoritmo. En segundo lugar, el proceso de normalización de las calificaciones a nivel nacional mediante la imputación de calificaciones de corte pseudo-numéricas (*notional cut-scores*) tuvo como consecuencia que las más altas calificaciones de los alumnos pertenecientes a centros pequeños hicieron bajar las calificaciones de los alumnos de las escuelas de centros grandes, que ya de por sí tenían su calificación estimada a la baja por el algoritmo. Dado que los centros con mayor número de alumnos tienen más probabilidades de ser centros públicos, y los pequeños tienen más probabilidades de ser centros privados, ello incrementa aún más la injusticia y la disparidad en las calificaciones imputadas a los primeros. En tercer lugar, en el proceso

⁶³ *Idem*, parr. 61-62, 65-67.

⁶⁴ S. Wachter, B. Mittelstadt y C. Russell, “Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law”, *West Virginia Law Review*, Vol. 123, núm. 735, 2021, p. 739. <https://x.gd/hKkqS>.

⁶⁵ J. Faraway, *An Alternative...*, cit.

de emparejamiento de las distribuciones históricas de las calificaciones con su cohorte correspondiente, el modelo diseñado no tuvo en cuenta la variabilidad de las calificaciones de un año a otro, lo que especialmente acabó afectando a los centros públicos ubicados en zonas más deprimidas. En cuarto lugar, la influencia decisiva de los rankings junto con el proceso de estandarización a nivel nacional mediante calificaciones pseudo-numéricas de corte y el correspondiente sistema de imputación acabó generando situaciones paradójicas en la que dos estudiantes con rendimiento casi idéntico en el mismo centro y con el mismo CAG acabaran con calificaciones diferentes⁶⁶.

V. EN PARTICULAR, IMPORTANCIA DE LA VALIDACIÓN DE LOS MODELOS ALGORÍTMICOS Y EL PROBLEMA DE LA AMPLIFICACIÓN DEL SESGO. DE NUEVO, OFQUAL

Como señala Zlotnik para los sistemas de IA implementados por las organizaciones, incluidas las Administraciones Públicas, tanto los errores como sus potenciales impactos (individuales o sociales) deben ser asumibles en todo momento por la organización⁶⁷. Lo afirmado vale lo mismo para sistemas de IA, como para cualquier modelo predictivo, como el elaborado por Ofqual.

En este sentido, una de las limitaciones detectadas por los expertos en el modelo de Ofqual fue precisamente la falta de cuantificación de la incertidumbre, en el sentido de que las calificaciones imputadas por el algoritmo (predicciones) no tuvieron en cuenta en la práctica la relevancia del error en las estimaciones correspondientes. Esto último era particularmente problemático por dos motivos: por un lado, la precisión general estimada del modelo no era ni muchos menos buena; por otro, se obvió que la mayoría de los modelos estadísticos tien-

⁶⁶ Vid. T. Harkness, “How Ofqual Failed...”, Op. cit.; S. Bennett, *On A Levels...*, cit.; T. Paulden, “A cutting re-mark”..., cit.; G. Constantinides, “A-Levels and GCSEs in 2020”, *Thinking* [blog]. 15 de agosto 2020. <https://x.gd/yLHRp>.

⁶⁷ A. Zlotnik, “Inteligencia Artificial en las Administraciones Públicas: definiciones, evaluación de viabilidad de proyectos y áreas de aplicación”, *Boletic*, núm. 84, 2019, pp. 27-28.

den a ser limitados en su capacidad de predecir con precisión los resultados para los sujetos individuales (frente a grupos poblacionales).

Teniendo en cuenta que la precisión tanto en los modelos estadísticos como en los de IA, se refiere a la frecuencia con la que un modelo de IA acierta las respuestas correctas a partir de los datos de prueba correctamente etiquetados⁶⁸, ¿cómo había procedido Ofqual para validar el modelo y determinar su precisión? En el documento técnico publicado por el regulador se explicaba que el algoritmo en cuestión se había validado para diferentes asignaturas utilizando distintos modelos, entre otros, la de regresión lineal y de regresión logística —utilizados también en sistemas de IA⁶⁹.

En primer lugar, el método utilizado por el regulador fue calificado como “defectuoso” por los expertos⁷⁰.

En efecto, para evaluar la precisión del modelo, Ofqual había utilizado datos históricos de la cohorte de 2019. Sin embargo, no existían rankings elaborados por los centros para esta cohorte histórica, como sí ocurría en la cohorte de 2020. Esto generó circularidad en el proceso de validación del modelo pues, en la práctica, Ofqual utilizó los datos de calificaciones reales de 2019 —en lugar de los rankings, como así se hacía para la cohorte de 2020— para predecir las calificaciones de 2019. Este defectuoso proceso de validación de la precisión en realidad condujo a una sobreestimación de la precisión real del modelo utilizado para predecir las calificaciones de 2020. ¡Sería el equivalente en IA a darle a un modelo de aprendizaje supervisado las calificaciones de los exámenes A-Level de 2019 para predecir las calificaciones de 2019! Es decir, un evidente problema de *overfitting* o sobreajuste que invalidaría todo el modelo. En segundo lugar, la precisión del modelo final no era realmente fiable. Existía una variación significativa

⁶⁸ ICO, *Guidance on AI and data protection*, 2020, p. 38. <https://x.gd/ivtYM>.

⁶⁹ Vid. ICO y The Alan Turing Institute, *Explaining decisions made with AI*, 2020. <https://x.gd/gcYA2>. Entre las técnicas algorítmicas se incluyen tanto la regresión lineal como la logística.

⁷⁰ Un análisis detallado de los principales defectos en la validación del algoritmo de Ofqual puede verse en S. Bennett. *On A Levels...*, cit.; T. Paulden. “A cutting re-mark” ..., cit.

(aproximadamente entre el 40%-75%) en función de las asignaturas y el modelo de validación considerado por Ofqual en su Informe⁷¹.

Ello era consecuencia de predecir la distribución del año 2020 a partir de los años anteriores y dado que la validación del modelo se hizo del modo defectuoso descrito, sencillamente, no era posible conocer la verdadera precisión del modelo, pero desde luego algunos expertos estimaban que sería inferior a la oficialmente comunicada por Ofqual en su Informe.

En tercer lugar, la precisión disminuía considerablemente en función del tamaño de la escuela, teniendo en cuenta, además, que el uso variable del algoritmo en función del tamaño de la escuela introducía injusticia en el proceso. En efecto, esta validación no era realmente una validación de todo el algoritmo, en la medida en que el algoritmo real y final utilizado por Ofqual no se aplicó de manera uniforme o sencillamente no se aplicó.

Asimismo, la automatización de decisiones y, muy particularmente, la implementación de sistemas de IA puede amplificar los sesgos discriminatorios y desigualdades sociales ya existentes e, incluso, distorsionar la finalidad de las políticas públicas⁷², como consecuencia, por ejemplo, de un modelo desarrollado a partir de unos datos de entrada o entrenamiento incompletos o no representativos, o la selección de datos procedentes de fuentes sesgadas⁷³.

La denominada “automatización del sesgo” implicaría la complacencia y dependencia excesiva del usuario del modelo con los resultados generados por el modelo algorítmico sin entrar a cuestionar

⁷¹ Así, por ejemplo, para las asignaturas de Biología, Francés, Drama y Estudios religiosos, con el enfoque de validación núm. 8 basado en un modelo de regresión lineal se obtenía una precisión inferior al 0,4 (e.g. Biología). Y, con el enfoque de validación núm. 3 —basado en una variante del modelo utilizado y el finalmente tomado en consideración—, se obtuvo una precisión por debajo del 0,7 (e.g. Biología). Vid. Ofqual. *Interim Report...*, cit., pp. 52-54.

⁷² D. Citron y D. Calo, *The Automated Administrative State*. Harvard Kennedy School, Shorenstein Center, sp, 2019.

⁷³ D. Bourcier y P. De Filippi “La transparence des algorithmes face à l’Open Data: Quel statut pour les données d’apprentissage?”, *Revue française d’Administration Publique*, ENA, 2018, pp. 7, 12-14, 15.

la adecuación, razonabilidad, validez y justicia de los resultados del modelo⁷⁴.

Teniendo en cuenta lo anterior, Ofqual evaluó si el algoritmo sesgaba injustamente a ciertos grupos de individuos. Sin embargo, sólo analizó lo que ocurre cuando todos los estudiantes y escuelas reciben las calificaciones estimadas por el algoritmo, pero no si algunas escuelas —predominantemente las más acomodadas— se benefician de obtener las calificaciones más altas, al utilizarse el CAG, en lugar del algoritmo. Además, sólo se testó el impacto del algoritmo en la igualdad en un pequeño subconjunto de los modelos que se probaron⁷⁵.

A causa de las deficiencias en el diseño ya comentadas y en la validación del modelo, los alumnos de alto rendimiento de los centros de alto rendimiento recibieron calificaciones altas, mientras que los alumnos de alto rendimiento en escuelas de bajo rendimiento vieron cómo sus calificaciones fueron limitadas por el rendimiento anterior de otros estudiantes y recibieron una nota más baja a la que merecían. En la práctica, esta tendencia afectó de forma más desproporcionada a los estudiantes negros, asiáticos y minorías étnicas⁷⁶.

No sorprende, en este sentido, que el propio ICO considerase esta circunstancia en su resolución estimatoria sobre los resultados desagregados del algoritmo de Ofqual al ponderar el interés público en el acceso: “Se temía que el propio algoritmo fuera ilegal, no solo por infringir la legislación antidiscriminatoria⁷⁷, sino también el artículo

⁷⁴ R. Binns y V. Gallo, *Automated Decision Making: the role of meaningful human reviews*, Information Commissioner's Office, sp, 2019.; M. Veale, M. van Kleek, y R. Binns, “Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making”, *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, April. Paper No. 440, 2018, p. 4.

⁷⁵ Bennett. *On A Levels...*, cit.

⁷⁶ S. Wachter et al. “Bias Preservation in Machine Learning...”, cit., pp. 738-739. <https://researchrepository.wvu.edu/wvlt/vol123/iss3/4>

⁷⁷ En particular, la *Equality Act* 2010, que protege legalmente a las personas frente a la discriminación directa, indirecta, el acoso o la victimización en el lugar de trabajo, en la educación, como usuarios de los servicios públicos y en la sociedad en general por motivos de edad, cambio de sexo, situación matrimonial o de pareja de hecho, embarazo o baja por maternidad, discapacidad, raza (incluido el color, la nacionalidad, el origen étnico o nacional), religión o creencias, sexo, y orientación sexual.

22 del RGPD que reconoce el derecho a no ser objeto de una toma de decisiones totalmente automatizada que afecte significativamente a las personas. El recurrente ha realizado esta solicitud con base en la preocupación existente de que los estudiantes que asisten a centros de menor rendimiento procedentes de zonas más desfavorecidas se hubiesen visto perjudicados por el algoritmo”⁷⁸.

VI. SOBRE EL “VALOR IUSFUNDAMENTAL” DEL DERECHO A SABER QUÉ TRATAMIENTOS ALGORÍTMICOS APLICA LA ADMINISTRACIÓN. LECCIONES FINALES DE PARCOURSUP

A pesar de las disposiciones contenidas en los arts. L. 311-3-1 y L. 312-1-3 del CRPA y de las numerosas resoluciones de la CADA sobre la plataforma nacional Parcoursup, la realidad es que ni todo el código fuente ni todos los tratamientos algorítmicos implementados eran públicos. Así lo acreditó la auditoría externa encargada por el Tribunal de Cuentas francés que evaluó la eficacia y equidad de los procedimientos de clasificación de candidatos.

En efecto, la auditoría concluía que, en la práctica, los elementos publicados por el Ministerio de Enseñanza Superior, Investigación e Innovación (MESRI) en mayo de 2018 sólo representaban el 1% de las líneas de código y menos de un 2% de los ficheros JAVA y SQL del código fuente⁷⁹. Sin embargo, en el código fuente de la aplicación existirían dos tipos de tratamientos algorítmicos. Por un lado, los “algoritmos locales” incluidos en la herramienta de apoyo a la decisión que podían utilizar los actores locales, como las Comisiones Académicas de Evaluación de las Solicitudes (CEV) o los servicios de rectorado de cada institución, que clasificaban las solicitudes de los candidatos. Por otro lado, los algoritmos que generaban la clasifica-

⁷⁸ ICO. IC-70514-H7K5, de 5 de agosto 2021, par. 55.

⁷⁹ De hecho, los archivos y líneas de código en SQL se habían hecho públicos eran cuantitativamente inferiores a los publicados en Java. La relevancia de dar publicidad a esta información estriba en que los archivos en SQL permiten estructurar y analizar datos, mientras que los escritos en Java permiten desarrollar aplicaciones e implementar cálculos algorítmicos en conexión con los archivos SQL.

ción final de las solicitudes a partir de las clasificaciones efectuadas por los actores locales. Solo estos últimos se habían hecho públicos⁸⁰, —a pesar de que en palabras del Tribunal de Cuentas— tal información tenía un “interés limitado para garantizar la transparencia del sistema”. Respecto de los primeros, mediante técnicas de aprendizaje automatizado basadas en bosques aleatorios, la auditoría analizó los conjuntos de datos utilizados por la aplicación, llegando a identificar hasta “15.000 algoritmos locales”. Entre las conclusiones recogidas por el Informe del Tribunal de Cuentas, se incluía la recomendación al MESRI de que se generalizase la publicidad para una parte más amplia de la aplicación, en particular con respecto a los elementos relacionados con el funcionamiento de los CEV, a fin de informar el debate público sobre los “algoritmos locales” y el apoyo a la decisión a través de medios automáticos⁸¹.

Junto a los numerosos recursos presentados por la Union Nationale des Étudiants de France (UNEF) en vía contenciosa contra las resoluciones desestimatorias de las Universidades de comunicar sus algoritmos locales, algunas instituciones relevantes venían recomendando también la necesidad de publicar los algoritmos locales.

La propia CADA así lo había hecho, criticando la opción adoptada del legislador francés, al establecer un régimen especial de acceso restringido exclusivamente a los candidatos que solicitaran información sobre los procesos algorítmicos cuando estos fuesen implementados

⁸⁰ A saber, el código Java que permitía el cálculo y la implementación del orden de llamada, de las propuestas de formación, de las propuestas de alojamiento en residencias universitarias, la aplicación del esquema “Mejores bachilleres”, y la aplicación del contestador automático.

⁸¹ Cour des Comptes, *Un premier bilan de l'accès à l'enseignement supérieur dans le cadre de la loi orientation et réussite des étudiants. Communication au comité d'évaluation et de contrôle des politiques publiques de l'Assemblée nationale*, febrero 2020, pp. 53-54, 66 y 142. Para el Tribunal de Cuentas, la publicación de los algoritmos locales utilizados por las CEV respondía primeramente a un objetivo pedagógico evidente para los usuarios de la plataforma que podían conocer cómo se ponderaban las distintas disciplinas de su formación previa en cada Universidad y estudios universitarios. En segundo lugar, dicha publicidad respondía, además, a un “imperativo de transparencia”, pues si bien no se cuestionaba la utilización de algoritmos locales para clasificar las candidaturas —debido a su gran volumen, la ausencia de publicidad y el rechazo a su comunicación a terceros generaba un “riesgo de desconfianza.”

por los equipos educativos encargados de examinar las solicitudes, con exclusión del derecho de acceso de terceros. La Autoridad francesa consideraba que, si bien las disposiciones Código de Educación eximían a los establecimientos universitarios “de la obligación de difundir en línea las reglas que definen los principales tratamientos algorítmicos utilizados [...] cuando son la base de decisiones individuales [...], no prohíben, sin embargo, asegurar su difusión espontánea”⁸².

Por su parte, la CNIL había considerado que, si bien en teoría la fase de tratamiento de los dosieres de los candidatos por las CEV no estaba totalmente automatizada a los efectos del art. 22 del RGPD, sin embargo, no tenía una certeza absoluta de que esto fuera así en la práctica. En consecuencia, la CNIL recomendaba como “buena práctica” la difusión de los algoritmos locales⁸³.

La cuestión finalmente quedó dirimida por el Consejo Constitucional francés en su Decisión n° 2020-834, de 3 de abril 2020, al resolver la cuestión de inconstitucionalidad (“question prioritaire de constitutionnalité”) planteada por la UNEF. El sindicato reclamante estimaba que el último párrafo del art. L. 612-3. I del Código de Educación vulneraba los derechos y libertades garantizados por la Constitución de 1958, y otros textos a los que reenvía el Preámbulo del texto constitucional, en particular, la Declaración de Derechos del Hombre y del Ciudadano de 1789.

El precepto impugnado establece que: “Para garantizar la necesaria protección del secreto de las deliberaciones de los equipos educativos encargados del examen de las solicitudes presentadas en el marco del procedimiento nacional de preinscripción [...], las obligaciones derivadas de los artículos L. 311- 3-1 y L 312-1-3 del Código de Relaciones entre el Público y la Administración se considerarán satisfechas cuando se informe a los candidatos de la posibilidad de obtener, si así lo solicitan, la comunicación de la información relativa a los criterios y procedimientos para el examen de sus solicitudes, así como las razones pedagógicas que justifican la decisión tomada”.

⁸² CADA. Opinión 20184400, de 10 de enero de 2019.

⁸³ Audición de 17 de julio de 2019 de Mme Marie-Laure Denis, presidenta de la CNIL.

Según la jurisprudencia del Consejo de Estado, al excluir el régimen general previsto en el CRPA, el precepto impugnado establece un acceso cualificado a los documentos administrativos relativos a los tratamientos algorítmicos para la evaluación de las candidaturas, únicamente en favor de los candidatos que lo solicitan, una vez adoptada la correspondiente decisión, y únicamente respecto a la información relativa a los criterios y las modalidades de examen de su candidatura.

Pues bien, el sindicato reclamante consideraba que estas disposiciones serían contrarias al derecho a la comunicación de documentos administrativos reconocido en el artículo 15 de la Declaración de los Derechos del Hombre y del Ciudadano de 1789, al excluir el acceso de terceros y de los candidatos a los algoritmos susceptibles de ser utilizados por las instituciones universitarias para evaluar las solicitudes de preinscripción en el primer ciclo formuladas a través de la plataforma “Parcoursup”. Sin embargo, tal exclusión no estaría justificada ni por el secreto de las deliberaciones de los órganos colegiados ni por ninguna otra razón. Además, estas disposiciones violarían el derecho a la tutela judicial, al impedir no sólo la impugnación efectiva contra la falta de comunicación de los algoritmos implementados, sino privando también a los afectados de la posibilidad de atacar el fondo de las denegaciones de preinscripción.

El Consejo Constitucional recordaba, en primer lugar, que en virtud del art. 15 de la Declaración de 1789: “La sociedad tiene derecho a pedir cuentas a cualquier funcionario público de su Administración”. Este precepto garantiza el derecho de acceso a los documentos administrativos. Corresponde al legislador la posibilidad de establecer limitaciones a este derecho vinculadas a exigencias constitucionales o justificadas por el interés general, siempre que ello no se traduzca en restricciones desproporcionadas con respecto al objetivo perseguido. En este sentido, consideraba el Consejo que las limitaciones al acceso general previstas en la disposición impugnada se justificaban, a priori, por un motivo de interés general —la protección del secreto deliberatorio de las Comisiones de Evaluación— y resultaban proporcionadas a dicho objetivo.

En segundo lugar, el Consejo estimó que el procedimiento nacional de preinscripción a través de la plataforma no estaba totalmente automatizado. En primer lugar, porque el uso de algoritmos para realizar

la evaluación de las candidaturas era una facultad discrecional de las instituciones universitarias.

En segundo lugar, y en caso de implementar algoritmos, la decisión adoptada sobre cada candidatura no podía fundamentarse exclusivamente en tales tratamientos, sino que se exigía una evaluación de los méritos de las candidaturas por las CEV y, a continuación, por el Rector de la institución.

En tercer lugar, una vez adoptada una decisión de denegación al respecto, los candidatos podrán obtener por parte de la institución, cuando lo soliciten, la comunicación de la información sobre la jerarquización y la ponderación de los diferentes criterios establecidos por las instituciones, las modalidades de examen de sus candidaturas, así como de los motivos pedagógicos que justifican la decisión adoptada al respecto. La comunicación prevista en el precepto cuestionado también podrá incluir información sobre los criterios utilizados por los posibles tratamientos algorítmicos implementados por las comisiones de examen.

Sin embargo, este acceso cualificado sólo beneficiaría a los candidatos. En consecuencia, la exclusión del derecho de acceso de los terceros a cualquier información sobre los criterios y las modalidades de examen de las candidaturas efectivamente exigidos por las instituciones constituiría una vulneración desproporcionada del derecho garantizado en el art. 15 de la Declaración de 1789 con respecto al objetivo de interés general derivado de la protección del secreto de las deliberaciones de los equipos docentes. Por consiguiente, la disposición impugnada no puede interpretarse en el sentido de eximir a las instituciones universitarias de la obligación de publicar un informe con los criterios en función de los cuales se evaluaron las candidaturas, especificando, en su caso, en qué medida se utilizaron tratamientos algorítmicos para proceder a dicha evaluación.

El “valor constitucional” del pronunciamiento —en palabras de la CADA⁸⁴— no puede pasar desapercibido. Al establecer cuál ha de ser la interpretación del precepto impugnado, el Consejo Constitucional anuda el derecho constitucional de acceso a los documentos administrativos con la obligación positiva de la Administración de publicar

⁸⁴ CADA. Opinión 20213847, de 13 de enero de 2022.

de forma proactiva la información sobre los tratamientos algorítmicos aplicados por ésta en las decisiones individuales que afectan a los ciudadanos.

Algoritmo y transparencia de la Administración pública en el ordenamiento jurídico italiano. Algunas reflexiones críticas a partir del caso “Buona Scuola”

SALVATORE PALUMBO

*Doctor por la Universitat de València y
la Università degli Studi di Palermo*

I. EL CASO “BUONA SCUOLA”

El caso judicial en cuestión surge en el contexto de los cambios regulatorios que se han producido en Italia sobre la disciplina para la contratación de personal de escuelas públicas, por la ley mencionada “Buona Scuola” o más técnicamente Ley n°107 de 13 de julio de 2015. Este texto legislativo —en particular los apartados 94-105 del artículo 1— preveía un plan extraordinario para la contratación de personal docente a nivel nacional, mientras que anteriormente la contratación se realizaba a nivel provincial. De esta manera, cualquier maestro que calificara para la contratación podría ser colocado en cualquier provincia donde hubiera plazas disponibles. Al rechazar la propuesta de trabajo, se impusieron una serie de sanciones, a saber, la pérdida de la oportunidad de contratar, la expulsión de la lista, así como el desplazamiento de la misma al siguiente profesor en espera de plaza, hasta la cobertura de todos los puestos que el plan extraordinario había previsto cubrir. Luego, cada maestro tenía que indicar la provincia de preferencia y las demás elegibles para la elección. La solicitud debía presentarse exclusivamente de forma electrónica, en una plataforma digital preparada por el “Ministero della Università e della Ricerca” (MIUR) para la recogida de datos, la tramitación de los mismos y la formulación de una propuesta de contratación para cada persona legitimada.

Por lo tanto, el procedimiento se llevó a cabo en las siguientes fases: la primera se refiere a los movimientos de docentes nombrados

hasta el 2014/2015 dentro de la provincia de titularidad (fase A), y la segunda se refiere al movimiento de docentes entre diferentes provincias, y se divide a su vez en tres fases adicionales (fase B, C y D).

Los profesores se quejaron, sin embargo, de que habían sido asignados a lugares alejados del elegido, y que habían sido ocupados por competidores con puntuaciones más bajas. En este sentido, los sindicatos actuaron en contra del Ministerio para tener acceso al algoritmo que guió el procedimiento de selección. Al principio, el MIUR denegó esta solicitud de acceso justificándola sobre la base de la disciplina de la propiedad intelectual, ya que el *software* está protegido como una obra del ingenio, como las obras literarias, cinematográficas, etc. De hecho, el algoritmo fue desarrollado por una empresa privada (HPE s.r.l., agente de RTI entre HPE srl y Finmeccanica Spa), en nombre del Ministerio para gestionar las operaciones de movilidad del personal docente para el año 2016/2017.

El razonamiento es que, en la medida en que se trata de una propiedad intelectual sujeta a su protección y dado que no constituye en sí misma un documento administrativo, no puede ser objeto de una solicitud de acceso en virtud de la Ley n° 241, de 7 de agosto de 1990 (Nuove norme in materia di procedimento amministrativo e di diritto di accesso ai documenti amministrativi)¹.

Durante algunos juicios, la administración produjo entonces parte del código fuente², lo que permitió a algunos técnicos aportar una “pericia técnica en el análisis del algoritmo que gestiona el software

¹ E. Prosperetti, “Accesso al software e al relativo algoritmo nei procedimenti amministrativi e giudiziali. Un’analisi a partire da due pronunce del Tar Lazio”, *Il Diritto dell’Informazione e dell’Informatica*, núm. 4, 2019, pp. 979 y ss.

² La controversia ha sido objeto de múltiples y diferentes recursos para requerir la intervención de le Sezioni Unite de la Corte di Cassazione en tema de reparto de jurisdicción. Con la sentencia núm. 8821, de 10 de abril de 2018, la Corte ha establecido que, en relación con la movilidad del profesorado, las medidas con las que estos últimos se asignan a ámbitos territoriales no indicados principalmente en sus solicitudes de movilidad se refieren a la fase ejecutiva de la relación laboral. Con la consecuencia obvia de que, en caso de oposición a estas medidas, el juez competente para pronunciarse sobre la oposición es el ordinario y no la autoridad judicial administrativa.

de movilidad docente”³. Algunos pasajes del informe técnico parecen útiles para comprender el funcionamiento real del *software* “acusado”. Del examen realizado por los técnicos de parte sobre los expedientes facilitados por el Ministerio se desprende que el código fuente en cuestión, relativo a la fase A, se desarrolla en el lenguaje de programación denominado COBOL (Common Business-Oriented Language), un lenguaje de alto nivel⁴. En concreto, se trata de un lenguaje diseñado a principios de los años 60 y utilizado principalmente para el desarrollo de *software* de aplicación comercial como, por ejemplo, el *software* bancario y contable. En cuanto a las fases B, C y D, sin embargo, cabe señalar que el código relativo se ha desarrollado de una manera diferente, en lenguaje C.

A diferencia de COBOL, los programas hechos en C tienen un tamaño más pequeño y son más eficientes gracias al lenguaje estructurado del propio código, incluso presentan un método más pobre para identificar errores⁵.

A primera vista se observa que la redacción de los algoritmos relacionados con la fase A y las fases B, C y D se llevó a cabo, no solo en diferentes lenguajes, sino presumiblemente por diferentes sujetos que tienen un código estilístico diferente, así como semántica literal en la programación de *software*⁶. En la actualidad no es posible hacer un

³ El informe realizado por técnicos cualificados se puede consultar en <https://www.orizzontescuola.it/algoritmo-mobilita-2016-confuso-lacunos-perizia-technica-svela-pasticcio-dirottato-insegnanti-lontano-dalla-provincia-scelta/>

⁴ Se han introducido lenguajes de alto nivel en informática con el fin de acelerar la fase de programación del *software* y permitir realizar varias tareas. Un programa escrito en lenguaje de alto nivel, sin embargo, para ser decodificado por la computadora (que es capaz de leer solo secuencias de dígitos binarios) debe estar necesariamente sujeto a un programa de servicio, capaz de traducir las instrucciones de inicio al lenguaje ensamblador. Por lo tanto, los pasos de la programación actual están marcados de la siguiente manera: el código escrito en lenguaje de alto nivel es traducido por el programa de traducción a instrucciones en lenguaje ensamblador; a su vez, el programa ensamblador convierte estas instrucciones en *bits* para cargarlas y ejecutarlas en la memoria del ordenador (E. Peres, *Che cosa sono gli algoritmi*, Milano, Salani, 2019). Los programas de traducción son esencialmente de dos tipos, compiladores e intérpretes.

⁵ Véase la nota 3.

⁶ El informe pericial continúa revelando varias cuestiones críticas que deben leerse en todo caso con vista a una defensa experimentada por la parte procesal. En

análisis completo sobre el formato de los datos de entrada y salida de las distintas fases ya que esta información no ha sido proporcionada por la empresa y el Ministerio, como también ha sucedido con otros elementos dentro del código.

Por lo tanto, el procedimiento se caracterizó por una clara falta de transparencia, en el sentido de que cuando la administración proporcionó el software en el curso del juicio, lo hizo produciendo archivos que contenían el código fuente, pero esos archivos como se desprende del informe son solo parciales y parcialmente oscurecidos. La actitud reticente se debe principalmente a la incertidumbre que rige la propiedad intelectual del software que se desarrolla para la administración pública (en adelante también a. p.).

Precisamente de esta última criticidad trataremos de ocuparnos en las próximas páginas.

II. USO DE SOFTWARE EN LA ADMINISTRACIÓN PÚBLICA Y PROTECCIÓN DE LA PROPIEDAD INTELECTUAL

El derecho de propiedad intelectual sobre las obras de propiedad intelectual desarrolladas por particulares para la administración pública —como los programas informáticos para llevar a cabo procedimientos—⁷ encuentra una base directa en el artículo 42 de la Consti-

particular, se observa que la práctica para combinar programas de diferentes autores “lo es a menudo utilizado por los fabricantes de *software*, pero sigue siendo delicado cuando no se define un único estándar de referencia para la estructura del código que se va a desarrollar; estándar al que todos los desarrolladores deben volver. Especialmente si el objetivo es desarrollar múltiples sistemas que deben comunicarse entre sí y compartir datos y resultados (como determinar las vacantes disponibles en cada área que queda como resultado de las operaciones a las que se hace referencia en la fase A cuyos “datos de salida” deben convertirse en “datos de entrada” en las fases posteriores). En el caso de autos, esta norma no se respeta en toda la lista examinada”.

⁷ La legislación italiana está orientada a incluir la protección de los programas informáticos en la legislación sobre derechos de autor, excluyendo como norma general la patentabilidad. En concreto, el Decreto Legislativo n°30/2005 (Código de la Propiedad Industrial), dispone en el art. 45 que los denominados métodos matemáticos no se consideran invenciones y, por lo tanto, no son patentables.

tución italiana, que establece la protección de la propiedad privada⁸. Las limitaciones de la misma pueden ser previstas por la ley con el fin de garantizar su utilidad social.

De la lectura general de la disposición constitucional se desprende, por tanto, que las formas de limitación de la propiedad privada deben estar expresamente previstas y reguladas por la ley⁹.

Sin embargo, se llega a conclusiones insatisfactorias en cuanto a la protección de este derecho analizando el complejo del ordenamiento jurídico resultante de las normas de derecho positivo según la interpretación de los tribunales administrativos.

Originalmente el Decreto legislativo n°39 de 1993, que contiene “Normas relativas a los sistemas de información automatizados de las administraciones públicas”, estableció que las administraciones, incluso si hubieran confiado a terceros el diseño, desarrollo y gestión de sistemas de información automatizados, conservaban la propiedad de los programas (artículo 2, apartado 3). Esta disposición era coherente con lo que (todavía) prevé la Ley de Derecho de Autor, según la cual “las administraciones del Estado, las provincias y los municipios tienen derecho a los derechos de autor sobre las obras creadas y publicadas bajo su nombre y en su nombre y gastos”¹⁰.

En tiempos relativamente recientes, sin embargo, el legislador, con el Código de Administración Digital (Decreto legislativo n°82 de

Según una parte de la doctrina, esta elección se debe a la necesidad de no sobrecargar a las oficinas de patentes cfr. M. Capparelli, “Le nuove frontiere del diritto d'autore alla prova dell'Intelligenza Artificiale”, en U. Ruffolo, *Intelligenza artificiale: Il diritto, i diritti, l'etica*, Milano, Giuffrè, 2020., pp. 345 y ss.

⁸ Puede ser útil recuperar el texto del art. 42: “La propiedad es pública o privada. Los bienes económicos pertenecen al Estado, a entidades o a particulares.

La propiedad privada está reconocida y garantizada por la ley, que determina las formas de compra, disfrute y límites para asegurar su función social y hacerla accesible a todos.

La propiedad privada podrá, en los casos previstos por la ley, y sujeta a indemnización, ser expropiada por razones de interés general.

La ley establece las normas y límites de la sucesión legítima y testamentaria y los derechos del Estado sobre las herencias”.

⁹ El debate que ha implicado formas de abuso de autoridad pública sobre el derecho a la propiedad es bien conocido cfr. Corte costituzionale, Sentencia n°71 de 2015.

¹⁰ Art. 11, Ley n°633 de 1941.

2005 o más sencillamente CAD) ha trabajado por un enfoque parcialmente diferente. En concreto, el art. 2 citado fue derogado por d.lgs. 179 de 2016, incluido en la reforma más amplia llamada Madia¹¹ y su disciplina se ha transpuesto al CAD.

Más en detalle, el legislador italiano a través del Decreto Legislativo n°82 de 7 de marzo de 2005, ha obrado para un marco de soluciones diferenciadas.

En particular, el texto regulatorio ahora citado en el art. 68 (Análisis comparativo de soluciones), establece que la a. p. debe realizar una evaluación comparativa de carácter técnico y económico, de acuerdo con los criterios especificados en el apartado 1-bis; sólo si como resultado de esta evaluación se comprueba que es imposible acceder a soluciones ya disponibles en la administración pública, o a *software* libre o códigos fuente abiertos¹², adaptados a las necesidades a satisfacer, se permite la adquisición de programas informáticos propietarios mediante el recurso a “licenza d’uso”.

Como demuestra la doctrina sectorial italiana¹³, esta última es una solución negociadora caracterizada por la característica esencial de no ceder los derechos de explotación económica del producto.

El Decreto legislativo n° 217 de 13 de diciembre de 2017 ordenó la introducción de las letras l-bis) y l-ter) al art. 1 del CAD. Concretamente, la disposición establece las siguientes definiciones: l-bis)

¹¹ Sobre este tema véase G. Amoroso, “La riforma Madia: rapporto tra legge delega e decreti attuativi”, en *Il libro dell'anno del diritto 2018*, Roma, Istituto della Enciclopedia italiana, 2018, pp. 335-340.

¹² En realidad, estos términos a menudo se confunden, pero usando las palabras del autor que dio a luz a estas categorías podemos decir que los dos términos describen aproximadamente la misma categoría de *software*: “Pero se basan en valores fundamentalmente diferentes. El código abierto es una metodología de desarrollo; el *software* libre es un movimiento social. Para el movimiento del *software* libre, el *software* libre es un imperativo ético, el respeto esencial a la libertad de los usuarios. Por el contrario, la filosofía del código abierto piensa en cómo “mejorar” el *software* solo desde un punto de vista práctico. Dice que el *software* no libre es una solución no óptima. A menudo, las discusiones sobre el “código abierto” no consideran lo que está bien o mal, sino solo el éxito y la popularidad” (R. Stallman, *Perché l’“Open Source” manca l’obiettivo del Software Libero*, en <https://www.gnu.org/philosophy/open-source-misses-the-point.html>).

¹³ M.G. Jori, *Diritto, nuove tecnologie e comunicazione digitale*, Milano, Giuffrè, 2013, pp. 33 y ss.

formato abierto: un formato de datos hecho público, plenamente documentado y neutro con respecto a las herramientas tecnológicas necesarias para el uso de los datos; l-ter) datos de tipo abierto: los datos que tienen las siguientes características: 1) están disponibles bajo los términos de una licencia o una disposición reglamentaria que permite su uso por cualquier persona, incluso con fines comerciales, en un formato desagregado; 2) ser accesibles a través de las tecnologías de la información y la comunicación, incluidas las redes telemáticas públicas y privadas, en formatos abiertos en el sentido de la letra l-bis), sean adecuadas para el uso automático por programas de ordenador y estén provistas de los metadatos pertinentes; 3) se pongan a disposición gratuitamente a través de las tecnologías de la información y la comunicación, incluidas las redes telemáticas públicas y privadas, o se pongan a disposición a los costos marginales incurridos para su reproducción y difusión, salvo lo dispuesto en el artículo 7 del Decreto legislativo n°36 de 24 de enero de 2006.

Así, el Código da una definición de datos abiertos, pero no de los programas pertinentes. Para estos últimos, se trata de programas que no prevén una licencia propietaria, por lo que potencialmente cualquier persona puede acceder al código fuente y a su funcionamiento, para consultarlo o reutilizarlo.

En la práctica, se ha generalizado el hábito de adquirir licencias de los titulares de programas informáticos comerciales exigiendo al proveedor que los adapte a las necesidades de la administración, pero de los que el proveedor ha conservado la titularidad¹⁴.

El CAD ha avalado así esta costumbre, permitiendo a las administraciones “la adquisición de programas informáticos propietarios”. En estos casos, por lo tanto, encuentra aplicación el art. 69 (apartado 2), que dispone que, a menos que ello sea excesivamente gravoso por razones técnicas y económicas demostradas, la administración contratante es siempre la titular de todos los derechos sobre los programas y servicios de las tecnologías de la información y la comunicación, desarrollados específicamente para ella.

¹⁴ Cf. C. Flick, G.A. Cignoni, y V. Ambriola, “Il riuso del software nella pubblica amministrazione”, *Dir. Internet*, núm. 1, 2008.

Por lo tanto, de conformidad con las disposiciones combinadas de los artículos. 68 y 69 del CAD sobre la reutilización, una administración puede considerarse propietaria de un programa informático en dos ocasiones: (1) programa realizado sobre sus propias indicaciones específicas, siempre que haya encargado tal solución —a través de un contrato de licitación u otros casos de negociación de naturaleza similar, incluso independientemente del título dado al contrato— y el contrato prevé la adquisición por él de todos los derechos de propiedad intelectual e industrial sobre el *software* desarrollado por cuenta de la Administración Pública, o en todo caso no atribuye esta titularidad al contratista o a terceros; o (2) el *software* ha sido desarrollado por recursos dentro de la propia administración¹⁵.

La normativa citada parece exigir a toda administración, al negociar un contrato destinado a encargar el desarrollo de un *software*, que se asegure la plena y exclusiva titularidad de todos los derechos sobre el *software* objeto de desarrollo, salvo que resulte excesivamente oneroso por razones técnicas y económicas justificadas.

Sin embargo, en cuanto a los resultados de la jurisprudencia, no son particularmente homogéneos. Más específicamente, una parte de la jurisprudencia italiana sobre el tema ha señalado que “no puede ser relevante la confidencialidad invocada de las empresas que producen los mecanismos informáticos utilizados que, al poner estas herramientas al servicio del poder autoritario, evidentemente aceptan sus consecuencias en términos de la transparencia necesaria” (así el Consiglio di Stato, Sezione VI, Sentencia n°8472 de 2019).

Esta solución parece contradicha por otra parte de la jurisprudencia, según la cual, aun cuando la administración pública “haya gestionado el procedimiento selectivo con un desarrollo “informatizado” —mediante la utilización de un programa, teniendo como código fuente el algoritmo pertinente en la presente sentencia— no afecta a la necesidad de buscar, sobre la base del documento al que se va a acceder, la existencia de eventuales partes contrarias, pudiendo existir contrapartes interesadas aunque el documento se utilice en el contex-

¹⁵ En este sentido AGID, *Linee guida su acquisizione e riuso di software per le pubbliche Amministrazioni*, 9 mayo 2019.

to de actividades gestionadas directamente por la Administración” (así el Consiglio di Stato, Sezione VI, Sentencia n°30 del 2020).

En este sentido, confirmó que el tercero proveedor de la aplicación digital está calificado como contraparte interesada y como autor material del *software* está obligado a mostrarlo, siendo el titular del algoritmo solicitado por la parte.

La doctrina no siempre está de acuerdo con respecto a las soluciones de la jurisprudencia, y se nota la falta de una disposición legislativa que imparta la exhibición del código del programa. En este sentido, se argumenta que el derecho de acceso, en sí mismo, no es capaz de sobrepasar los límites del derecho de autor y que, por tanto, incluso ante una solicitud de acceso, es necesario, en teoría, plantear el problema de los derechos para permitir a quienes acceden utilizar la obra a la que se concede el acceso¹⁶.

Una solución similar está presente en parte de la jurisprudencia española. En concreto, el Juzgado Central de lo contencioso administrativo, solicitado por el recurso de una asociación de consumidores, se ocupó de quién ostenta el código fuente de la aplicación informática denominada BOSCO, utilizada por el Gobierno español para reconocer determinados descuentos en la factura de la luz¹⁷.

Más en detalle, la aplicación informática verifica principalmente los datos de ingresos en poder de la administración y, sobre la base de su procesamiento, transmite al proveedor de energía la indicación sobre la presencia o ausencia de los requisitos para otorgar un descuento en la factura (Bono Social). Ante la negativa de la administración a proporcionar información sobre el funcionamiento de ese sistema informático, la fundación independiente Civio denunció ante el tribunal la violación de diversos parámetros normativos como el principio de legalidad, el derecho de acceso o el derecho de motivación.

El citado Tribunal con Sentencia n°143/2021 desestimó las quejas de la demandante, al considerar que la denegación de acceso al código fuente cumple con los límites reglamentarios establecidos para proteger la propiedad intelectual y la seguridad pública.

¹⁶ En este sentido E. Prosperetti, cit.

¹⁷ La disciplina principal está establecido por el Real Decreto n°897/2017, del 6 octubre, especialmente art. 8.

Finalmente, volviendo al ordenamiento jurídico italiano, cabe señalar que, por un lado, el legislador admite la compra de *software* propietario con licencia de uso, dejando a negociación cualquier perfil relacionado con el ejercicio del derecho de explotación, por otro, especialmente en la lectura dada por una parte de la jurisprudencia, parece vaciar la propiedad intelectual de cualquier contenido preceptivo. Este marco normativo lleva a la conclusión de que en la actualidad no existe una normativa completa que regule los límites de servidumbre relativos a la propiedad intelectual privada, lo que genera una posible tensión con el art. 42 de la Constitución italiana¹⁸.

Dudas similares también son compartidas por el Parlamento Europeo, que ha reiterado la necesidad de supervisar la eficacia continua de la legislación vigente sobre derechos de propiedad intelectual frente al desarrollo de la IA en el campo del procesamiento de contenidos creativos¹⁹.

III. ACCESO A LA INFORMACIÓN PÚBLICA Y LENGUAJE DE PROGRAMACIÓN

Hemos visto cómo la relación entre el derecho de acceso a los programas informáticos utilizados por la administración pública y la protección del titular del programa es aún incierta²⁰.

¹⁸ Véase S. Guizzardi, “La protezione d’autore dell’opera dell’ingegno creata dall’Intelligenza Artificiale”, en *Aida*, 2018, donde el autor señala que, con la excepción de algunos países, la mayoría de los ordenamientos jurídicos no prevén una disciplina expresa y sistemática del tema.

¹⁹ Resolución de 12 de febrero de 2019, 2018/2088(INI). Véase ampliamente M. Capparelli, cit., donde el autor argumenta que es posible incluir entre las obras intelectuales también las generadas de forma autónoma por la misma IA.

²⁰ Para un examen en profundidad de los límites del derecho de acceso a la información pública en la legislación española ver L. Cotino Hueso y A. Boix Palop, coords., *Los límites al derecho de acceso a la información pública*, Tirant lo Blanch, Valencia, 2020. Para los límites específicos del secreto industrial y de la propiedad intelectual ver en el mismo libro la contribución de M.E. Gutiérrez David, “El derecho de acceso a la información pública contractual y sus límites”, especialmente pp. 277-293.

Todavía no se sabe cuál será el futuro de este debate, pero el único hecho constante de la jurisprudencia italiana que ha tratado el caso²¹ está relacionado con la conclusión “de que el uso de procedimientos “robotizados” no puede ser motivo de elusión de los principios que conforman nuestro sistema y que regulan la realización de la actividad administrativa. En este sentido, “la caracterización multidisciplinar” del algoritmo (una construcción que ciertamente no requiere solo habilidades jurídicas, sino técnicas, informáticas, estadísticas, administrativas) no exime de la necesidad de que la “fórmula técnica”, que de hecho representa el algoritmo, vaya acompañada de explicaciones que lo traduzcan en la “regla jurídica” subyacente y haciéndola legible y comprensible. Con las consecuencias ya identificadas en términos de conocimiento y cuestionabilidad”²².

La ya mencionada referencia al lenguaje surge como un momento de contacto entre jurista, ciudadano y administración. Sobre el hecho de que los lenguajes de alto nivel (por ejemplo, COBOL o C utilizadas en el procedimiento en cuestión) o el lenguaje de máquina puedan ser

²¹ Todavía no se puede hablar de un derecho vivo sobre los procedimientos administrativos informáticos, basta pensar que las dos sentencias del Consiglio di Stato que se han ocupado de la materia parten de premisas lógico-jurídicas muy diferentes. En particular, mientras que la Sent. n°2270 de 2019, llega a calificar el propio algoritmo como acto administrativo, sigue un enfoque totalmente diferente la citada Sent. n° 8472. En efecto, esta última sentencia acogiendo una reconstrucción autorizada de la doctrina ha entendido el procedimiento automatizado en clave estrictamente de instrucción; en este sentido se habían expresado M.C. Cavallaro y G. Smorto, “Decisión pública y responsabilidad de la administración en la sociedad algorítmica”, *Federalismi.it*, 04 Septiembre 2019. Para una reconstrucción profunda del debate interno sobre el tema ver A.G. Orofino, “The Implementation of the Transparency Principle in the Development of Electronic Administration”, *European Review of Digital Administration & Law-Erdal*, Volume 1, Issue 1-2, 2020, pp. 123-142.

Se trata, sin embargo, de una cuestión presente también en el contexto español cfr. para profundizar C. Velasco Rico, “Dossier sobre l’Administració a l’era digital”, *Revista catalana de dret públic*, núm. 58, 2019, pp. 208-230; J. Ponce, “Inteligencia artificial, derecho administrativo y reserva de humanidad: algoritmos y procedimiento administrativo debido tecnológico”, *Revista General de Derecho Administrativo*, núm. 50, 2019.

²² Así de acuerdo con las anteriormente citadas Sentencias del Consiglio di Stato n°2270 y n°8472.

examinados seriamente por las competencias de la administración o del juez, es legítimo tener serias dudas.

Por su parte, la jurisprudencia ha aceptado por último que este tipo de procedimientos automatizados también pueden aplicarse a actividades discrecionales, y también recurriendo a formas avanzadas de aprendizaje automático, siempre en cumplimiento de los requisitos de transparencia mencionados²³.

También a la luz de lo que se ha examinado con respecto a las capacidades de aprendizaje automático, es útil reiterar aquí que este mecanismo no siempre es transparente. En particular, como se ha destacado en la doctrina²⁴, mientras que los métodos de aprendizaje basados en sistemas expertos (o los llamados basados en el conocimiento) permiten a través de la reconstrucción hacia atrás reanudar los pasos lógicos que se han llevado a cabo, no se puede decir lo mismo de las redes neuronales más recientes. Además, ante la imposibilidad de la administración de proporcionar los pasos lógicos que se han transpuesto al algoritmo “Buona Scuola”, nos parece que pone de manifiesto todos los índices de criticidad típicos de la implementación de sistemas de IA aún más complejos dentro de las instituciones públicas.

Estos límites se refieren a la misma posibilidad de traducir el ordenamiento jurídico, en este caso las normas de la Ley n°104 de 2015 y las de negociación colectiva, al lenguaje formal de un algoritmo funcional y operativo²⁵.

De la misma manera, no parece haber habido conciencia de los funcionarios sobre los límites y el funcionamiento de la herramienta informática. Además, el acceso a la información del procedimiento y su funcionamiento por parte del ciudadano no fue garantizado desde la etapa procedimental como exige la legislación sobre actividad administrativa²⁶. Aún más relevante es el hecho de que, dados los

²³ Cfr. nota 22.

²⁴ G. Sartor, *L'informatica giuridica e le tecnologie dell'informazione*, Corso di *informatica giuridica*, Torino, Giappichelli, 2016, p. 304.

²⁵ I. Trujillo, “Linguaggio: Il discorso giuridico tra linguaggio e azioni”, en A. Andronico, T. Greco, y F. Macioce, *Dimensioni del diritto*, Torino, Giappichelli, 2019, p. 276.

²⁶ Crf. Párrafo 4

resultados alcanzados por el *software* en el caso que examino, los mecanismos de resolución de errores que presentan las tecnologías de la información están lejos de demostrarse.

Por lo tanto, las dudas sobre la implementación de la IA en el ámbito del derecho, y del derecho administrativo en particular, están lejos de estar apagadas.

Sobre cómo puede garantizar la debida transparencia en estos casos es un tema central de este trabajo al que nos dedicaremos más ampliamente en los siguientes apartados.

IV. LA TRANSPARENCIA COMO PRINCIPIO FUNDAMENTAL DE LA ACCIÓN ADMINISTRATIVA

El principio de transparencia se ha reconducido en el debate interpretativo italiano al principio fundamental de legalidad de la administración pública²⁷.

Son reflexiones que han sido llevadas a la madurez incluso en la jurisprudencia constitucional más reciente. En este sentido, la Corte costituzionale en la sentencia n°69 de 2018 aclaró que:

“Es en el foro procedimental, por lo tanto, donde puede y debe tener lugar la apreciación sincrónica de los intereses públicos implicados y dignos de protección, tanto en comparación con el interés del operador económico privado como de nuevo (y no menos importante) con los intereses adicionales de los que los ciudadanos individuales y las comunidades, y que encuentran su disposición y protección en los principios constitucionales. La estructura del procedimiento administrativo, de hecho, hace posible que surjan estos intereses, su perspectiva adecuada, así como la publicidad y transparencia de su evaluación, en aplicación de los principios mencionados en el art. 1 de la Ley n°241 de 7 de agosto de 1990: eficacia, imparcialidad, publicidad y transparencia. De este modo se garantiza, en

²⁷ Para la tesis del derecho de acceso a la información pública como principio fundamental en el contexto español, véase L. Cotino Hueso, “Los derechos fundamentales de acceso a la información pública, sus límites y la supletoriedad de la legislación general de transparencia”, en L. Cotino Hueso y A. Boix Palop, coords., *Los límites al derecho de acceso...*, cit., p. 21 y ss.; el autor vincula este derecho a disposiciones precisas de la Constitución española, como el art. 20 en materia de libertad de información, así como al derecho a la participación política de los mencionados en el art. 23, es decir, a la libertad sindical del artículo 28.

primer lugar, la imparcialidad de la elección, al igual que el art. 97 Cost., pero también la búsqueda, de la manera más adecuada y eficaz, del interés primario, en aplicación del principio de buen funcionamiento de la administración, contemplado en el mismo art. 97 Cost.

En definitiva, de este modo se garantiza el respeto del principio de lealtad también deducible del art. 97 Cost. en sentido no solo formal, como atribución normativa del poder, sino también sustancial, como ejercicio del poder de manera coherente con la fuente normativa de atribución. En efecto, al cierre del sistema, existe la posibilidad de someter las decisiones tomadas y sus modalidades de adopción al escrutinio jurisdiccional”.

A este respecto, por tanto, la Ley n° 241/1990 representó un cambio disruptivo, a través de la subversión de la anterior regla general del secreto de los actos en la administración pública con el derecho general de acceso garantizado al interesado²⁸. Más en detalle, como se reconstruye cuidadosamente en doctrina con la entrada en vigor de la citada ley “el legislador ha optado por enuclear, entre los principios generales de actuación administrativa a que se refiere el art. 1 de la misma Ley de procedimiento, el principio de publicidad que luego fue acompañado por la reforma de 2005 de la transparencia”²⁹.

Estas reflexiones son útiles aquí para resaltar aún más el vínculo entre el principio de transparencia y los otros principios constitucionales de la acción administrativa. La entrada en vigor de la Constitución, de hecho, eleva la transparencia a un principio general de acción de los poderes públicos³⁰. En este orden de razones, “el principio de transparencia se consideró crucial para asegurar el correcto ejercicio del poder público, con el fin de garantizar una dimensión más participativa, así como imparcial”³¹.

²⁸ Ver para una reconstrucción detallada A.G. Orofino, *Profili giuridici della trasparenza amministrativa*, Bari, Caucci editore, 2013; M.C. Cavallaro, “Garanzie della trasparenza amministrativa e tutela dei privati”, *Dir. amm.*, núm. 1, 2015, p. 121 y ss.

²⁹ M.C. Cavallaro, *cit.*, en particular la autora señala las cuestiones interpretativas críticas de distinguir el concepto de publicidad del de transparencia.

³⁰ En este sentido V. Fanti, “La trasparenza amministrativa tra principi costituzionali e valori dell’ordinamento europeo: a margine di una recente sentenza della Corte costituzionale (n. 20/2019)”, *federalismi.it*, 4 de marzo de 2020.

³¹ M.C. Cavallaro, *cit.*

Más concretamente, se considera que la plena y total accesibilidad y usabilidad de los datos e información de que dispone una administración es capaz de lograr una participación activa y colaborativa de los ciudadanos respecto a la actuación administrativa. En este sentido, la participación se deriva del conocimiento y uso de la información en poder de la administración, convirtiéndose en “funcional al establecimiento de un proceso virtuoso, de tal manera que la transparencia asegure un diálogo y una comparación constante entre el ciudadano y la administración”³².

Más recientemente, a través del Decreto legislativo n°33 de 2013, los institutos jurídicos de los que se compone el principio de transparencia han sido votados para la función de una barrera a los fenómenos de corrupción, vinculando así este asunto aún más firmemente con el principio constitucional de buen desempeño de la a. p.

Este último es un tema ampliamente probado en la doctrina, donde se ha reconocido el valor de la transparencia como “un papel fundamental en la transición del Estado absoluto al Estado democrático, permitiendo una nueva dimensión, más democrática e inclusiva de las relaciones entre la a. p. y los ciudadanos, y que hoy puede representar, a través del equilibrio con el principio de buen desempeño de la administración pública, una herramienta eficaz para combatir la corrupción”³³.

Como ha destacado el Consiglio di Stato “la transparencia puede convertirse así en un “punto de confluencia” de los principios jurídicos constitucionalmente establecidos de actuación administrativa (desde el buen desempeño hasta la imparcialidad, pasando por el respeto a los llamados principios de legalidad sustancial, al método de participación democrática), del que derivan institutos jurídicos, de tipo transversal, que pueden considerarse dirigidas a lograr la transparencia”³⁴.

³² *Ibidem*.

³³ V. Fanti, “La pubblicità e la trasparenza amministrativa in funzione del contrasto alla corruzione: una breve riflessione in attesa del legislatore delegato”, *Giustamm*, núm. 3, 2016.

³⁴ Reunión de la “Sezione Consultiva per gli atti normativi” del Consiglio di Stato, de 18 de febrero de 2016, n° 00515/2016.

Se trata, por tanto, de instrumentos normativos que prescriben la “visibilidad del poder”, y como veremos mejor en el transcurso de la discusión representan elementos instrumentales esenciales para el ejercicio efectivo de las garantías participativas³⁵.

Sobre la utilización de las nuevas tecnologías, bajo una perspectiva positiva, se han leído las posibilidades de utilizar aplicaciones de inteligencia artificial para descifrar los datos en posesión de la administración con el fin de prevenir posibles fenómenos corruptos, contribuyendo así al buen funcionamiento de la a. p³⁶. Sin embargo, en este escenario, como se ha puesto de manifiesto en la doctrina, se trata de hacer transparentes y controlables los indicadores que subyacen al sistema automatizado de detección del riesgo de corrupción³⁷.

También hay que valorar positivamente las posibilidades de la IA de garantizar un acceso funcional a los datos de la administración pública, acceso que se hace instrumental a una participación efectiva. En este sentido, se ha puesto de manifiesto que las tecnologías de la IA permiten transformar las masas de datos en conocimiento difundido y compartido, de manera que la P.A. sea transparente para los ciudadanos y, sobre todo, para sí misma, garantizando a los ciudadanos y a los administradores no sólo el acceso semántico a la información y la interoperabilidad de los procesos, sino una mejor comprensión de la relación entre el Estado y el ciudadano³⁸.

Si se observa bien, una parte considerable de la literatura considera que, en la actualidad, estos son los principales resultados de la democracia digital, o una mejora significativa del acceso a la in-

³⁵ Ese es el “*Gobierno Público en Público*” usando otra definición prestigiosa de N. Bobbio para describir la democracia en *Il futuro della democrazia*, Torino, Einaudi, 1995, p. 86.

³⁶ Ver en este sentido la Resolución del Parlamento Europeo, de 14 de marzo de 2017, sobre las implicaciones de los macrodatos para los derechos fundamentales: privacidad, protección de datos, no discriminación, seguridad y aplicación de la ley (2016/2225(INI).

³⁷ J. Ponce, “La prevención de riesgos de mala administración y corrupción, la inteligencia artificial y el derecho a una buena administración”, *Encuentros multidisciplinarios*, vol. 22, núm. 65, 2020.

³⁸ AGID, *Libro Bianco sull'Intelligenza Artificiale al servizio del cittadino*, 2018.

formación pertinente con fines públicos y del intercambio de dicha información³⁹.

En este momento, sin embargo, en nuestro sistema esta apertura parece estar orientada principalmente hacia dentro, es decir, en las relaciones entre las diferentes administraciones. Un claro ejemplo de este enfoque es la Plataforma Nacional de Datos Digitales (PDND) prevista en el art. 50-ter del CAD, y dirigida a promover el conocimiento y uso de los activos de información poseídos, con fines institucionales, lo que de hecho limita el acceso solo a sujetos específicos previamente habilitados.

Esta tendencia restrictiva tiene una explicación válida si tenemos en cuenta que en la administración pública italiana existe una fuerte noción de propiedad de los datos, faltan aquellos requisitos de disponibilidad total e incondicional de información en la base de las aplicaciones tecnológicas que permiten la comprensibilidad de los datos externamente⁴⁰.

Desde un punto de vista problemático, se destaca la cuestión adicional de la buena administración, donde la explicabilidad o transparencia del algoritmo asume una importancia decisiva, condición que corre el riesgo de eclipsar masivamente los beneficios producidos. En otras palabras, el acceso a la lógica de los procesos públicos de toma de decisiones puede decirse que es un derecho instrumental para el ejercicio de todos los derechos políticos del individuo. El uso de *software* reconducibles a la IA corre el riesgo de socavar este ejercicio en su raíz, ya que en la actualidad sus métodos de operación no son comprensibles. Además, incluso si se garantizara la transparencia formal del código fuente, sigue siendo su escaso conocimiento por parte del ciudadano común del lenguaje de programación. Por lo tanto, se puede decir que la oscuridad de algunos de los llamados algoritmos

³⁹ J. van Dijk, "Digital democracy: Vision and reality. Innovation and the Public Sector", *Public Administration in the Information Age*, Amsterdam, IOS Press, 2012. El autor señala también que la participación digital se limita en gran medida a las etapas tempranas y finales del procedimiento deliberativo y que rara vez da acceso a las etapas centrales del proceso de toma de decisiones y de ejecución.

⁴⁰ G. Vetere, *Accesso ai dati della PA, l'utilità dell'intelligenza artificiale*, en www.agendadigitale.eu, 10 octubre 2018.

inteligentes es ontológica, mientras que para las formas de IA clásica siguen siendo los límites externos derivados de la comprensibilidad del lenguaje de programación⁴¹.

Como destacan los mismos jueces administrativos, la luz de la transparencia fertiliza la semilla del conocimiento entre los ciudadanos y contribuye, por un lado, al buen funcionamiento de la administración pública, pero, por otro, también a la satisfacción de los derechos fundamentales de la persona si es cierto que la organización administrativa y los derechos administrativos están estrechamente interrelacionados⁴².

Las mismas conclusiones fueron alcanzadas por el juez constitucional, quien tuvo la oportunidad de precisar que “los principios de publicidad y transparencia, deben ser referidos no sólo, como corolarios del principio democrático (art. 1 Cost.), a todos los aspectos relevantes de la vida pública e institucional, pero también, de conformidad con el art. 97 Cost., para el buen funcionamiento de la administración” (Corte costituzionale, Sentencia n°20 de 2019)⁴³.

⁴¹ Véase G. Lo Sapio, “La trasparenza sul banco di prova dei modelli algoritmici”, *federalismi.it.it*, 21 de abril de 2021, donde la autora distingue tres niveles de opacidad “a) todos los algoritmos se caracterizan ciertamente por el “opacidad lingüística”, debido a que las instrucciones son comunicadas por el programador a la máquina en un lenguaje informático, y no en el lenguaje natural con el que se encuentran expresas las normas jurídicas; b) muchos de ellos tienen un nivel adicional de opacidad, que puede definirse como “legal”, ya que la innovación tecnológica basada en algoritmos, especialmente si es adquirida por la administración en el mercado, puede ser objeto de derechos de propiedad intelectual, o derechos industriales reconocidos por sus “creadores”, conectados a la dinámica de explotación comercial sobre el secreto de los códigos fuente (con la consiguiente necesidad necesaria de equilibrar el interés cognitivo y la protección. de esos derechos); c) para una categoría específica de algoritmos, pertenecientes a una subárea específica de la inteligencia artificial, y en particular, para los de *machine learning* y de *deep learning*, una forma de opacidad “estructural” que deriva del mismo mecanismo de funcionamiento del sistema, quedando oscuro e impenetrable incluso para que los propios programadores entiendan cómo la máquina, a partir de los datos proporcionados, ha llegado a un determinado resultado (es el fenómeno conocido como “caja negra”)”.

⁴² Así recientemente el Consiglio di Stato, Adunanza Plenaria, Sentencia n°10 de 2020.

⁴³ La sentencia aborda el delicado equilibrio entre transparencia y privacidad en cuanto a la publicación de datos relativos a gestores públicos. Emblemáticas son

Además, el derecho a que la decisión se adopte sobre la base de una motivación comprensible puede derivarse directamente, para los actos jurisprudenciales, de la misma obligación de motivación de las sentencias previstas en el art. 111 de la Constitución italiana⁴⁴.

Consideraciones similares pueden llevarse a cabo valorando la relación con los derechos de acceso y participación en la administración pública, aceptando la perspectiva aquí defendida para la cual tales valores encuentran hoy un fundamento directo en la Constitución, a través de la lectura sistemática de sus disposiciones⁴⁵.

Que el acceso y la transparencia están orientados a garantizar, a través del acceso a la información, el derecho a una buena administración fue apoyado recientemente por el Consiglio di Stato en la conocida Sentencia n° 10 de la Adunanza Plenaria 2020 donde se destacó que el carácter fundamental del derecho de acceso cívico generalizado es invencible de los artículos 1, 2, 97 y 117 de la Constitución⁴⁶.

las consideraciones realizadas en la exposición de motivos donde se destaca que “los derechos a la confidencialidad y a la transparencia se afrontan sobre todo en el nuevo escenario digital: un ámbito en el que, por un lado, los derechos personales pueden verse amenazados por la circulación indiscriminada de información y, por otro, precisamente la circulación más amplia de datos puede permitir a todos informar y comunicarse mejor”. Para un comentario en profundidad sobre la sentencia, véase V. Fanti, *La trasparenza amministrativa tra principi costituzionali*, cit.

⁴⁴ En el art. 120.3 de la Constitución Española que establece que: “Las sentencias serán siempre motivadas y se pronunciarán en audiencia pública”. Ver sobre estos aspectos A. Cerrillo i Martínez. Y C. Velasco Rico, “Jurisdicción, algoritmos e inteligencia artificial”, en F. López Ramón y J. Valero Torrijos, eds., *20 años de la Ley de lo Contencioso-administrativo, actas del XIV Congreso de la Asociación Española de Profesores de Derecho Administrativo*. Murcia, 8-9 de febrero de 2019, Madrid, Instituto Nacional de Administración Pública, 2019, pp. 291–303; S. De la Sierra Morón, “Inteligencia artificial y justicia administrativa: una aproximación desde la teoría del control de la Administración Pública”, *Revista General de Derecho Administrativo*, núm. 53, 2020.

⁴⁵ Cfr. G. Pastori, “Amministrazione e procedimento”, en *Il cittadino e la pubblica amministrazione, Giornate di Studi in onore di Guido Corso*, Palermo 12 e 13 dicembre 2014, Napoli, Editoriale Scientifica, 2016, p. 24 y ss.

⁴⁶ Otras referencias se encuentran en el art. 42 de la CDF y art. 10 del CEDH.

V. ACCESO A LA INFORMACIÓN DE LA ADMINISTRACIÓN PÚBLICA Y PROCEDIMIENTO ALGORÍTMICO: DESDE LA LEY N°241 DE 1990 HASTA EL DECRETO LEGISLATIVO N°97 DE 2016

Habiendo enmarcado el principio de transparencia en la esfera de los principios fundamentales, ahora es necesario verificar su extensión en la disciplina con detalle.

En este sentido, ya hemos mencionado que la ley general sobre el procedimiento administrativo (241/1990) ha supuesto un cambio de época en la actuación administrativa, subvirtiendo el principio de secreto con el de accesibilidad de la información de las administraciones públicas.

El artículo 22 de la citada ley, en su versión modificada por el artículo 10 de la Ley n°69 de 18 de junio de 2009, establece que el acceso es un principio general de la actividad administrativa “con el fin de fomentar la participación y garantizar su imparcialidad y transparencia”⁴⁷. Por lo tanto, como se señala en la literatura jurídica, la transparencia requiere, por un lado, el cumplimiento de las normas

⁴⁷ Puede ser útil recordar el texto del acuerdo:

“(Artículo 22) A los efectos del presente capítulo, se entenderá por: a) “derecho de acceso”: el derecho de los interesados a inspeccionar y extraer copias de documentos administrativos; b) “interesados”: todas las entidades privadas, incluidas aquellas con intereses públicos o generalizados, que tengan un interés directo, concreto y actual, correspondiente a una situación legalmente protegida y vinculado al documento al que se solicita el acceso; c) “partes contrarias a los interesados”: todas las personas, identificadas o fácilmente identificables sobre la base de la naturaleza del documento solicitado, que se verían comprometidas por el ejercicio del acceso a su derecho a la confidencialidad; d) “documento administrativo”: toda representación gráfica, fotocinematográfica, electromagnética o de cualquier otra índole del contenido de los actos, incluidos los internos o no relativos a un procedimiento específico, en poder de una administración pública y relativos a actividades de interés público, con independencia del carácter público o privado de su disciplina sustantiva; e) “administración pública”: todas las personas de Derecho público y las personas de Derecho privado limitadas a su actividad de interés público regida por el Derecho nacional o comunitario.

2. El acceso a los documentos administrativos, habida cuenta de sus importantes fines de interés público, constituirá un principio general de la actividad administrativa con el fin de fomentar la participación y garantizar su imparcialidad y transparencia.

procedimentales y, por otro, la comprensibilidad de la acción administrativa por parte del administrado⁴⁸.

Conscientes de los límites ontológicos de la opacidad que se derivan de la aplicación de sofisticados procedimientos algorítmicos, no podemos sino adherirnos a esa necesidad de mayor transparencia adoptada por la jurisprudencia, que implica el pleno conocimiento de una norma expresada en un lenguaje distinto al jurídico.

Con el fin de comprobar que los criterios, los supuestos y los resultados del procedimiento robotizado se ajusten a las prescripciones y a las finalidades establecidas por la ley o por la propia administración en la fase anterior a dicho procedimiento y a fin de que sean claras —y por consiguiente susceptibles de ser controladas— las modalidades y las normas con arreglo a las que algoritmo se ha establecido⁴⁹. Esto se vuelve aún más esencial también para la parte pública, en un contexto en el que se afirma la tesis de que “una vez que la administración decide remitir las determinaciones a un algoritmo, limita el ejercicio residual de la discrecionalidad administrativa de las administraciones competentes individuales, en la aplicación del algoritmo, dentro de un alcance muy limitado”⁵⁰.

3. Todos los documentos administrativos serán accesibles, con excepción de los contemplados en el artículo 24, apartados 1, 2, 3, 5 y 6.

4. La información en poder de una administración pública que no adopte la forma de un documento administrativo no es accesible, salvo lo dispuesto en el Decreto Legislativo n°196 de 30 de junio de 2003 sobre el acceso a los datos personales por parte de la persona a la que se refieren los datos.

5. La adquisición de documentos administrativos por parte de entidades públicas, cuando no estén comprendidas en el párrafo 2 del artículo 43 del texto refundido de las leyes y reglamentos sobre documentación administrativa, a que se refiere el Decreto del Presidente de la República de 28 de diciembre de 2000, n. 445, se basa en el principio de leal cooperación institucional.

6. El derecho de acceso será ejercitable mientras la administración pública esté obligada a poseer los documentos administrativos a los que se solicita el acceso”.

⁴⁸ A. Simonati, “La ricerca in materia di trasparenza amministrativa: stato dell’arte e prospettive future”, *Dir. amm.*, núm. 2, 2018, p. 311 y ss.

⁴⁹ En este sentido Consiglio di Stato, Sentencia n°8472 citada.

⁵⁰ En este sentido Consiglio di Stato, Sezione IV, Sentencia n°2927 de 2021, sobre la posibilidad de que la a. p. (es decir, la región de Lombardía) se desvíe del resultado generado por el algoritmo utilizado.

Como hemos señalado, la ley general de procedimiento contiene en sí misma la semilla de la transparencia, habiendo invertido la relación norma-excepción entre el secreto y el acceso a la actividad administrativa. Más detalladamente, se han sometido a este principio los institutos de la motivación de la medida administrativa (art. 3), de la comunicación de incoación del procedimiento (art. 7) y las consiguientes reivindicaciones participativas (artículos 8, 9, 10, incluida la presentación de escritos y documentos y la accesibilidad de los documentos dentro del propio procedimiento), del responsable del procedimiento como interlocutor del ciudadano (art. 5 y 6)⁵¹. Por lo tanto, como hemos reiterado en repetidas ocasiones, la transparencia de la administración pública puede considerarse un requisito previo sustancial para garantizar la participación procedimental efectiva de la parte interesada⁵².

El límite principal de la Ley n°241 de 1990, sin embargo, se encontró en la disciplina de acceso a los actos administrativos según lo regulado por su art. 22. Más concretamente, la Ley ha seleccionado las exigencias del denominado acceso defensivo, identificándolas en inmediatez, concreción y actualidad (artículo 22, apartado 1, letra d), de manera que el juicio sobre el interés que legítima se base en parámetros fijos, rígidos y predeterminados en cuanto a su contenido objetivo⁵³.

Se trata de requisitos que han sido mejor desentrañados en la jurisprudencia donde se ha aclarado que “la “correspondencia” circunscribe exactamente el interés en el acceso a los documentos en el sentido de “correspondiente a una situación jurídicamente protegida”. El único interés que legitima el acceso defensivo será el que corresponda de forma directa, concreta y actual al tratamiento o incluso defensa en juicio de tales casos predeterminados, en clave estrictamente defensiva. Este último aspecto, más concretamente, se aclara con el segundo de los parámetros a los que se ha hecho referencia, a saber, el relativo a la denominada “conexión”. El legislador ha circunscrito ulterior-

⁵¹ En este sentido M.C. Cavallaro, *cit.*, p. 121 y ss.

⁵² Ver para todo M. Bombardelli, *Trasparenza amministrativa*, in *Diritto on line*, Roma, Treccani, Istituto della Enciclopedia italiana, 2015.

⁵³ En este sentido y para las extensas referencias doctrinales y jurisprudenciales, véase A.G. Orofino, *Profili giuridici della...*, *cit.*, p. 9 y ss.

mente el objeto de la situación legitimadora el acceso defensivo, exigiendo que la misma, además de corresponder al contenido del abstracto paradigma legal, también esté vinculada al documento al que se solicita el acceso (art. 24, apartado 7, de la n°241 de 1990) de manera que se ponga de manifiesto de manera directa e inequívoca el nexo de instrumentalidad que afecta la situación subjetiva final al documento cuya exhibición se solicita, y para cuya obtención el acceso defensivo, en cuanto situación instrumental, actúa de intermediario”⁵⁴.

Estos requisitos se resumen, a nivel de procedimiento, en el siguiente art. 25, apartado 2, de la Ley n°241 de 1990, según la cual “la solicitud de acceso a los documentos debe estar motivada”.

En cuanto al objeto del derecho, éste está fuertemente circunscrito por el legislador considerando que el art. 22, apartado 1, letra (d)⁵⁵ dicta un concepto de documento administrativo que, por muy amplio que sea, excluye ciertamente una parte del patrimonio cognoscitivo de la administración⁵⁶.

Por estas razones, en el debate interno y supranacional se ha manifestado durante mucho tiempo una cierta impaciencia por este enfoque normativo tan restrictivo, una limitación corroborada por una

⁵⁴ Consiglio di Stato, Adunanza Plenaria, n° de 2021. La sentencia es significativa porque intervenir en animados debates jurisprudenciales finalmente ha dictaminado que “la administración pública titular del documento y el juez administrativo que conozca del procedimiento de acceso de conformidad con el art. 116 c.p.a. no debe, en cambio, efectuar ninguna otra apreciación *ex ante* de la admisibilidad, influencia o decisión del documento solicitado en ninguna sentencia dictada, ya que dicha apreciación es responsabilidad, en su caso, únicamente de la autoridad judicial que conoce del asunto y, desde luego, no de la administración pública titular del documento o del juez administrativo en la sentencia de acceso, salvo en el caso de una manifiesta, absoluta, falta de conexión entre el documento y las necesidades defensivas y, por lo tanto, en el caso de un ejercicio pretextual o imprudente del propio acceso defensivo debido a la ausencia radical de las condiciones legitimadoras previstas por la Ley n°241 de 1990”.

⁵⁵ ““documento administrativo”: cualquier representación gráfica, fotocinematográfica, electromagnética o de cualquier otra índole del contenido de los actos, incluidos los internos o no relacionados con un procedimiento específico, en poder de una administración pública y relativo a actividades de interés público, independientemente del carácter público o privado de su disciplina sustantiva”.

⁵⁶ De ahí la definición generalizada en doctrina y jurisprudencia italiana de “acceso documental”.

actitud generalizada de reticencia por parte de las individuales administraciones públicas⁵⁷.

Esto justificó la promulgación del Decreto legislativo n°33 de 14 de marzo de 2013⁵⁸ y su modificación con el Decreto legislativo n°97 de 25 de mayo de 2016, una articulación normativa comparable a lo que en los sistemas anglosajones se denomina Freedom of Information Act (FOIA)⁵⁹.

La nueva disciplina mejora la transparencia, legitimando formas generalizadas de control por parte de los ciudadanos, adaptándose a los estándares internacionales. De esta manera se fortalece la relación participativa entre ciudadanos e instituciones, combinando a la perfección “la contribución cognitiva de todos” con el objetivo de contrastar los fenómenos de corrupción “a favor de unos pocos”.

En la medida en que es de mayor interés aquí, esta normativa interfiere con la opacidad algorítmica del procedimiento automatizado, ya que amplía el concepto esbozado anteriormente de transparencia, definida ahora como “acceso total a los datos y documentos en poder de las administraciones públicas con el fin de proteger los derechos de los ciudadanos, promover la participación de los interesados en la actividad administrativa y favorecer formas generalizadas de control de la realización de las funciones institucionales y de la utilización de los recursos públicos” (art. 1)⁶⁰.

En este sentido, el acceso defensivo al documento va acompañado de un acceso cívico generalizado⁶¹.

⁵⁷ En este sentido y para profundizar ver R. Chieppa y R. Giovagnoli, *Manuale di diritto amministrativo*, Milano, Giuffrè, 2018, p. 703 y ss.

⁵⁸ El Decreto implementa la delegación contenida en la ley de 6 de noviembre de 2012, sobre medidas para combatir la corrupción.

⁵⁹ Ver recientemente M. Savino, “Il FOIA italiano e i suoi critici: per un dibattito scientifico meno platonico”, *Diritto Amministrativo*, núm. 3, 2019, p. 453 y ss.

⁶⁰ Estos son aspectos que la doctrina ha comenzado a probar recientemente: ver G. Lo Sapio, *cit.*

⁶¹ Sobre las relaciones entre ambos institutos, véase más recientemente la sentencia de la Adunanza Plenaria n°10/2020 citada en la que se aclara que “el art. 5, del Decreto legislativo n°33 de 2013 admite claramente la competencia entre las diferentes formas de acceso, cuando precisa que las diferentes formas de acceso de los interesados previstas por el Capítulo V de la Ley n° 241 de 7 de agosto de

A diferencia del primero, en el nuevo instituto la solicitud de acceso puede ser propuesta por “cualquier persona”, es decir, el solicitante no tiene que demostrar en la solicitud que remita a la administración, ninguna conexión “cualificada” con los documentos, información y datos que pretende conocer (art. 5, apartado 2). El acceso cívico generalizado, de hecho, no está sujeto a ninguna limitación en cuanto a la legitimación subjetiva del solicitante, así como la solicitud no debe estar motivada (art. 5, apartado 3).

El aspecto más interesante para nuestros propósitos, sin embargo, se refiere al alcance objetivo del instituto. Una lectura global de las normas (en particular, los artículos 1 y 5) muestra que la administración, tras la presentación de la solicitud, tiene la obligación de proporcionar “documentos, información o datos”⁶².

En este caso normativo, la administración no puede limitarse a proporcionar una copia del documento, como lo exige la Ley 241/1990, sino que también debe proporcionar los “datos” y sobre todo la “información”. Esta disciplina, por lo tanto, legitima al individuo para solicitar acceso, además de los “datos” que se colocan en la base del *software* (en términos de entrada y salida), también a la “información” inherente a la lógica de funcionamiento del procedimiento algorítmico utilizado. En este sentido, creemos que del complejo normativo en materia de transparencia, surge un deber oportuno por parte de la administración pública de aportar los datos e información que preceden al procedimiento administrativo automatizado.

Se trata, por tanto, de un instituto jurídico que contribuye a realizar la necesidad, destacada en la jurisprudencia, de garantizar al principio de transparencia un valor reforzado, asegurando que el mecanismo a través del cual se realiza la decisión robótica (es decir, el algoritmo) sea “cognoscible”. Lo “que implica también el del pleno conocimiento de una norma expresada en una lengua distinta a la jurídica”⁶³.

1990 siguen siendo firmes, además del llamado acceso cívico simple (apartado 1) y el llamado generalizado (apartado 2).

⁶² Adicionales a los publicados en sitios web institucionales, en el apartado denominado “Amministrazione Trasparente”.

⁶³ Así Consiglio di Stato, Sent. n° 8472 de 2019 cit.

A falta de respuesta positiva de la administración pública en los términos de la ley (30 días), o en los casos en que ésta sea parcial e insatisfactoria, el particular podrá recurrir al rito específico relativo al acceso regulado por el art. 116 del Código de Proceso Administrativo⁶⁴, y se refiere a la competencia exclusiva del juez administrativo en virtud del artículo 133 del mismo texto.

En cuanto al potencial del *software* inteligente con respecto a este complejo regulatorio, son en gran medida similares a los ya destacados. Desde una perspectiva positiva, de hecho, la IA es capaz de ordenar la complejidad de los datos en poder de la administración pública para proporcionar al usuario información comprensible de la que a menudo carecen los sitios y documentos de los que hacen alarde las instituciones públicas. La principal limitación, sin embargo, sigue siendo superar el obstáculo relacionado con la comprensibilidad del procedimiento algorítmico, o el acceso a la información que preside la lógica subyacente al software utilizado.

En este último aspecto quedan abiertas las preguntas, que ya hemos abordado en los párrafos anteriores, relativas al equilibrio entre el derecho de acceso del ciudadano y los derechos de protección del *software*.

Aquí puede ser útil destacar cómo este equilibrio es muy diferente “en el caso del acceso previsto por la Ley n° 241 de 1990, donde la protección puede permitir un acceso más profundo a los datos relevantes (cfr. art. 24, apartado 7); en comparación con el caso del acceso generalizado, donde las necesidades de control generalizado del ciudadano permiten un acceso menos profundo (si es necesario, en relación con el funcionamiento de los límites), pero más extenso, teniendo en cuenta que el acceso en este caso implica, de hecho, una amplia cognoscibilidad (y difusión) de datos, documentos e información (ver artículo 5 bis, Decreto Legislativo n°33 de 2013)”⁶⁵.

Sin embargo, este sigue siendo un tema controvertido y destinado a adquirir una importancia cada vez mayor, debido al amplio uso del

⁶⁴ Decreto legislativo n°104 de 2 julio de 2010.

⁶⁵ Así el Consiglio di Stato, Adunanza Plenaria, n°10 de 2020.

mercado privado que caracteriza las opciones de las administraciones italianas con respecto a la contratación de aplicaciones ITC⁶⁶.

VI. CONCLUSIONES PROVISIONALES

Tratando de llegar a una síntesis final, por lo tanto, la disciplina detallada de la transparencia administrativa tiene fuertes características de innovación debido al momento en que se emitió. Sin embargo, sigue existiendo la dificultad de conciliar esas normas con la realidad profunda y rápidamente modificada de los procedimientos totalmente digitalizados. En este sentido, se puede suponer razonablemente que en ausencia de indicaciones regulatorias claras y precisas, la administración no es capaz, o más bien no posee las herramientas técnicas y regulatorias para proporcionar esos datos e información sobre el funcionamiento del proceso gestionado por *software*. En este escenario regulatorio, algunas respuestas deberían provenir de la reciente propuesta de Reglamento de la Comisión Europea COM(2021) 206 final, denominada Ley de Inteligencia Artificial.

En efecto, en el informe con el que la Comisión ha presentado este instrumento normativo, se prevé como principio que “Las obligaciones que exigen una mayor transparencia tampoco afectarán de manera desproporcionada al derecho a la protección de la propiedad intelectual (artículo 17, apartado 2), puesto que únicamente se aplicarán a la información mínima necesaria para que las personas ejerzan su derecho a una compensación efectiva y solo exigirán la transparencia necesaria hacia las autoridades de supervisión y las encargadas de la aplicación de la ley, conforme a sus respectivos mandatos. La información se divulgará siempre con arreglo a la legislación pertinente en la materia, entre la que se incluye la Directiva 2016/943 relativa a la protección de los conocimientos técnicos y la información empresarial no divulgados (secretos comerciales) contra su obtención, utilización y revelación ilícitas. Cuando las autoridades públicas y los orga-

⁶⁶ En este sentido, ver El informe Istat, *Pubblica amministrazione locale e ICT, 2018*, de 17 de abril de 2020, de donde emerge que las funciones informáticas se confían en gran medida a proveedores privados (94,2%), mientras que hay una disminución en el uso de *software* de código abierto.

nismos notificados deban tener acceso a información confidencial o al código fuente para examinar el cumplimiento de las obligaciones sustanciales, tendrán que cumplir obligaciones de confidencialidad vinculantes”⁶⁷.

Por lo tanto, esperamos con interés que estas indicaciones iniciales se traduzcan en disposiciones vinculantes para los Estados miembros⁶⁸.

⁶⁷ <https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:52021PC0206>

⁶⁸ De hecho, las indicaciones más significativas parecen proceder por el art. 70 de la propuesta de Reglamento, que orienta el equilibrio entre la propiedad intelectual y las obligaciones de transparencia hacia el primer polo. El resultado de esta evaluación comparativa entre los intereses opuestos sin embargo, aparece en esta norma muy matizada y en todo caso devuelto en gran parte a las autoridades nacionales.

Bloque III
Transparencia y explicabilidad
en la empresa, la contratación y
prestación de servicios de internet y
en el ámbito laboral

Explicabilidad e inteligencia artificial responsable. Recomendaciones “aterrizadas” al día a día de las empresas

JUAN MANUEL BELLOTO GÓMEZ
Director de GuIA
OdiseIA

I. A MODO DE INTRODUCCIÓN

Cuando una tecnología es asimilada tras años de recorrido, los tecnólogos se convierten en meros transcriptoros de las necesidades empresariales, ya que las empresas saben lo que dicha tecnología puede ofrecer, y los perfiles tecnológicos no necesitan explicar nuevas grandes capacidades de la misma.

Este contexto cambia de manera importante cuando aparecen nuevas tecnologías como la inteligencia artificial, ya que provoca la necesidad de una comunicación fluida y constante entre perfiles de negocio y tecnología. Esta comunicación es fundamental para entender rápidamente la manera en la que aprovechar las capacidades de dichas nuevas tecnologías y sus limitaciones, ya que solo así se consiguen soluciones tecnológicas que aporten valor a los procesos de negocio de las empresas y un buen *time to market* que proporcione una ventaja competitiva de mercado. Los perfiles tecnológicos y de negocio evolucionan así su relación de cliente-proveedor a una relación de *partnership*, trabajando de manera mucho más estrecha.

Además, desde hace tiempo, no solo son necesarios perfiles de negocio y tecnología. La eclosión en las dos últimas décadas de tecnologías digitales donde la experiencia de usuario es fundamental (*Web*, *Mobile*, *Conversational*, realidades extendidas, etc.) ha provocado que a los perfiles de negocio y tecnológicos se uniesen nuevos compañeros. A su mesa de trabajo se unieron perfiles necesarios para la conceptualización y creación de estas nuevas soluciones digitales: diseñadores, sociólogos, psicólogos, lingüistas, etc. La mesa es por tanto cada vez es más diversa.

Con la inteligencia artificial, a todos estos perfiles se ha unido al menos un perfil más: el jurista. En este contexto la multidisciplinariedad es cada vez más necesaria y todos tienen que saber al menos un poco de la disciplina de su compañero. Y, lo más importante, estos perfiles tan diversos necesitan llegar a un lenguaje común comprensible por todos ellos, dejando de lado cuando interactúan los tecnicismos necesarios para cuando se comunican con sus compañeros de disciplina.

II. EXPLICABILIDAD E INTELIGENCIA ARTIFICIAL RESPONSABLE

1. El porqué de la necesidad de una IA responsable, ética y normativa

Tras la eclosión de internet a finales de los 90, hemos adoptado tecnologías que han cambiado por completo nuestra forma de entender empresa y sociedad. Pero la inteligencia artificial es una tecnología diferencial respecto de esas que hemos asimilado en las últimas décadas porque tiene la capacidad de aprender por sí misma, hacer predicciones, puede tomar decisiones, e incluso emular capacidades cognitivas del ser humano.

Por primera vez tenemos a nuestra disposición una tecnología que tiene similitudes (que no equivalencias directas, no de momento) con las capacidades del ser humano. Es por tanto natural pensar que para su correcta adopción sea necesario tener en cuenta los principios éticos (por las implicaciones de sus capacidades) como aspectos regulatorios (por los riesgos que puedan derivarse de su uso). De igual manera que dichos principios éticos y aspectos regulatorios son necesarios en las empresas y en cualquier sociedad de bienestar.

Cuando hablamos de competencias que se requieren a las personas en su desempeño profesional nos referimos por ejemplo, y entre otras, a:

- Que su actividad no esté sujeta a estereotipos sociales, y que no realice ningún tipo de discriminación ya sea por raza, género, edad, lugar de nacimiento, pertenencia a alguna minoría, o cualquier otro concepto.

- Que sean seguros, que tengan criterio para no ser engañados en su aprendizaje y actividad diaria manteniendo así su fiabilidad.
- Que puedan asegurar la privacidad de los datos que manejan.
- Por supuesto, que sean transparentes, y que puedan explicar sus razonamientos a la hora, por ejemplo, de hacer una predicción o tomar una decisión.

Esas competencias tienen un trasfondo ético y normativo. Y si volvemos a recoger esa similitud entre las capacidades de las personas y las de la IA, es natural pensar que estas competencias anteriormente expuestas y que se piden a las personas en su desempeño profesional deban ser igualmente aplicadas a la IA.

La ética por tanto es necesaria en la IA de igual manera que lo es en las personas. Las empresas, al igual que tienen sus políticas de comportamiento para sus empleados, tienen la responsabilidad de que sus soluciones inteligentes estén desarrolladas bajo principios éticos como la equidad, la privacidad y la explicabilidad, entre otros.

Además, al igual que ocurre en la sociedad, la ética es condición necesaria pero no suficiente. Es necesario acompañarla de legislación. Y por primera vez desde la revolución digital en la que nos hallamos inmersos en este siglo (no olvidemos, por ejemplo, los 20 años que pasaron tras la eclosión de internet hasta la aparición de la GDPR para gestionar la Privacidad), la legislación no se está dejando esperar y se está desarrollando prácticamente a la par del desarrollo tecnológico de la IA. Este es un claro síntoma de la importancia de una IA responsable, sujeta a ética y normativa.

2. La explicabilidad. Contexto y su relación con la IA responsable

Anteriormente explicaba como a las personas se nos pide en nuestro desempeño profesional que seamos transparentes y que podamos explicar nuestros razonamientos.

Si vamos más allá, podemos decir que esas transparencia y explicabilidad (a partir de ahora, para simplificar, me referiré a ellas únicamente como explicabilidad) son fundamentales en las relaciones interpersonales, ya sea en contextos personales o profesionales. Cual-

quier decisión tomada por una persona es difícil que tenga un consenso absoluto independientemente del contexto en el que se produzca, pero explicarla de manera transparente al menos puede generar confianza en el resto de las personas involucradas, incluso en situaciones de desacuerdo.

Por tanto, si lo proyectamos a la IA, la explicabilidad es la base sobre la que se construye la confianza en esta tecnología. Porque dicho principio es fundamental para comprender y poder auditar sus capacidades y mecanismos de razonamiento, y que así las personas que tienen su control (independientemente del nivel de autonomía que tenga la IA) puedan confiar en la información que maneja, en como la procesa para hacer predicciones y poder tomar decisiones, y en su proceso de aprendizaje.

Ya hemos asimilado la necesidad de adoptar en cualquier sistema de información un principio ético como la privacidad. Pero en la IA se manejan nuevos escenarios que hacen necesarios más principios éticos, entre ellos la explicabilidad, para asegurar una IA responsable.

3. Iniciativas que ejemplifican la necesidad de una IA responsable

De entre las muchas iniciativas para una IA responsable, ética y normativa, cabe destacar tres iniciativas a nivel español, europeo y global. De ellas resalto a continuación algunos titulares relacionados con la explicabilidad. Pero, como decía anteriormente y desarrollaré más adelante, este es solo uno de los principios que tienen en cuenta dichas iniciativas.

- A nivel nacional, la Estrategia Nacional de Inteligencia Artificial del gobierno de España (ENIA).

En diciembre de 2020, la Secretaría de Estado de Digitalización e Inteligencia Artificial del Gobierno de España (SEDIA) publicó la ENIA¹. Esta estrategia dedica por entero uno de sus seis ejes (el sexto) a la inteligencia artificial ética y normativa.

¹ <https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/ENIA2B.pdf>

Entre sus medidas está la de la creación de un observatorio para el impacto social de los algoritmos. Dicho observatorio, el *Obi-sal*, pretende ser la semilla de la agencia estatal de supervisión de la IA que definirá, entre otras funciones, los indicadores de referencia para las auditorías sobre IA. Auditorías que deberán contemplar aspectos normativos que tendrán un sustento ético, y donde la explicabilidad será uno de los indicadores a tener en cuenta, sobre todo en sectores regulados y en aplicaciones críticas independientemente de que el sector sea o no regulado.

- A nivel europeo, el “*AI Act*” de la Unión Europea

En abril de 2021 la comisión europea presentó la “*AI Act*”, su propuesta de regulación sobre Inteligencia Artificial². El proyecto de ley pretende establecer una regulación horizontal sobre la IA y todas las empresas deberán comenzar a ajustarse en menos de dos años.

Dicha propuesta de regulación propone una clasificación con cuatro niveles de riesgo en función de la criticidad de la aplicación práctica para la que sea utilizada la IA. Asegurar la explicabilidad de la solución inteligente desde varias perspectivas es fundamental para asegurar que su comportamiento se atiene al nivel de riesgo con el que haya sido categorizada.

- A nivel mundial, la declaración de la UNESCO.

En noviembre de 2021, los 193 Estados Miembros de la UNESCO adoptaron el primer acuerdo mundial sobre la ética de la Inteligencia Artificial³. En los nueve capítulos de dicha declaración existen treinta y seis menciones a la transparencia y explicabilidad, siendo especialmente explícita en su capítulo tercero, donde indica que “La 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”.

² <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52021PC0206>

³ https://unesdoc.unesco.org/ark:/48223/pf0000380455_spa

En resumen:

- Las empresas tienen la responsabilidad de que su inteligencia artificial esté desarrollada bajo principios éticos, de igual manera que sus trabajadores deben regirse por marcos de comportamiento corporativos.
- Como siguiente paso, ya están en marcha marcos normativos que instarán a las empresas a cumplir pautas en este ámbito. Es fundamental que las empresas estén preparadas y sepan cómo afrontarlo para cuando entren en vigor sus recomendaciones y obligaciones.
- Ética y normativa se retroalimentan, ya que ambas por separado son condición necesaria pero no suficiente.

En este contexto, el principio ético de explicabilidad y su normativa asociada, entre otros, tendrá un papel fundamental.

4. El necesario papel de entidades independientes como OdiseIA

Vuelvo a retomar la declaración de la UNESCO, ya que recoge de manera explícita la necesidad de que una IA responsable sea impulsada por entidades independientes para garantizar que el objetivo final sea conseguido de manera lo más objetiva posible, sin atender a intereses empresariales, y anteponiendo siempre las necesidades de la sociedad.

“Se alienta a los Estados Miembros, las empresas del sector privado y las organizaciones de la sociedad civil a que creen una red de responsables independientes de la ética de la IA”⁴.

Por otra parte, según un informe de la Fundación Seres *“En el escenario español, destaca la iniciativa independiente OdiseIA”* (*“Decálogo IA responsable e inclusiva”*)⁵.

⁴ Capítulo IV. “Ámbito de actuación 2: Gobernanza y Administración éticas”.

⁵ 2021, Enunciado 3, página 24. https://www.fundacionseres.org/Repositorio%20Archivos/Informes/210128_Decalogo%20LabS%20IA%20seres%20everis.PDF

OdiseIA (Observatorio del Impacto Social y Ético de la Inteligencia Artificial) es una asociación sin ánimo de lucro creada a finales de 2019, independiente y no sujeta a otros intereses comerciales. Es una plataforma de colaboración abierta a empresas, administraciones públicas, centros de investigación, escuelas de negocio, universidades y profesionales independientes que se identifiquen con su misión, que es “trabajar activamente por un uso ético y responsable de una Inteligencia Artificial que tenga en su centro el bienestar del ser humano”.

5. La necesidad de aterrizar los conceptos de la IA responsable

Existen muchas investigaciones alrededor de la IA ética. La conclusión a día de hoy (julio de 2022), sigue siendo la misma que hizo la Universidad de Harvard en su informe “*Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*” publicado en 2020⁶: “*existe una amplia brecha y complicada articulación entre conceptos de alto nivel y su aplicación en el mundo real*”.

Este reto, esta necesidad de aterrizar estos conceptos, es recogido también por la anteriormente mencionada declaración de la UNESCO. Dicha declaración indica cómo “*Las empresas deberían aplicar medidas adecuadas para vigilar todas las etapas del ciclo de vida de los sistemas de IA.*”⁷. Dicha declaración recalca también como es necesario centrarse en la “*articulación de valores y principios, sino también en su aplicación práctica, mediante recomendaciones concretas*”⁸.

Estos dos ejemplos permiten concluir como es necesario profundizar en planteamientos que permitan a las empresas asimilar la inteligencia artificial responsable de manera pragmática en su día a día. Aterrizar estos conceptos mediante recomendaciones, tecnologías, herramientas metodológicas y modelos de gobierno que ayuden a las empresas a saber cómo hacerlo.

⁶ <https://dash.harvard.edu/handle/1/42160420>

⁷ Capítulo IV. “Ámbito de actuación 1: evaluación del impacto ético”.

⁸ Capítulo II. “Fines y Objetivos”

6. Un ejemplo de aterrizaje. La iniciativa GuIA de OdiseIA

GuIA es una iniciativa de OdiseIA lanzada el 27 de noviembre de 2020⁹, contando con la presencia de la Secretaría de Estado de Digitalización e Inteligencia Artificial del gobierno de España (SEDIA), y los responsables de IA ética de Google, Microsoft, IBM y Telefónica.

GUIA
UNA INICIATIVA DE ODISEIA

COMIENZA EL ATERRIZAJE DE LA ÉTICA EN LA INTELIGENCIA ARTIFICIAL

Acompáñanos en el arranque del proceso que generará una **GuIA** personalizada para cada sector con las mejores prácticas éticas y responsables en Inteligencia Artificial. Con la participación de importantes compañías tecnológicas y sectoriales, y de la Secretaría de Estado de Digitalización e Inteligencia Artificial.

VIERNES 27 DE NOVIEMBRE / 9:30 - 11:30 h.
(Hora española)

Registro

CARME ARTIGAS
Secretaría de Estado de Digitalización e Inteligencia Artificial

PILAR MANCHÓN
Senior Director of AI Research Strategy

ENRIC DELGADO SAMPER
Chief Technology Officer IBM Spain, Portugal, Greece and Israel

DAVID CARMONA
General Manager, Artificial Intelligence & Innovation

FRANCISCO MONTALVO
Chief Data Officer

El evento fue una declaración de intenciones y objetivos, al que siguieron tres meses para definir la estrategia para conseguirlos, dotando a la iniciativa de estructura conceptual y operativa.

La iniciativa contó con el interés de varias compañías que quisieron acompañar a OdiseIA en su desarrollo, señal del interés y confianza en OdiseIA en general y por la iniciativa GuIA en particular. Y señal de lo más importante, la necesidad de su objetivo: la definición de buenas prácticas aterrizadas para una IA responsable.

En junio de 2021 comenzó su *delivery* con un equipo de OdiseIA al que se incorporó *PwC* España. Esta compañía multinacional presentó una excelente propuesta técnica, al igual que el resto de las recibidas, como respuesta a la RFP lanzada por OdiseIA. Además, *PwC* propuso realizar el proyecto de manera pro-bono, ejemplo del compromiso de una gran compañía por tener impacto positivo en empresa y sociedad.

⁹ Vídeo del evento: <https://youtu.be/m62Y9jubitck>

El 17 de febrero de 2022 fue presentado el primer resultado de la iniciativa, un documento de 232 páginas, en un evento híbrido¹⁰ con más de 300 asistentes del que se hicieron eco importantes medios de comunicación nacionales e internacionales como *Business Insider*, *Cinco Días*, *ComputerWorld*, *Dircom*, *El Confidencial*, *El Mundo*, *Europa Press*, *msn*, *The global frontier*, *The observatorio*, etc¹¹.

El documento presentado en dicho evento¹² expone el *FrameWork* GuIA, el primer marco que integra de manera trazable tres dimensiones: la ética, la jurídica y la tecnológica. Las dos primeras porque, como indicaba anteriormente, por sí solas son condición necesaria pero no suficiente. La tercera, para saber cómo “GuIAr” la gestión de las dos primeras.



- Un marco ético con ocho principios éticos, entre ellos la explicabilidad, resultantes de la destilación de decenas de iniciativas de todo el mundo en el ámbito académico y empresarial.
- Un marco jurídico que incluye las normativas que aplican a cada principio ético.
- Un marco tecnológico que homogeneiza y resume los extensos planteamientos de *Google*, *Microsoft* e *IBM*, y que además agrega la experiencia de *Telefónica*, una compañía española pionera en el mundo en la adopción de IA de manera responsable en sus procesos de negocio.

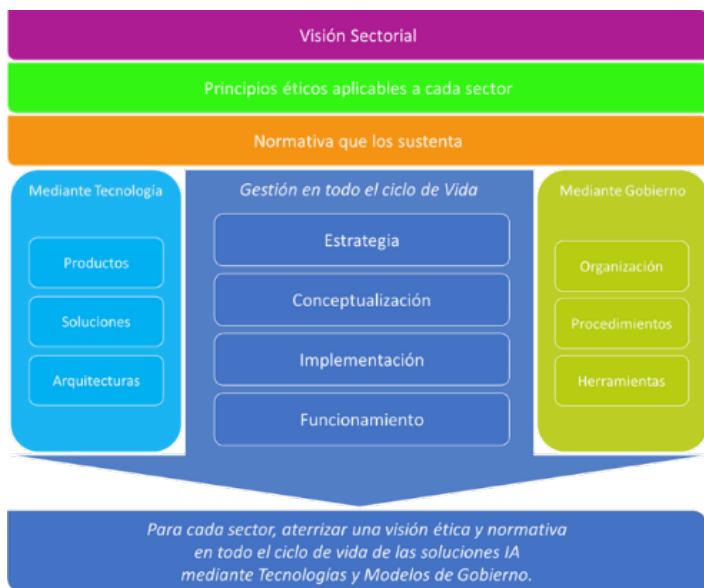
En resumen, un marco que aterriza mediante una visión 360 los principios éticos y su normativa asociada con las tecnologías y modelos de gobierno que permiten su gestión, que lo hace en todo el ciclo de vida de las soluciones Inteligentes (visión *end-to-end*), y de manera verticalizada a los casos de uso de cada sector empresarial. Un aterri-

¹⁰ Vídeo con la grabación del evento: <https://youtu.be/NxPUKJDqVxE>

¹¹ Como ejemplo, la reseña en *Cinco Días*, https://cincodias.elpais.com/cincodias/2022/02/17/companias/1645124315_805328.html

¹² <https://www.linkedin.com/feed/update/urn:li:activity:6900345521412608001/>

zaje desde el principio ético y su normativa asociada hasta, en muchos casos, el algoritmo necesario para automatizarlo.



Este primer documento GuIA es una base sólida válida para todos los sectores empresariales. Sin embargo, cada sector tiene sus necesidades particulares. Por ejemplo, por resaltar algunas diferencias globales:

- Un sector regulado necesita de un mayor soporte en el marco jurídico.
- Los principios éticos tienen distinta relevancia en un sector industrial donde prevalece la componente “máquina” física gestionada mediante IA, que en el sector *retail* donde prevalece la sensibilidad de los consumidores, que en el sector salud donde el uso de la IA puede salvar vidas, o en cualquier sector regulado donde la explicabilidad será fundamental para exponer a los auditores el funcionamiento de la misma.

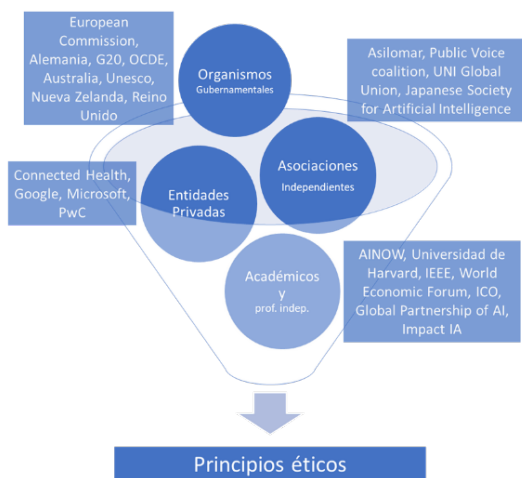
Por eso actualmente se está adaptando dicho *framework* a 10 sectores empresariales, de la mano de más de 50 empresas. Un trabajo entre entidades con intereses empresariales solapados, que no dudan

en colaborar ya que todas entienden la importancia del objetivo de la iniciativa GuIA en pro del beneficio de empresa y sociedad en su totalidad. Todas tienen un mismo interés común alineado con la misión de OdiseIA, que recuerdo es: *trabajar activamente por un uso ético y responsable de una inteligencia artificial que tenga en su centro el bienestar del ser humano.*

7. La explicabilidad es solo uno de los principios éticos necesarios en la IA

Como ya he adelantado anteriormente, además del principio de explicabilidad, en la IA existen otros principios necesarios para que esta tecnología pueda ser asimilada de una manera responsable. Los principios éticos (y por tanto su normativa asociada) difieren en número y descripción respecto de la iniciativa que consultemos, pero en esencia son muy parecidos entre todas ellas.

En el caso de la iniciativa GuIA anteriormente mencionada, su marco ético contempla ocho principios éticos, entre ellos el de la explicabilidad. En ese caso, dicho marco ético ha sido generado destilando 150 iniciativas de todo mundo, 27 de ellas con detalle. Iniciativas gubernamentales, empresariales, académicas y de asociaciones independientes para así tener una visión lo más complementaria posible.





Todos estos principios son necesarios en la IA. Tenemos por tanto que ir acostumbrándonos a ellos, y a saber cómo operativizarlos cuando las empresas sectoriales construyan sus propias soluciones inteligentes.

La mayor parte de los principios aplican de una manera u otra a casi cualquier aplicación práctica de la IA, siendo fundamental conocer de qué manera aplica a cada caso de uso en particular, y como solucionar dicho escenario.

Además, confluye una derivada que aporta mayor complejidad, y es que estos principios en la IA deben verse como un todo, ya que tienen relaciones más o menos explícitas entre ellos.

De hecho, así lo recoge la declaración de la UNESCO indicando como “Los valores y principios fundamentales deben considerarse complementarios y relacionados entre sí”¹³. Sobre la necesidad de esta visión holística teniendo en cuenta todos los principios y su relación entre ellos profundizaremos más adelante en este capítulo, en el apartado “*Relación de la Explicabilidad con otros principios éticos para una IA responsable*”

¹³ Capítulo III. “Disposiciones Finales”.

Un resumen de lo visto en el apartado que aquí concluye que:

- La IA es una tecnología diferencial respecto de las que hemos asimilado en las últimas décadas.
- Las organizaciones tienen la responsabilidad de que sus soluciones inteligentes estén desarrolladas bajo principios éticos, siendo la explicabilidad uno de los más importantes, pero no el único.
- Ética y normativa están relacionadas. Ambas por separado son condición necesaria, pero no suficiente.
- Ya están en marcha marcos a nivel nacional, europeo e internacional (ENIA, *AI Act*, UNESCO) que instarán a las organizaciones a cumplir determinadas pautas. La explicabilidad será uno de los principios éticos a tener en cuenta, pero también otros. Es fundamental estar preparados para cuando entren en vigor.
- El papel de las entidades independientes como OdiséIA y de iniciativas como GuIA es importante y necesario.
- Enfoque holístico: Hay que tener en cuenta todos los principios éticos, y la relación entre ellos. Existen relaciones directas entre la explicabilidad y otros principios necesarios en la IA.

Es necesario aterrizar los conceptos de IA responsable al día a día de las empresas sectoriales. Necesitan entender cómo hacerlo al construir sus propias soluciones inteligentes *ad-hoc* que soportan sus procesos de negocio.

III. EXPLICABILIDAD. RECOMENDACIONES ATERRIZADAS AL DÍA A DÍA DE LAS EMPRESAS

A continuación me dispongo a explicar, enfocado en el principio de explicabilidad:

- Principales recomendaciones de *Google*, *Microsoft*, *IBM* y *Teléfonoica*.
- Recursos, tecnologías y modelos de gobierno para gestionarla.
- Relación de la explicabilidad con otros principios para una IA responsable.

1. Recomendaciones de Google, Microsoft, IBM y Telefónica

El lector se preguntará porqué atender a las recomendaciones de las tres primeras compañías. El motivo es sencillo: estas empresas son el primer eslabón de la cadena, los creadores de los productos y servicios inteligentes que después son utilizados en las empresas sectoriales para crear sus soluciones *ad-hoc*, esas que soportan sus procesos de negocio de manera personalizada.

Estas tres compañías llevan realizando desde finales de los '10 un notable esfuerzo por ofrecer tecnologías, herramientas y recomendaciones alrededor de la IA responsable que permitan aterrizar al día a día de las operaciones de las empresas sectoriales los conceptos académicos desarrollados durante décadas.

A continuación detallaré un decálogo que resume algunas de las muchas recomendaciones que estas empresas realizan acerca de la explicabilidad, utilizando para ello como base el primer documento GuIA mencionado anteriormente¹⁴.

1. Visión sectorial
2. Explicar el modelo y también los datos que maneja
3. Explicar lo global y lo particular
4. Adaptar la explicación al interlocutor
5. En todo el ciclo de vida de la solución inteligente
6. Menos puede ser más
7. Caja negra vs caja blanca
8. Explicaciones claras y prudentes
9. Usando causalidad, minimizando correlaciones
10. La importancia del modelo de gobierno

1.1. Visión sectorial

Como exponía al principio de este capítulo, estamos asimilando una nueva y compleja tecnología que además tiene connotaciones especiales dadas sus similitudes con las capacidades del ser humano.

¹⁴ <https://www.linkedin.com/feed/update/urn:li:activity:6900345521412608001/>

Por tanto es fundamental conocer cuáles son las necesidades del sector (seguros, finanzas, sanidad, *retail*, transporte, industria, AAPP, etc.) en relación a la explicabilidad en la IA, identificar las características claves de explicabilidad que se han de tener en cuenta en cada caso de uso (obligatoriamente en los casos donde existe una auditoría para entender el funcionamiento de la solución inteligente), el por qué, y el límite de las explicaciones, un concepto que recogeré más adelante en este mismo decálogo.

1.2. Explicar el modelo y también los datos que maneja

Es fundamental poder explicar el funcionamiento del modelo y su proceso de aprendizaje.

Pero no solo hay que comprender el razonamiento de la solución inteligente. Es fundamental comprender también los datos que maneja (estructurados o no estructurados —imágenes y lenguaje natural, por ejemplo—) para después explicar mejor los resultados que proporcione dicha solución inteligente.

Para ello es importante analizar muestras de datos, conocer su esencia, los valores críticos o atípicos de los mismos, etc., entendiendo su significado ya que, a veces, las características en un conjunto de datos dado son visibles a primera vista pero otras veces están entrelazadas, y esto puede tener implicaciones para explicar el funcionamiento del modelo.

1.3. Explicar lo global y lo particular

Es necesario explicar el funcionamiento de la solución inteligente y los datos que maneja desde lo más general a lo más particular, abordando aspectos como:

- El mecanismo de razonamiento global de la solución inteligente.
- Las predicciones y decisiones tomadas sobre un subconjunto de información con características parecidas, para garantizar la homogeneidad de su funcionamiento.
- E incluso de cada una de dichas predicciones y decisiones de manera individual.

1.4. Adaptar la explicación al interlocutor

Hay que cubrir las necesidades de explicabilidad para todos los actores que intervienen en el ciclo de vida de la solución inteligente, desde el responsable de negocio que la define, la explota, y es responsable de su funcionamiento, hasta el usuario final que la utiliza y el auditor externo que la revisa, pasando por los técnicos que la implementan.

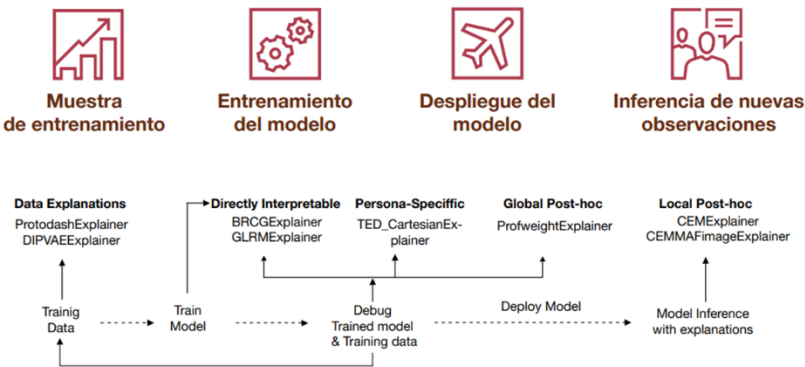
Todo ello, adaptando la explicación al tipo de interlocución de dichos actores.

- El responsable de negocio, cuyo lenguaje es técnico desde un punto de vista empresarial, en el diseño tiene que garantizar el cumplimiento de las políticas de la empresa y poder explicarlo, por ejemplo, en los correspondientes comités. Además, cuando está en funcionamiento, tiene que poder explicar, por ejemplo, la precisión del modelo, que las predicciones y decisiones que está proporcionando la solución son homogéneas en usuarios finales con características similares, etc.
- Los técnicos que la implementan, cuyo lenguaje es técnico desde un punto de vista tecnológico, cuando la están construyendo necesitan explicar a los responsables de negocio el razonamiento de sus modelos mediante un lenguaje entendible por éstos, y así asegurar entre ambos que la solución cumple con las políticas de la empresa. Y en toda la vida de la solución inteligente deben conocer con un lenguaje técnico el razonamiento que está siguiendo la solución inteligente para asegurar su correcto funcionamiento e incluso mejorar su precisión.
- El usuario final, que en muchas ocasiones no tiene porqué disponer de un lenguaje técnico de cualquier tipo, necesita conocer, por ejemplo, los motivos por los cuales no se le ha concedido un producto o servicio y saber en qué condiciones aplicadas a su realidad se le concedería.

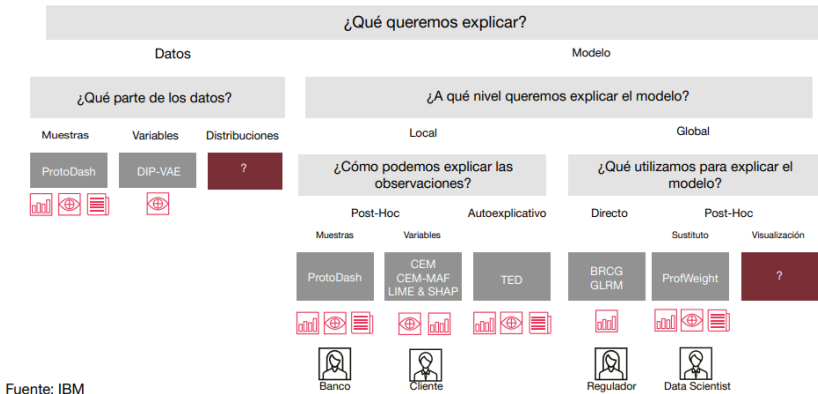
Los auditores entran en juego en todos los sectores cuando se tratan datos personales, pero especialmente en aquellos sectores globalmente regulados (banca, seguros, etc.). Ellos requieren explicación acerca de la precisión del sistema, así como una explicación acerca de todas las predicciones realizadas y decisiones tomadas por la solución inteligente.

1.5. En todo el ciclo de vida de la solución inteligente

Es necesario poder explicar la solución inteligente desde que se diseña, pasando por su implementación, y especialmente cuando está en funcionamiento. Ya existen algoritmos que permiten explicar de manera automatizada el funcionamiento de la solución inteligente en todas las fases de su ciclo de vida.



Y, además, recogiendo las recomendaciones anteriores, estos algoritmos de explicabilidad pueden explicar automáticamente datos, los modelos, de lo más global a lo más particular, y hacerlo para todo tipo de usuarios, adaptando su explicación a su nivel de interlocución.



Fuente: IBM

1.6. Menos puede ser más

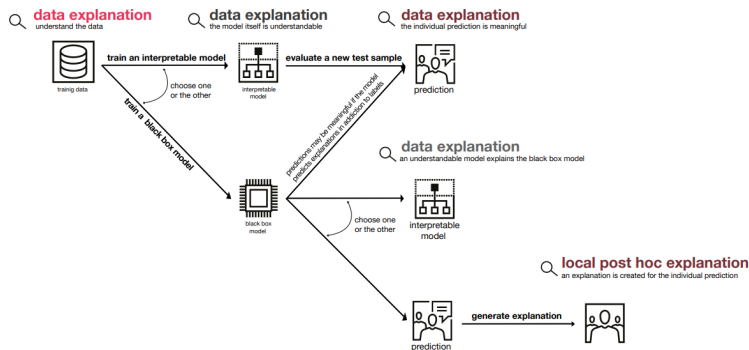
Entender el funcionamiento de soluciones inteligentes puede implicar el análisis de cantidades ingentes de variables y operaciones matemáticas. Por tanto, un buen diseño de la solución entendido desde la simplicidad de la misma puede ser clave para la lograr una mayor explicabilidad. Por ejemplo, en ocasiones se tiende a aumentar la complejidad de la solución inteligente para buscar más precisión, lo cual puede implicar la pérdida de explicabilidad acerca de su funcionamiento. Por tanto es necesario medir el retorno de inversión de complicar el modelo por obtener un poco más de precisión a costa posiblemente de perder explicabilidad.

Es recomendable usar el modelo más sencillo que cumpla con los objetivos de negocio, simplificando así la explicabilidad y el entendimiento de su funcionamiento. Así es más fácil identificar qué factores afectan al modelo y poder explicarlo. Además, esta simplicidad en la solución redunda en un menor consumo energético, algo muy importante en estas soluciones inteligentes, cuya necesidad de computación implica un alto impacto medioambiental.

1.7. Caja negra vs caja blanca

Los modelos de caja negra (redes neuronales, por ejemplo) ofrecen una gran precisión y pueden solucionar casos de uso complejos, pero no son directamente interpretables dada su complicada construcción. Su falta de explicabilidad puede implicar tener que descartarlos para recurrir a modelos directamente interpretables, de caja blanca, sobre todo en aplicaciones críticas de alto riesgo donde hay que asegurar que su funcionamiento no se degenera, o en sectores regulados donde la explicabilidad es una necesidad ineludible.

La explicabilidad en modelos de caja negra es un reto sobre el que se está investigando. Actualmente, una de las posibles soluciones para explicarlos (no al 100%) pasa por algoritmos que sondean la red neuronal analizando las entradas y las salidas correspondientes que producen. Una vez analizados estos datos, se genera un modelo directamente interpretable equivalente que sirve para explicar el funcionamiento de la caja negra.



1.8. Explicaciones claras y prudentes

Es necesario explicar el funcionamiento de la solución inteligente de manera clara y concisa. Si no es así, los actores que interactúan con ella en todo su ciclo de vida pueden tender a especular y elaborar sus propias teorías sobre cómo funciona, lo que puede reducir su confianza en la misma si su funcionamiento finalmente no cumple con sus expectativas.

Además, las explicaciones tienen que ser prudentes, identificando los escenarios en los que éstas puedan no ser apropiadas, al menos de manera pública si, por ejemplo:

- Derivan en mayor confusión en los usuarios.
- Pueden ser aprovechadas por agentes externos para vulnerar la seguridad del sistema y así “pervertir” su funcionamiento, sus procesos de aprendizaje y razonamiento. Esto es algo particular en las soluciones inteligentes respecto de los sistemas de información tradicionales, y es conocido como “ataques de envenenamiento”.
- Revelan información privada y confidencial.

En estos casos, se recomienda considerar alternativas a través de otros mecanismos, como auditorías internas, o el permitir que los usuarios impugnen o influyan en decisiones futuras mediante el *feedback* que compartan acerca del funcionamiento de la solución.

1.9. Usando causalidad, minimizando correlaciones

La correlación no implica causalidad, ya que dos variables pueden tener una alta correlación, pero pueden no ser una causa de la otra. Un sencillo ejemplo de estos conceptos es usar directamente la estatura y no la edad para predecir si es seguro físicamente para un niño subirse a una montaña rusa. Con la estatura estaríamos utilizando una causalidad directa, mientras que con la edad estaríamos utilizando una correlación teórica que, de no cumplirse, podría producir accidentes.

Esto es importante en la IA teniendo en cuenta la gran cantidad de variables que manejan los modelos, y las muchas relaciones explícitas e implícitas que pueden existir entre las mismas. La correlación además puede estar sujeta a la subjetividad de quien diseña e implementa la solución inteligente y las reglas de correlación que considera, o puede no contemplar escenarios no habituales pero relevantes. Es por tanto una potencial fuente de sesgo, y como consecuencia un riesgo de no tener en cuenta el también necesario principio de equidad.

1.10. La importancia del modelo de gobierno

Para finalizar el decálogo de recomendaciones sobre explicabilidad, hemos visto como muchas de ellas son gestionables desde la tecnología, ya sea de manera automática o con decisiones acerca del diseño de la misma.

Más allá de dichas recomendaciones, es necesario acompañarlas de un buen modelo de gobierno que incluya estructuras organizativas y procedimientos dentro de la empresa que garanticen la explicabilidad en todo el ciclo de vida de la solución inteligente.

Como hemos visto en este decálogo, se puede automatizar la gestión de algunas componentes de explicabilidad, utilizando para ello tecnologías que detallaremos en el siguiente punto. Pero todos los principios, entre ellos la explicabilidad, deben ser gestionados mediante modelos de gobierno con ejemplos como el de *Telefónica*, que veremos también más adelante.

2. Recursos, tecnologías y modelos de Gobierno para gestionarla

Esta es una muestra de algunas de las herramientas de Google, Microsoft e IBM que más relación directa tienen con la explicabilidad, con la descripción de sus principales capacidades.

		
Tecnologías		
Explainable AI. Explicación de predicciones en tiempo real, detectar y resolver sesgos, desvíos y vacíos en datos y modelos. What-if tool. Explicabilidad de modelos de caja negra. Language Interpretability Tool. Interpretabilidad para modelos de procesamiento de lenguaje natural. MLMetadata. Permite rastrear el origen de un mal funcionamiento.	InterpretML, para explicar el funcionamiento de modelos de caja blanca y de caja negra. De manera global, local y por subconjunto de la muestra. DiCE, para gestionar la contrafactualidad, “interrogando” al modelo para encontrar los cambios necesarios que cambiarían su decisión y así ofrecer una explicación complementaria a la decisión tomada. EconML, para gestionar la inferencia causal y conocer la causa de predicciones concretas.	IBM AI Explainability 360. Suite para la explicabilidad, orientada a perfiles técnicos. IBM Watson Openscale. Tiene como base las librerías de IBM AI Explainability 360 a las que se ha añadido, entre otras funcionalidades, una interfaz visual que permite su uso por perfiles no técnicos. IBM Uncertainty Quantification 360, para conocer el nivel de (in)certidumbre de los modelos.
Recursos metodológicos		
Model Card Toolkit. Proporciona información acerca de cómo utilizar el modelo. Model Cards. Modelo documental para proporcionar información que facilite la explicación.	Datasheets for Datasets. Para explicar los datos usados en el entrenamiento del modelo.	FactSheet. Información sobre la creación y el despliegue de un modelo IA.

Tengamos en cuenta que, como comentaba anteriormente, la mayor parte de los principios éticos aplican de alguna manera a la mayor parte de los casos de uso. Y, además, tienen relación entre ellos. Como consecuencia, algunas de las citadas herramientas que permiten

la gestión de principios éticos sirven para gestionar varios de ellos, o bien es necesario integrar varias de ellas para asegurar una IA responsable de manera holística. Todo ello dependiendo del caso de uso práctico en el que se esté aplicando la IA. Estas herramientas y sus capacidades son descritas con detalle en el informe GuIA mencionado anteriormente.

Por otra parte, como indicaba también anteriormente en el decálogo de recomendaciones, además de tecnologías y recursos metodológicos es necesario un modelo de gobierno que garantice la explicabilidad (y el resto de los principios) en todo el ciclo de vida de la solución inteligente (recomendación #10).

Y, a este respecto, la experiencia de *Telefónica* es fundamental ya que esta multinacional española es pionera en su aplicación. *Telefónica* ya dispone de un modelo de gobernanza que incluye organización y procedimientos que involucran a todos los perfiles en el desarrollo de productos internos que incluyan IA y/o datos masivos.



Estos son algunos de los aspectos principales de este sistema de gobernanza:

- Cuestionario de autoevaluación para que los responsables de producto y equipos de desarrollo lo completen durante la fase

de diseño y a posteriori en sucesivas fases durante todo el ciclo de vida del producto.

- Proceso de escalado entre niveles para garantizar el análisis detallado de casos de riesgo detectados.
- La figura de los *RAI Champions* (*Responsible Artificial Intelligence*), como parte del proceso de gobierno de la IA responsable, los cuales están cerca de toda la organización informando, formando, guiando, coordinando, creando conciencia y gestionando el cambio hacia una IA responsable.

Además, *Telefónica* tiene automatizada la privacidad en sus sistemas (denominados “*cuarta plataforma*”) mediante una arquitectura tecnológica (“*Privacy by design*”). Según este enfoque, esta automatización le va a permitir escalar en la adopción del resto de principios para una IA responsable, entre ellos la explicabilidad, algo que ya ha comenzado mediante la integración de un marco ético sobre dicho marco de privacidad (“*Ethics by design*”).

En el mencionado informe GuIA se detalla el caso de éxito de *Telefónica* en la adopción de la IA de manera responsable, describiendo su contexto, historia, principios considerados, aprendizajes y beneficios obtenidos.

3. Relación de la explicabilidad con otros principios para una IA responsable

Recuerdo como la explicabilidad no es el único principio necesario en la IA, y como todos estos principios nos resultarán en poco tiempo tan cercanos en la inteligencia artificial como el ya muy desarrollado principio de privacidad en los sistemas de información tradicionales.

Y como, además, estos principios en la IA deben verse como un todo, tal y como recoge la declaración de la UNESCO. Por tanto, es importante considerar las relaciones más o menos explícitas entre los mismos.

Esas relaciones son particulares en función del sector y del caso de uso en particular, pero a continuación indico algunas ideas que permiten identificar dichas relaciones en la mayor parte de las aplica-

ciones prácticas. Ideas que de alguna manera ya he esbozado en las recomendaciones y contenidos vistos anteriormente.

- Justicia. Las correlaciones son una potencial fuente de sesgo y, como consecuencia, un riesgo de no tener en cuenta el también necesario principio de justicia (también denominado “equidad”). Por tanto, hay que minimizarlas en favor de la causalidad.
- Control y vigilancia humana. Hay que cubrir las necesidades de explicabilidad para los responsables de la solución inteligente, ya que ellos son los que deben vigilarla y mantenerla bajo su control.
- Promoción de derechos humanos. Explicar la solución inteligente desde varias perspectivas y mediante lenguaje entendible por todos los usuarios y responsables es fundamental para asegurar que se contemplan los derechos de las personas.
- Privacidad. Las explicaciones claras, pero también prudentes, requieren de un equilibrio para así no revelar información privada utilizada por la solución inteligente.
- Fiabilidad. Hacer públicos determinados aspectos de la solución inteligente puede facilitar la vulnerabilidad de la misma ante ataques orientados a “pervertir” su funcionamiento, sus procesos de aprendizaje y razonamiento. Esto es algo particular en las soluciones inteligentes respecto de los sistemas de información tradicionales, y es conocido como “ataques de envenenamiento”.
- Sostenibilidad ambiental. Un modelo sencillo que cumple los objetivos de negocio, pese a que pueda perder un poco de precisión, facilita su explicabilidad. Como consecuencia de esta simplicidad, además se consigue un menor consumo energético, algo muy importante en estas soluciones inteligentes, cuya necesidad de computación implica un alto impacto medioambiental.
- Responsabilidad y rendición de cuentas. Hay que cubrir las necesidades de explicabilidad para todos los actores que intervienen en el ciclo de vida de la solución inteligente, puesto que son los responsables últimos de la misma.

En resumen, estas son las principales ideas de este apartado del presente capítulo:

- Las recomendaciones de empresas tecnológicas como *Google*, *Microsoft* e *IBM* son fundamentales para abordar la explicabilidad al crear nuestras soluciones de negocio *ad hoc* a partir de sus productos, cuya gestión será particular en función del sector y caso de uso.
- Enfoque holístico: Hay que tener en cuenta todos los principios, y la relación entre ellos. Existen relaciones directas e indirectas entre la explicabilidad y otros principios necesarios en la IA.
- Como consecuencia, muchas de las herramientas que permiten la gestión de principios éticos sirven para gestionar varios de ellos, o bien es necesario integrar varias de ellas para asegurar una IA responsable, todo ello dependiendo de los casos de uso en los que se aplique la IA.

La gestión de algunos principios éticos se puede automatizar con tecnología. Todos los principios éticos se pueden gestionar con modelos de gobierno como el de *Telefónica*.

IV. EPÍLOGO

Existen principios éticos, como la privacidad, que tenemos más arraigados dado su desarrollo normativo en los últimos años y su aplicación automatizada en los sistemas de información tradicionales. Pero, y ahí radica una novedad muy importante, en la IA responsable no solo existe este principio. Existen más de cinco principios tan importantes como el de la privacidad, y por tanto tenemos que ir acostumbrándonos a todos ellos.

De entre estos principios, el de explicabilidad es fundamental porque es la base sobre la que se construye la confianza en la IA. Dicho principio es necesario para comprender y poder auditar sus capacidades y mecanismos de razonamiento para que las personas que tienen su control (independientemente del nivel de autonomía que tenga la IA) puedan confiar en la información que maneja, como la utiliza en su proceso de aprendizaje, y en como la procesa para hacer predicciones y poder tomar decisiones.

Y, por último, debemos tener muy presente que todos los principios tienen relación entre sí, lo cual hace más interesantes y complejos los mapas conceptuales de negocio, éticos, normativos y tecnológicos necesarios para gestionarlos. Es fundamental visualizarlos de manera holística e integrada, y saber cómo operativizarlos, como aterrizarlos al día a día de las soluciones inteligentes que soportan de manera personalizada los procesos de negocio de las empresas.

La moderna configuración de la transparencia en la prestación de servicios y en la contratación en Internet

JAVIER PLAZA PENADÉS
Catedrático de Derecho Civil
Universitat de València

I. LA TRANSPARENCIA COMO VALOR Y COMO PRINCIPIO GENERAL

Comenzaré con una conclusión evidente y es que la transparencia es un valor y un principio que se proyecta a instituciones, productos y servicios.

Como valor, podría tener perfectamente rango de constitucional, junto con el resto de valores superiores del ordenamiento jurídico, como son la libertad, la igualdad, la justicia y el pluralismo político. Piénsese que a fecha de hoy, la transparencia se exige tanto a los distintos poderes del Estado y al sector público, pero también a las principales empresas del sector privado (con independencia de que se relacionen con el sector público y estén obligadas por Ley) e incluso a los partidos políticos y al tercer sector, conformado por las asociaciones y a las fundaciones¹.

Pero la transparencia, aparte de a las instituciones, también se debe aplicar a aquellos productos y servicios que se ofertan a través de la contratación con condiciones generales (especialmente en la contratación electrónica a través de Internet), puesto que cumplir con un riguroso principio de transparencia evidencia ya un cumplimiento normativo o de legalidad, que en el moderno ámbito jurídico se denomina *Compliance*, pero también es verdad que dicho cumplimiento normativo de la transparencia añade otros valores intangibles a la

¹ F.J. Orduña Moreno, “La transparencia como valor del cambio social: su alcance constitucional y normativo”, *Actualidad jurídica Aranzadi*, núm. 944, 2018, pp. 12 y ss.

institución o al producto o servicio del que se predicara transparencia, como son la calidad de la institución o del producto y servicio y la confianza que genera dicha institución o dicho producto y servicio en los terceros.

Así pues, la transparencia, como principio, conlleva la necesidad de hacer visibles y comprensibles las actuaciones y decisiones de una institución, o las principales características y singularidades de un producto o de un servicio, por lo que se recomienda el seguimiento de una especie de “protocolo de transparencia”, *compliance* que permita verificar y hacer prueba de su cumplimiento normativo.

Sentada esta premisa, paso a desarrollar la aplicación práctica de la transparencia en Internet, que se inició en 1969 como proyecto militar, y que hacia la década de los ochenta, con el impulso del MIT y de las instituciones universitarias, se convirtió en una gran red de comunicación, los que nos ha permitido hacer el tránsito de la sociedad de la información a la sociedad del *Big Data*, puesto que prácticamente la totalidad del conocimiento humano se encuentra ya disponible en internet en todos los idiomas. Sin embargo, a efectos de una aplicación práctica del principio de transparencia, internet ha sido objeto de una regulación exhaustiva por parte de la Unión Europea, que todavía no ha culminado y que lo hará en breve; regulación de la que daré cuenta a continuación.

II. PRINCIPIO DE TRANSPARENCIA Y PROTECCIÓN DE DATOS

Como bien sabemos el derecho de protección de datos de carácter personal es un derecho fundamental, que tiene por objeto los datos personales de concretas personas físicas, y que a su vez se materializa en otros derechos, como son los de acceso, rectificación, cancelación, oposición, portabilidad y conocimiento respecto de la existencia de perfiles personales.

En ese sentido, la anterior Ley Orgánica 15/1999, en su artículo 5 ya establecía un principio de calidad de los datos, lo que servía para delimitar la transparencia exigible en la información de los mismos, obligando a la persona que quería tratar o ceder datos personales a justificar qué datos necesitaba tratar o ceder y con qué finalidad, no

pudiendo exigir el tratamiento de aquellos datos que no fueran necesarios para la finalidad perseguida, ni hacer uso de dichos datos más allá de la finalidad exacta para lo que se obtuvieron, y mucho menos hacerlo con finalidades incompatibles con aquella respecto de la que se había informado, por lo que sobre este principio de transparencia y calidad de los datos el titular de los datos otorgaba su consentimiento informado cuando no había una base legal para el tratamiento o para la cesión de los mismos.

Ese principio de transparencia en la protección de datos personales tiene hoy una nueva dimensión en el Reglamento General de Protección de Datos de la Unión Europea y en la Ley Orgánica 3/2018, de 5 de diciembre, porque al principio de transparencia y calidad de los datos se ha añadido la llamada “*Accountability*” o responsabilidad desde el diseño (*by design* y por defecto (*by default*). Y si bien es cierto que el tratamiento y la cesión de datos por defecto ya estaba en nuestro Ordenamiento Jurídico, con ese principio de minimización a la hora de obtener los datos personales, en el sentido de no exigir más datos de los estrictamente necesarios para la finalidad perseguida, la novedad viene en la *Accountability* o transparencia desde el diseño, lo que obliga a elaborar cualquier producto o a diseñar la prestación de cualquier servicio en el que intervengan datos personales involucrando a todas las personas de la institución que directa o indirectamente van a participar en el tratamiento. Por tanto, la Unión Europea exige que las instituciones diseñen su transparencia con un principio de “responsabilidad proactiva”, es decir con carácter previo a su implementación, buscando el menor impacto y el mayor respeto al derecho fundamental de protección de datos, e involucrando en la implementación del mismo a toda la institución, que deberá estar informada y deberá recibir la formación necesaria para cumplir este objetivo.

Además, las exigencias de transparencia en materia de protección de datos obligan a utilizar una doble capa o nivel de información, con un nivel básico y un nivel más desarrollado, buscando información gráfica o visual que permita comprender el contenido escrito del que se quiere informar de forma transparente.

Todo ello nos permite ya trasladar los requisitos que se exigen para satisfacer los deberes de transparencia en materia de protección de datos a cualquier producto servicio o institución a la que se quiera

aplicar un protocolo de transparencia: buscando siempre una responsabilidad proactiva (en el que tiene que participar toda la institución, participando en el propio diseño y recibiendo la información necesaria para ello), y tratando de hacer comprensible la información con un sistema de doble capa y con dibujos y diseño, incluso con material audiovisual que favorezcan su comprensión.

III. “LOS DERECHOS FUNDAMENTALES DE LA PERSONA PREVALECE SOBRE LA TECNOLOGÍA”

Pero además, el derecho de protección de datos nos ha brindado un nuevo derecho: el derecho al olvido, que es el que tiene cualquier ciudadano para que las informaciones que se ofrecen a través de las redes y que se hacen accesibles a través de los motores de búsqueda (como *Google*, *Yahoo*, o el *Bing* de *Microsoft*) cuando ya no son exactas o están desactualizadas y causan daño al honor a la intimidad de la persona, desaparezca. Este es el famoso caso Mario Costea versus *Google*, en la STJUE (Gran Sala) de 13 de Mayo de 2014, asunto C-132/12, donde más allá del reconocimiento de dicho derecho, lo que estaba en juego era si el sistema tecnológico y de inteligencia artificial diseñado por el conocido buscador, sistema complejo que funciona mediante una tecnología al margen de la voluntad humana, basado en una serie de algoritmos complejos que sitúan las informaciones en primer lugar en atención al número de visitas o en atención a la relevancia de la propia información, hacían inviable en principio que la compañía *Google* pudiera satisfacer este tipo de derechos, precisamente porque el diseño tecnológico escapaba a la voluntad de la propia compañía, que había diseñado ese sistema de búsqueda y de localización, de tal suerte que si no se suprimían las informaciones en origen, las propias máquinas, servidores y motores de búsqueda, volverían a ubicar esa información no deseada entre las primeras que aparecen a la hora de introducir el nombre y apellidos de una persona.

En definitiva, lo que estaba en juego era si el progreso científico y tecnológico debían de sacrificarse para satisfacer los derechos fundamentales de un simple ciudadano de la Unión Europea. Y no olvidemos que el informe del abogado general en este asunto era favorable

a los intereses de *Google*, pero también es cierto que la gran sala del Tribunal de Justicia de la Unión Europea dicta esta sentencia para sentar un principio mayor, y que debe de impregnar cualquier sistema tecnológico de inteligencia artificial, y que podemos enunciar así:

“Los derechos fundamentales de la persona prevalecen sobre la tecnología”.

Por tanto, cualquier diseño de un producto de un servicio con *big data* o inteligencia artificial o el diseño de un “smart contract,” deberá de tener en cuenta este principio de prevalencia de los derechos fundamentales (como la protección de datos) sobre la tecnología, principio que debe extenderse al resto de derechos fundamentales como el derecho al honor, a la intimidad, o la propia imagen, entre otros. De tal suerte que implementar una tecnología, que amparada en el anonimato de los que interviene, lesione derechos fundamentales básicos de las personas como el honor o la intimidad, o no pueda atender las reclamaciones de las personas damnificadas por las intromisiones ilegítimas a sus derechos respecto de la retirada de dicha información, es una tecnología que no respeta los derechos fundamentales y de la que su titular puede hacer dos cosas: suprimirla, o responder de los daños causados con el mismo grado de responsabilidad de la que lo serían las personas que advierten esas opiniones anónimas y difamatorias o lesivas a derechos fundamentales.

Por tanto, podemos señalar que este principio de primacía de los derechos fundamentales sobre la tecnología resulta básico y esencial en el diseño de cualquier programa o plataforma informática y debe ser tomado en consideración en cualquier protocolo de transparencia.

IV. LA TRANSPARENCIA EN INTERNET

Es curioso comprobar como Internet, cuyo nacimiento se vislumbró como una red que escapaba del control de la legislación de los diferentes Estados, configurándose como una realidad supra nacional, que además debería de ser el reducto y el exponente del ejercicio de una libertad pura y sin cortapisas o condicionantes, libre de intereses espurios, hoy requiere sin duda de un notable ejercicio de transparen-

cia, precisamente para garantizar el adecuado ejercicio de las libertades y derechos fundamentales en Internet².

Por tanto, podemos decir sin caer en ningún tipo de exageración, que la transparencia no se configura como un límite a las libertades y a los derechos fundamentales en Internet, sino como una garantía de imparcialidad y objetividad para el adecuado ejercicio de dichas libertades y derechos, evitándose así las actuaciones arbitrarias, discriminatorias e injustificadas, especialmente cuando inciden en ámbitos como la libertad de expresión e información, donde los prestadores de servicios de intermediación deben de garantizar el adecuado respeto a los derechos fundamentales de la persona, teniendo la posibilidad de cerrar algunos espacios o limitar algunas ideas y opiniones expresadas en dichos espacios. Por ello, la Directiva 2000/31/CE, de servicios de la sociedad de la información y comercio electrónico, va a ser modificada precisamente en la parte que refiere a la responsabilidad de los prestadores de servicios de Internet (los llamados *Internet Services Providers*), especialmente aquellos prestadores que permiten el acceso y el alojamiento de espacios en la red, y que por tanto pueden suprimir o limitar dicho alojamiento de forma temporal o permanente; así como de la responsabilidad por esos mismos servicios de acceso, transmisión de información y alojamiento de las plataformas *online*. En ese sentido, se exige una máxima transparencia en la actuación de estos prestadores de servicios de intermediación a la hora de decidir los criterios objetivos por los que se decide suprimir los servicios de intermediarios, pero ahora se va exigir que se actúe en estos casos con la máxima transparencia, con su consiguiente protocolo de actuación, que va a ser regulado en la Unión Europea a través de un Reglamento, que se encuentra formulado actualmente como propuesta de Reglamento (en concreto, la Propuesta de Reglamento de 15 de diciembre de 2020 COM(2020) 825 final) relativo a un mercado único de servicios digitales (Ley de Servicios Digitales o *Digital Services Act*), con un contenido muy completo.

Por tanto, la moderna regulación de la transparencia la encontramos ahora focalizada en la prestación de servicios, con la necesidad

² J. Plaza Penadés, “La libertad el Internet”, *Revista Aranzadi de derecho y nuevas tecnologías*, núm. 56, 2021, pp. 19 y ss.

de establecer y publicar los criterios que pueden llevar a restringir o suprimir un servicio, y en la propia regulación de la transparencia en la contratación, tanto en la que se realiza a través de condiciones generales de la contratación como la que utiliza los llamados *dark patterns* o patrones oscuros.

1. Transparencia en la contratación con condiciones generales

La relevancia que tiene el control de abusividad en orden a referir la especialidad o peculiaridad jurídica que informa al fenómeno de las condiciones generales ha hecho, por lo general, que el análisis de esta figura se centre, de forma casi autónoma, en este instrumento de control de la eficacia de la reglamentación predispuesta referenciándose, hasta la irrupción reciente del control de transparencia, en la tipicidad normativa que presenta el control de contenido o de abusividad, el cual se centraba en el desequilibrio de derechos impuesto por la parte predisponte³.

³ J. Alfaro Águila-Real, y J.C. Paz-Ares Rodríguez, *Las condiciones generales de la contratación: estudio de las disposiciones generales*, Civitas, Madrid, 1991; R. Bercovitz Rodríguez-Cano (coord.), *Comentarios a la Ley de Condiciones Generales de la Contratación*, Aranzadi, Cizur Menor (Navarra), 2000; R. Bercovitz Rodríguez-Cano (coord.), *Comentario del Texto refundido de la Ley general para la defensa de los consumidores y usuarios y otras leyes complementarias: (Real Decreto Legislativo 1/2007)*, Aranzadi, Cizur Menor (Navarra), 2009; S. Cámara Lapuente, “La codificación del derecho de consumo: ¿refundación o refundición? (Modelos y enseñanzas desde el Derecho comparado)”, *Revista de Derecho Civil*, vol. 2, núm. 1, 2015, pp. 105-151; S. Cámara Lapuente (dir.), *Comentarios a las normas de protección de los consumidores: Texto refundido (RDL 1/2007) y otras leyes y reglamentos vigentes en España y en la Unión Europea*, Editorial Constitución y Leyes, Colex, 2011; F. De Castro y Bravo, *Las condiciones generales de los contratos y la eficacia de las leyes*, Madrid, Anuario de Derecho Civil, 1961, pp. 295 a 342; L. Díez-Picazo (coord.), *Las condiciones generales de la contratación y cláusulas abusivas*, Civitas, Madrid, 1996; L. Díez-Picazo, *Los principios del Derecho Europeo de los Contratos*, Civitas, Madrid, 2002; G. Ferri, “Condizioni generali di contratto e tutela del contraente debole”, *Atti della tavola rotonda tenuta presso l’Istituto di Diritto Privato della Università di Catania*, Milano, 1970; M. García Amigó, “Condiciones generales de los contratos”, *Revista de Derecho Privado*, Madrid, 1969; C. Martínez De Aguirre, “Transcendencia del principio de protección de los consumidores en el Derecho de obligaciones”, *Congreso internacional sobre reforma del Derecho contractual y protección de los consumidores*, Zaragoza, 1993; J.M. Miquel González,

Sin embargo, tal y como desarrolla la STS de 8 de septiembre de 2014, el avance en los planteamientos metodológicos seguido nos demuestra que esta visión parcial o fragmentada del fenómeno debe ser también corregida.

En este sentido, debe señalarse que el control de abusividad, bien en su vertiente de control de contenido, o bien en su proyección de control de transparencia, no opera en el vacío o de un modo autónomo, sino que responde a un previo y necesario presupuesto de imbricación contractual que justifica su aplicación, esto es, a unos especiales deberes de configuración contractual que incumben al predisponente y que en el caso del control de transparencia deben quedar plasmados en orden a la comprensibilidad real de los aspectos básicos de la relación negocial que reglamenten las condiciones generales; tal y como reza la citada sentencia.

Esta precisión, correctamente analizada, constituye una exigencia lógico-jurídica que se deriva tanto de la unidad conceptual del fenómeno, de su calificación de auténtico modo de contratar y, en suma, del debido desarrollo y concreción jurídica de la aplicación nuclear del principio de buena fe en este marco del sistema legal referido a la contratación seriada.

Por lo demás, y con las matizaciones necesarias, debe señalarse que estos procesos de imbricación contractual constituyen en la actualidad un claro exponente de la reciente doctrina jurisprudencial del Tribunal Supremo en aras a perfilar la proyección contractual de

“Condiciones generales abusivas en los préstamos hipotecario”, *Revista jurídica Universidad Autónoma de Madrid*, núm. 27, 2013, pp. 223-255, J.M. Miquel González, “Algunos aspectos del control de contenido de las condiciones generales en la Ley española”, *Revista española de seguros: Publicación doctrinal de Derecho y Economía de los Seguros privados*, núm. 117, 2004, pp. 7-46; F.J. Orduña Moreno, *Contratación y consumo*, Tirant lo Blanch, Valencia, 1998; F.J. Orduña Moreno, “Derecho de la contratación y condiciones generales”, *RDP*, 2000-1, núms. 4 y 5; F.J. Orduña Moreno, *Contratación y servicios financieros*, Tirant lo Blanch, Valencia, 2000; F.J. Orduña Moreno, *Contratación bancaria*, Tirant lo Blanch, Valencia, 2001; E. San Juan y Muñoz, *Guía sobre contratación bancaria de productos complejos*, Madrid, 2015; C. Vattier Fuenzalida, *Las cláusulas abusivas en los contratos de adhesión*, Burgos, 1994; E. Vázquez De Castro, *Ilícitud Contractual. Supuestos y Efectos*, Tirant lo Blanch, Valencia, 2003; M. Wolf y W.F. Lindacher, *AGB-Gesetz Kommentar*, Munich, 1994.

determinadas figuras y, con ella, su incidencia en el plano de la responsabilidad contractual.

Esta relación o conexión del control de transparencia con la estructura contractual que presenta el fenómeno de las condiciones generales no sólo interesa a la lógica del planteamiento metodológico, sino también al desarrollo mismo del control proyectado, del que pueden extraerse algunas consideraciones en torno a su alcance.

En este sentido, y en primer lugar, desde señalarse que este especial deber de transparencia que incumbe al predisponente en la configuración de las condiciones generales, como proyección cualificada del principio de buena fe en la contratación seriada, responde a una obligación principal del contrato en orden a la validez y eficacia de la cláusula predispuesta; de forma que su vulneración o incumplimiento reporta la sanción de la ineficacia contractual de la misma.

En segundo lugar, y estrechamente relacionado con la anterior consideración, debe señalarse que la carga de la prueba respecto al cumplimiento de este deber corresponde al predisponente en el marco o curso de la reglamentación llevada a cabo. En el caso resuelto por la Sentencia de 8 de septiembre de 2014, el Tribunal Supremo puntualiza que este deber no resultó acreditado por las entidades predisponentes pues el alcance de las cláusulas suelo no formó parte de las negociaciones y tratos preliminares, ni tampoco resultó destacado y diferenciado específicamente, ni en el seno de la oferta comercial realizada, ni en el contexto de las escrituras públicas de los préstamos hipotecarios; en donde su referencia se llevó a cabo sin resalte o diferenciación alguna dentro de la cláusula genérica rubricada, significativamente, en atención al “interés variable” del préstamo suscrito.

Por último, y en tercer lugar, debe señalarse que como obligación principal, que incumbe al predisponente en orden a la misma configuración contractual de la reglamentación predispuesta, este deber de transparencia o comprensibilidad real de las condiciones generales debe inferirse del desarrollo que presente la propia reglamentación predispuesta; en donde debe acreditarse que el predisponente articuló los criterios precisos y comprensibles para que el consumidor adherente pueda evaluar, directamente, las consecuencias económicas derivadas a su cargo. De ahí, como señala la citada Sentencia de 8 de septiembre de 2014 que, con carácter general, y sin perjuicio de la im-

portante función preventiva que los notarios realicen sobre el control de estas condiciones, debe advertirse que la mera lectura de la escritura pública y, en su caso, el contraste de las condiciones financieras de la oferta vinculante con la del respectivo préstamo hipotecario, no suplen, por ellas solas, sin protocolo o actuación específica al respecto, el cumplimiento de este deber especial de transparencia.

El cumplimiento de estos especiales deberes de configuración negocial que incumben al predisponente no se realiza desde el plano tradicional de los criterios interpretativos que ofrece el contrato por negociación. En este ámbito, el cumplimiento, en sentido general, se presenta como un juicio de contraste de la ejecución del programa prestacional llevado a cabo por las partes, cuya configuración negocial no es puesta en duda y, por tanto, resulta plenamente válida y eficaz.

Por contra, en el fenómeno de las condiciones generales, como modo diferenciado de contratar, el cumplimiento de estos especiales deberes de transparencia afecta a la validez e idoneidad del marco negocial predispuesto, de forma que su valoración encierra, como se ha destacado, un auténtico juicio de eficacia contractual en orden a la validez e idoneidad de la cláusula predispuesta.

Desde esta perspectiva, el control de transparencia, como proyección natural o extensión del control de abusividad, participa, al igual que el control de contenido, de la caracterización del control de legalidad que acompaña al fenómeno de la contratación seriada, diferenciándose, de igual modo, por su naturaleza y alcance, de los fenómenos de la interpretación integrativa y de la integración del contrato que se infieren del proceso interpretativo del contrato por negociación.

Esta es la línea expresamente seguida por la doctrina jurisprudencial del Tribunal Supremo a la hora de establecer la caracterización del control de transparencia.

En efecto, la citada Sentencia de 8 de septiembre de 2014 sistematiza dicha caracterización en los siguientes criterios:

- a) Su fundamento en el marco del específico y diferenciado presupuesto causal que informa el régimen de validez y eficacia del fenómeno de las condiciones generales de la contratación.

- b) Su proyección, con base al principio de buena fe, en un previo y especial deber contractual de transparencia real que incumbe al predisponente y que debe quedar plasmado en la reglamentación predispuesta.
- c) Su aplicación como un control de legalidad a tales efectos (artículo 5 de la Directiva 93/13, artículos 5.5 y 7.b. de la LCGC y artículo 80.1.a. del TR LGDCU).

Conviene destacar que esta caracterización, que se desprende no sólo de la doctrina jurisprudencial del Tribunal Supremo, sino también como se ha señalado del análisis metodológico con arreglo al desarrollo analítico de los Principios generales del derecho, particularmente de su conexión con el desenvolvimiento actual de las directrices de orden público económico y su entronque con el principio de buena fe, así como del moderno derecho de la contratación europea y, en suma, de las diferencias y peculiaridades que presenta la contratación seriada como modo propio y diferenciado de la contratación, responde a una plena construcción jurídica de la figura objeto de estudio que disipa las dudas u objeciones planteadas por la tesis que, con carácter general, reconducen el alcance de este control de transparencia bien a la mera transparencia formal o documental de la cláusula en cuestión, o bien a la simple inteligencia gramatical que se deriva de su formulación (como reconducción al artículo 1281, párrafo primero, del Código Civil).

Al respecto, debe rechazarse el postulado que sostiene que la caracterización del control de transparencia, llevada a cabo por la doctrina jurisprudencial del Tribunal Supremo, constituye una creación jurisprudencial del derecho que vulnera el sistema de fuentes al establecer un criterio de transparencia, el de la comprensibilidad real, que no se encuentra comprendido en el estándar de transparencia del artículo 80.1 y la expresión de buena fe del artículo 82.1 del TRLGDCU.

Este planteamiento, como ya se ha dicho, aparte de desatender el necesario aspecto metodológico que presenta el análisis de la naturaleza y función del fenómeno de las condiciones generales, es representativo de una interpretación normativa parcial y sesgada que debe rechazarse por varias razones.

En síntesis, de la caracterización que dispensa la Sentencia de 8 de septiembre de 2014 no puede afirmarse, con rigor metodológico, que

se incurra en una creación judicial del derecho como propugnaba, a principios del siglo pasado, la “Escuela del derecho libre”, a través de la libre interpretación e investigación del Derecho. Por el contrario, como claramente se advierte, la doctrina jurisprudencial expuesta no tiene por objeto la libre decisión de lo justo en un caso concreto, sino el establecimiento de una caracterización general del control de transparencia aplicable a todos los casos que se deriven de este modo de contratar. Del mismo modo, de su articulación, como se señala, resulta conforme con la convergencia del Derecho contractual europeo. Aspectos o extremos que, analizados con cierto rigor, dan buena muestra de cómo la doctrina del Tribunal Supremo se aparta de la corriente metodológica indicada que es la que, precisamente, le imputan u objetan estas tesis disconformes con la misma.

Pero, además, también hay que señalar conforme al rigor técnico exigible, que la caracterización llevada a cabo por la doctrina jurisprudencial no se realiza fuera de la ley, si por tal entendemos el actual sistema de fuentes que cabe aplicar. En efecto, contrariamente a su simple asimilación a los deberes formales de la legislación sectorial, la caracterización expuesta refuerza su valor normativo al hacer descansar el fundamento de estos especiales deberes de transparencia real en el propio presupuesto causal del fenómeno que informa a la legislación especial aplicable y, con ella, la especial tutela del adherente-consumidor. De modo que el estándar de transparencia aludido, lejos de una interpretación literal de los preceptos en liza, encuentra su fundamento en la relevancia del principio de buena fe como fuente de especiales deberes de conducta; en el presente caso, como especiales deberes de configuración contractual del marco predispuesto.

Como puede observarse, fundamentos de índole causal y teleológica de la norma que alejan, aún más, la caracterización realizada del ámbito de la creación judicial del derecho.

En definitiva, también conviene resaltar, por si alguna duda quedara al respecto, que la doctrina jurisprudencial del Tribunal de Justicia Europeo concuerda tanto en el contenido como en el procedimiento con la caracterización dada por el Tribunal Supremo.

En este sentido la STJUE de 30 de abril de 2014 (apartado 60 a 75) declara como en relación al alcance interpretativo de la Directiva del 93/13 el único límite radica, precisamente, en la posición que pu-

dieran presentar la aplicación de los Principios generales del derecho, de suerte que la consideración de la directiva no podría servir de base para una interpretación *contra legem* del derecho nacional. Límite, o cuestión interpretativa, que sólo puede afectar a aquellos ordenamientos jurídicos que no contemplen específicamente esta exigencia de redacción clara y comprensible de la reglamentación predispuesta; pero no a los que ya la tienen incorporada a su legislación, como es el caso de nuestro ordenamiento.

De todas formas, ya resulte por la no oposición a los Principios generales del derecho, o bien por la previsión normativa al respecto, el alcance del control de transparencia, a la luz de la letra de la finalidad de esta directiva, está fuera de toda duda, esto es, tal y como señala la sentencia no se limita al plano formal o gramatical de la cláusula, sino que alcanza a su comprensibilidad real; estableciéndose una identidad de alcance de este plano de comprensibilidad real tanto en lo dispuesto en el artículo 4, apartado 2 de la Directiva, como en la formulación del artículo 5 de la misma, sin posibilidad de distinción alguna al respecto.

Como puede observarse, una interpretación claramente extensiva de la exigencia de transparencia que nuestro Tribunal Supremo no sólo ha seguido, sino que, además, ha dado las pautas necesarias para su pleno entendimiento en la dogmática contractual.

Por último, y en el contexto de las objeciones a la doctrina jurisprudencial expuesta, tampoco puede aceptarse que el alcance de este control de transparencia, así configurado, venga ligado, o pueda ser reconducido al tratamiento de las denominadas “cláusulas sorpresivas”, particularmente porque su pretendida previsión normativa decayó durante la discusión parlamentaria del Proyecto de la Ley de condiciones generales de la contratación.

Al margen de que su no previsión resulta criticable, (piénsese, que dicha regla se encuentra en las principales legislaciones comparadas), su equiparación con el control de transparencia es un absoluto despropósito conceptual, pues dicha regla no cumple la función del control de transparencia y se limita, en su caso, a servir como criterio de incorporación de la cláusula predispuesta. El hecho de que falte su previsión normativa para nada afecta al alcance del control de transparencia que exigirá, con mayor incidencia, si cabe, la comprensibi-

lidad real de la cláusula sorprendente o insólita que el predisponente haya empleado.

En este contexto, debe resaltarse que la Sentencia de 8 de septiembre de 2014 establece un planteamiento metodológico coherente con el espíritu y el objeto de la Directiva 93/13, al contemplar el control de transparencia como una extensión lógica del control de abusividad en orden a la idoneidad de la reglamentación predispuesta y por tanto, a la eficacia resultante de la misma. Todo ello, bajo la proyección del principio de buena fe, como fuente de creación de especiales deberes de configuración contractual, en este ámbito de la contratación que es calificado como un modo propio y diferenciado de contratar respecto del contrato por negociación.

La Sentencia del Tribunal Supremo de 18 de junio de 2012 fue pionera a la hora de establecer el contenido u objeto que presenta el control de transparencia, esto es, la finalidad y función que lo guía y le asiste.

Este contenido, tal y como describe la posterior Sentencia del Tribunal Supremo de 8 de septiembre de 2014, en armonía con la doctrina jurisprudencial del Tribunal de Justicia de la Unión Europea, (STJUE de 30 de abril de 2014. C26/13), responde, necesariamente, a la propia caracterización del control de transparencia que, como control de legalidad, se dirige a comprobar, primordialmente, que la cláusula en cuestión en el conjunto de la reglamentación predispuesta, refiere directamente una comprensibilidad real, que no meramente formal de los aspectos básicos del contrato. De forma que el consumidor y usuario conozca y pueda evaluar directamente las consecuencias jurídicas que, de acuerdo con el producto servicio ofertado, resulten a su cargo tanto respecto de la onerosidad o sacrificio patrimonial que realmente le suponga o pueda suponer la ejecución del contrato celebrado, como la posición jurídica que también realmente asume en los aspectos básicos que se deriven de la naturaleza y de la ejecución del mismo.

Queda claro, por tanto, por más que algunos se empeñen de lo contrario, que el objeto del control de transparencia, por imperio de su propia configuración normativa, no puede quedar limitado o ser reconducido a la comprobación de la mera transparencia formal o documental que desarrolle la normativa sectorial aplicable, sin per-

juicio de que dicha normativa deba ser tenida en cuenta a la hora de efectuar dicho control en el ámbito judicial.

Lo contrario, a tenor de la doctrina jurisprudencial expuesta, supone desnaturalizar “la ratio” o “médula” de la protección dispensada al consumidor, además de una clara vulneración de su derecho la tutela judicial efectiva (artículo 24 CE), esto es, a que judicialmente se le aplique el control de transparencia a la concreta reglamentación predispuesta que dio lugar al contrato celebrado.

Del mismo modo, este control tampoco puede quedar reconducido a un mero contraste de la inteligencia gramatical de la cláusula en cuestión, en sí misma consideraba, que resulte ajeno a los planos de comprensibilidad real anteriormente expuestos.

En esta línea, debe señalarse, de forma muy positiva, el avance que ha supuesto para la comprensión del proceso interpretativo, en general, la precisión del carácter “meramente instrumental” que presenta la denominada interpretación literal del contrato conforme a la reciente doctrina jurisprudencial del Tribunal Supremo, especialmente a tenor de la STS de 29 de enero de 2015 (núm. 27/2015).

Si el control de transparencia presenta este contenido es porque previamente, como se ha señalado, al predisponente, en este particular modo de contratar, le incumben unos especiales deberes de facilitar la comprensibilidad real de la reglamentación predispuesta. Deberes que obtienen su fundamento o fuente de creación en la propia aplicación del principio de buena fe en la contratación seriada determinando, de esta forma, la delimitación última del cumplimiento obligacional de estos especiales deberes de transparencia contractual.

Para la doctrina jurisprudencial la determinación básica en el cumplimiento de estos deberes viene configurada en orden a la inclusión, por el predisponente, de los criterios precisos y comprensibles que permitan, a su vez la evaluación o consideración directa del adherente respecto de las principales consecuencias económicas y jurídicas que se deriven a su cargo de la reglamentación predispuesta.

Con carácter general, se puede afirmar que estos “criterios de comprensibilidad”, que el predisponente debe articular, pueden concretarse en los siguientes planos de la contratación seriada:

A) En primer término, los criterios de precisión deben ser tenidos en cuenta cuando la cláusula en cuestión pueda afectar a la especifici-

dad o particularidad de la oferta comercial del producto o servicio de que se trate, de suerte que deba formar parte de la misma de un modo resultando o diferenciado.

En el caso de las denominadas cláusulas suelo, la Sentencia del Tribunal Supremo de 8 de septiembre de 2014 destaca este plano de concreción en los siguientes términos en el Fundamento de Derecho segundo, apartado nueve:

“9. La doctrina jurisprudencial expuesta, llevada al supuesto de enjuiciamiento, exige realizar las siguientes precisiones que conducen a la estimación del motivo planteado. En primer lugar, excluido el carácter negociado de la cláusula suelo, el análisis del presente caso se dirige a valorar si, conforme a la naturaleza y caracterización que se ha realizado del control de transparencia, el predisponente cumplió con el especial deber de comprensibilidad real de dicha cláusula en el curso de la oferta comercial y de la reglamentación contractual predispuesta. En este sentido, atendido el marco de la contratación realizado, no se observa que el predisponente incluyera los criterios precisos y comprensibles en orden a que los prestatarios pudieran evaluar, directamente, el alcance jurídico de la cláusula suelo respecto a la modulación de la oferta comercial que se realizaba. En efecto, fuera del debate acerca de si la denominada cláusula suelo (sujeción a un interés mínimo) desnaturaliza o no el concepto de interés variable, lo cierto es que, a los efectos del principio de transparencia real, constituye un elemento significativo en la modulación o formulación básica de la oferta de este tipo de contratos, que debe ser objeto de un realce específico y diferenciable. En el presente caso, esto no fue así pues el alcance de las cláusulas suelo no formó parte de las negociaciones y tratos preliminares que se llevaron a cabo, ni tampoco resultó destacado y diferenciado, específicamente, ni en el marco de la oferta comercial realizada, ni en el contexto de las escrituras públicas de los préstamos hipotecarios, objeto de estudio, en donde su referencia se realiza sin resalte o especificidad alguna, dentro de una cláusula más amplia y extensa rubricada, significativamente, en atención a la regulación del “interés variable” del préstamo.

Al respecto, también resulta significativo que la parte recurrida, fuera de probar los anteriores extremos en el curso de la reglamentación predispuesta, descargue el cumplimiento de su propio deber de transparencia en los protocolos notariales de los contratos celebrados. En este sentido debe señalarse, sin perjuicio de la importante función preventiva que los Notarios realizan sobre el control previo de las condiciones generales de la contratación que, conforme a la caracterización y alcance del control de transparencia expuesto, la comprensibilidad real debe inferirse del propio juego o desarrollo de la reglamentación predispuesta, de forma que la

lectura de la escritura pública y, en su caso, el contraste de las condiciones financieras de la oferta vinculante con la del respectivo préstamo hipotecario, no suplen, por ello solos, sin protocolo o actuación específica al respecto, el cumplimiento de este especial deber de transparencia”.

En segundo lugar, una vez que ha quedado excluido el cumplimiento, por parte del predisponente, del deber de transparencia en el propio curso de la oferta y de la reglamentación predispuesta cabe plantearse, en su caso, si este control queda acreditado en el ámbito de la “transparencia formal o documental” que acompaña a este modo de contratar, particularmente del documento en donde se contempla la llamada oferta vinculante. Al respecto, la respuesta debe ser también negativa pues el citado documento sigue el mismo esquema formal de las escrituras públicas analizadas en donde la cláusula suelo, referida a un “tipo mínimo anual”, queda encuadrada en el apartado correspondientemente rubricado con referencia excluida al “tipo de interés variable” (condición 3 bis de la oferta), sin mayor precisión y comprensibilidad de su alcance o relevancia y en un contexto caracterizado por la abundancia de datos y formulaciones bancarias, ausente, por otra parte, de simulaciones de escenarios diversos relacionados con el comportamiento razonablemente previsible del tipo de interés mínimo en el momento de la contratación; criterios, todos ellos, tenidos en cuenta por esta Sala en el caso similar que dio lugar a la Sentencia de 9 de mayo de 2013”.

B) En segundo término, y de forma sistemática con lo anteriormente expuesto, los criterios de precisión también deben articularse por el predisponente cuando la complejidad de la cláusula en cuestión requiera un detalle preciso y claro de su incidencia en la dinámica o ejecución del contrato.

Este es el caso de los denominados “préstamos hipotecarios multi-divisa”, en donde la Sentencia del Tribunal de Justicia de 30 de abril de 2014 exigió que el predisponente también expusiera de manera clara y comprensible, el concreto funcionamiento del mecanismo de conversión de la divisa extranjera y su incidencia en el préstamo realizado, esto es, tanto en la posible variación de la cuota, como del capital prestado que debe ser devuelto.

C) Por último, también debe puntualizarse que estos planos de precisión pueden exigirse tanto para la comprensión de las conse-

cuencias económicas de la reglamentación predispuesta, esto es, el sacrificio patrimonial que debe realizar el adherente, como para la comprensión de las consecuencias jurídicas que se deriven, es decir, la posición jurídica que realmente asume el adherente con la adquisición del producto o la realización del servicio ofertado. Caso paradigmático de las denominadas “acciones preferentes”, particularmente, de la distinción entre las entidades que las comercializaron y el verdadero titular de la emisión y, en consecuencia, de la valoración de su posible solvencia.

En todo caso, también conviene tener en cuenta, con relación a la concreción del control de transparencia, las siguientes consideraciones:

- La falta de información precontractual no puede ser suplida o subsanada por la mera obligación futura del predisponente de facilitar dicha información en la propia ejecución del contrato, STJUE de 21 de marzo de 2013.
- Cualquier modificación del contrato debe facilitarse en tiempo útil para la toma de decisión del consumidor. Igualmente, dicha modificación debe resultar comprensible y transparente, STJUE de 23 de octubre de 2014.

2. Transparencia en los servicios de intermediación en línea

Por otra parte, el Reglamento (UE) 2019/1150 del Parlamento Europeo y del Consejo, de 20 de junio de 2019, sobre el fomento de la equidad y la transparencia para los usuarios profesionales de servicios de intermediación en línea, realmente desconocido, pese a ser una norma directamente aplicable, regula las normas de transparencia entre los «usuarios profesionales», que son todo particular que actúa en el marco de una actividad comercial o profesional o toda persona jurídica que ofrece bienes o servicios a los consumidores a través de «servicios de intermediación en línea» con fines relativos a su comercio, negocio, oficio o profesión y los propios prestadores de servicios de intermediación en línea.

Pues bien, el artículo 3 establece las condiciones de validez de las condiciones generales que los prestadores de servicios de intermediación imponen a los usuarios profesionales, estableciendo que las

cláusulas sean claras, sencillas y comprensibles y que se establece con carácter previo y equitativo; especialmente las relativas a las causas objetivas en las que se basarán las decisiones de suspender, terminar o restringir de cualquier otro modo, de manera total o parcial, la prestación de los servicios de intermediación en línea a los usuarios profesionales.

Por tanto, hablamos de transparencia material entre empresarios, lo que supone una superación de la doctrina actual del Tribunal Supremo, que ha reducido y limitado el control de transparencia a las relaciones entre empresario y consumidor. Es más, la doctrina del Supremo debe asumir que, en la moderna contratación del Derecho de la Unión Europea, el control de transparencia y el control de contenido, relativo a la abusividad de las cláusulas predispuestas, también se da en las relaciones entre empresarios, especialmente en los contratos predispuestos entre los prestadores de servicios de intermediación en línea y los usuarios profesionales.

3. Protección frente a los patrones oscuros “dark patterns”

El Comité Europeo de Protección de Datos (CEPD) elaboró y publicó las Directrices 3/2022 sobre patrones oscuros (*dark patterns*) en las redes sociales.

Aunque ha sido la *Californian Consumer Privacy Rights Act* (CPRa), la primera en establecer explícitamente que no se entenderá válidamente prestado el consentimiento si este se obtuvo como consecuencia del uso de patrones oscuros.

Pero los patrones oscuros, aunque han sido muy habituales en materia de *cookies*, no afecta exclusivamente al ámbito de la protección de datos, sino que se proyectan a todos los ámbitos de la contratación predispuesta a través de medios electrónicos⁴.

⁴ M. Durovik, *European Law on Unfair Commercial Practices and Contract Law. Vol. Modern studies in European Law*, Bloomsbury Publishing, Forbrukerrådet, Londres, 2018; C.M. Gray, Y. Kou, B. Battles, J. Hoggat, y A.L. Tooms, “The Dark (Patterns) Side of UX Design” *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, AC, Nueva York; A. Vidal García, “La prohibición de la publicidad engañosa”, en *La promoción de medicamentos*

Los patrones oscuros son aquellas técnicas que influyen en el comportamiento de los usuarios, llevándolos a tomar decisiones involuntarias y potencialmente perjudiciales en relación con la emisión de consentimientos y la aceptación de condiciones generales predispuestas, para conseguir, mediante ciertas técnicas manipuladoras o fraudulentas, el consentimiento que quiere el predisponente y que el usuario se ve consciente o inconscientemente obligado a aceptar, mediante el diseño de algoritmos informáticos diseñados con tal fin.

Esta institución se relaciona en cierto modo con el dolo en el consentimiento, recogido en el artículo 1269 del Código Civil, según el cual, existe dolo (en el consentimiento) cuando, con palabras o maquinaciones insidiosas de parte de uno de los contratantes, es inducido el otro a celebrar un contrato que, sin ellas, no hubiera hecho. Es cierto que el usuario sí quiere contratar el servicio, pero no desea hacerlo con la cesión de datos, o con la modalidad de *cookies* que el predisponente ha diseñado, y en ese ámbito de la cesión de datos o de las *cookies* el consentimiento está viciado, y por tanto podría ser anulado. Evidentemente si el patrón oscuro se proyecta sobre el producto o servicio contratado, el efecto restitutorio inherente a la anulación del contrato podría incidir en dicho producto o servicio.

Dicho esto, existen distintos modelos y categorías de patrones oscuros, de los cuales también se han hecho diversos listados, pero evidentemente se trata de un listado ejemplificativo o “*numerus apertus*”.

Así existen “patrones oscuros en la interfaz”, como el modo en que se muestra la información o la forma de interactuar con ella, y “patrones oscuros de contenido”, como la redacción o el contexto.

A modo de ejemplo, podemos indicar los siguientes patrones oscuros.

Overloading: es la sobrecarga de información y opciones con el objetivo de que el usuario comparta más datos o permita inconscientemente el procesamiento de sus datos personales en contra de sus expectativas, aludiendo en muchas ocasiones a una falsa necesidad de recopilar ciertos tipos de datos basándose en el buen funcionamiento

del sistema o en una mejor experiencia, lo que incide en los principios de minimización de datos, transparencia, equidad y responsabilidad proactiva.

Ejemplos de ello los encontramos cuando se solicita concurrentemente un dato (como la agenda de contactos o nuestro número de teléfono) cada vez que avanzamos de pantalla a pesar de habernos negado en todas las ocasiones, con la única finalidad de ejercer presión sobre nosotros para que lo terminemos entregando. También hay quien enmascara la recogida de datos bajo legitimaciones erróneas, abusando del uso del consentimiento, en contra de lo dispuesto en el artículo 6 RGPD.

Otro ejemplo es cuando aparecen demasiadas opciones para elegir, aspecto que difiere de la especificidad que requiere el consentimiento, porque en este caso las mismas son irrelevantes, pues lo que se busca es provocar que el usuario se pierda manteniendo la opción por defecto, lo que en cierto modo conecta con el siguiente patrón.

Skipping: diseñado para que el usuario se olvide de la protección de sus datos. Esto se encuentra estrechamente vinculado a los principios de protección de datos por defecto y por diseño.

Stirring: es el uso de las emociones contra los propios intereses. En estos casos, y a pesar de que al inicio de las Directrices ya el CEPD muestra sus preocupaciones respecto al uso de patrones oscuros en menores, el impacto que producen este tipo concreto de tácticas se ven acentuadas por la plasticidad y la falta de comprensión a estas edades, elevando el riesgo para sus derechos y libertades, lo que podría derivar en daños físicos o psicológicos.

Ejemplos de ello los encontramos en el uso de frases e imágenes que acentúan la necesidad de terminar cuanto antes para empezar a utilizar la plataforma o servicio.

Hindering: pretenden dificultar la búsqueda de información sobre el procesamiento de los datos o administrar los controles una vez activados. Un caso típico es cuando se hace clic sobre el botón “omitir” y nos preguntan si estamos seguros, mientras que cuando se hace clic en aceptar no se hace reconsiderar la elección, lo que implica un desequilibrio en el tratamiento, pues sólo pide confirmación en uno de los casos (aquel que el predisponente no desea).

Fickle: es el uso del diseño de la interfaz para impedir acceder a las herramientas de configuración, bien porque no hay un enlace de redirección disponible o porque este no funcione. También suelen verse del diferente tamaño de la fuente de las letras o el uso de colores o destacado que no contrastan lo suficiente con el fondo como para ofrecer una legibilidad, yendo la vista del usuario a la opción que el predisponente quiere que acepte el usuario.

Left in the dark: la interfaz está diseñada para ocultar información haciendo uso de términos ambiguos, la extensión de la redacción o la aportación de información contradictoria que desorienta al usuario sobre lo que debe hacer, por lo que se suele dejar la configuración ya predeterminada por el predisponente.

Dentro de esta categoría también se encuentra la discontinuidad del idioma, que puede confundir a los usuarios incluso cuando estos confirman su capacidad para utilizar una lengua extranjera. Respecto a este punto, la AEPD ya ha tenido ocasión de pronunciarse, señalando que “el artículo 12.1 del RGPD requiere que se proporcione cualquier información a la que se refieren el artículo 13 y el artículo 14 del RGPD de forma concisa, transparente, inteligible y de fácil acceso, utilizando un lenguaje claro y sencillo”.

Como puede comprobarse, el uso de patrones oscuros resulta incompatible con los principios de transparencia e información, pues los interesados deben recibir información de manera clara y perceptible para que puedan comprender cómo se procesan y controlan sus datos.

Vinculado a esto último, el CEPD ha hecho referencia a las dificultades que encuentran los usuarios a la hora de borrar su cuenta o cancelar sus datos, lo que incluiría el derecho de supresión del Artículo 17.1.a) RGPD y la retirada del consentimiento del Artículo 7.3 RGPD. Entre las vicisitudes nos encontramos con la reafirmación de nuestras intenciones a lo largo de múltiples pantallas o la imperiosa necesidad de alegar motivos al cierre de la cuenta, obligándonos a escribir un número mínimo de palabras con la “finalidad” de mejorar la plataforma.

Pero el CEPD no sólo indica los problemas, sino que también enumera una serie de buenas prácticas, como son el uso de accesos directos a la información, índices o desplegables en las políticas de priva-

cidad, jerarquizar las opciones de forma homogénea o proporcionar definiciones cuando se utilice lenguaje técnico. También respecto al momento de borrado de la cuenta apunta que sería interesante que las plataformas ofrecieran en ese momento un medio para descargar nuestros datos.

Pues bien, toda la protección frente a los patrones oscuros está proyectada por la Unión Europea en un Reglamento, que se encuentra formulado actualmente como “Propuesta de Reglamento” (en concreto, la Propuesta de Reglamento de 15 de diciembre de 2020 COM (2020) 825 final) relativo a un mercado único de servicios digitales (Ley de Servicios Digitales o *Digital Services Act*), y en el que se ha incluido una regulación protectora frente al uso de patrones oscuros.

V. CONCLUSIONES

Podemos concluir señalando que la contratación a través de internet exige normas uniformes, y que dicha uniformidad se proyecta ahora sobre las llamadas condiciones generales de la contratación, cuyos excesos por la parte predisponente deben de ser corregidos frente a las cláusulas abusivas, y frente a la omisión o falta de la debida transparencia en la contratación, abusos que no solo de dan en las relaciones entre empresarios y consumidores (B2C), sino que también se da en las relaciones entre empresarios (B2B), como la que acontece entre los usuarios profesionales y los prestadores de servicios de intermediación en línea, regulados por el Reglamento 2019/1150 de la Unión Europea que es desconocido por la mayoría de los juristas, y que evidencia otro problema como es el desconocimiento de la normativa europea que se promueve a través del reglamento, a pesar de que dicho Reglamento es norma directamente aplicable a los Estados miembro.

Entre las principales novedades en materia de condiciones generales de la contratación y transparencia sobresale la identificación de los llamados “patrones oscuros”, con sus correspondientes sanciones tanto en el ámbito administrativo como en el ámbito del derecho civil (a través de las pertinentes acciones de anulación y de responsabilidad), lo que evidencia la relevancia que está adquiriendo la trans-

parencia contractual en la contratación realizada a través de redes sociales e Internet.

Transparencia algorítmica en el ámbito laboral

ADRIÁN TODOLÍ SIGNES¹

*Prof. Titular de Derecho del Trabajo y de la Seguridad Social
Universitat de València*

I. EL USO DE LOS ALGORITMOS EN EL TRABAJO

En la era de la información, gran parte del trabajo de los especialistas en recursos humanos consiste en recopilar la máxima información sobre el trabajador para mejorar de esta manera la toma de decisiones que deben adoptar (contratación, ascensos, despidos, incremento de horas de trabajo, movilidad geográfica, pago de bonus salariales, etc.)². En efecto, es crucial para la empresa reunir y compilar el máximo entendimiento posible sobre las habilidades, conocimientos, aptitudes, actitudes, etc., del trabajador para, de esta forma, poder adoptar las decisiones más convenientes en interés de la empresa³.

De la misma manera que una empresa desea conocer la mayor información posible sobre los consumidores para saber qué producto ofrecerle o qué anuncio publicitario exactamente conseguirá convencerle de que compre su producto, los empresarios desean obtener la máxima información sobre sus trabajadores para saber si serán pro-

¹ www.adriantodoli.com. Estudio en el marco de los proyectos, MICINN Retos “Derechos y garantías frente a las decisiones automatizadas...” (RTI2018-097172-B-C21), proyecto “Algorithmic law” (Prometeo/2021/009, 2021-24) Generalitat Valenciana”.

² L. Gensing-Pophal, “Recruiting 2.0”, *Credit Union Management Magazine*, 01/09/2009, p. 42; S. Sameen, y S. Cornelius, “Social networking sites and hiring: How Social Media profiles influence hiring decisions”, *Journal of Business Studies Quarterly*, 7 (1), 2013, p. 1.

³ P. C. Jackson Lewis, “Employee monitoring and workplace privacy law”, *American Bar Association, National Symposium on Technology in Labour & Employment Law*, 2016. https://www.americanbar.org/content/dam/aba/events/labour_law/2016/04/tech/papers/monitoring_ella.authcheckdam.pdf

ductivos, encajarán en el ambiente de la empresa o qué le motivará concretamente a permanecer o esforzarse más⁴.

El denominador común de estas situaciones (la reputación online, los *werables*, las videocámaras, etc.) es la recopilación de información. Una actividad —la recopilación de información para tomar mejores decisiones— que ha existido desde siempre⁵. En efecto, desde hace años las empresas realizan entrevistas en el proceso de selección, dinámicas de grupo, evaluaciones del desempeño, etc., con objeto de tomar estas decisiones relacionadas con el trabajo. No obstante, desde hace unos años los expertos en recursos humanos se han especializado en la recogida de información mediante el uso de las nuevas tecnologías (*Facebook*, *LinkedIn* y ahora reputación online)⁶. Ciertamente, la tecnología ha permitido que la empresa pueda acceder a mayor volumen de datos de forma muy barata, no obstante, siguió siendo el responsable de recursos humanos el que, reunida esa información, debía procesarla y tomar la decisión. Esto provocaba que a pesar de aumentar el nivel de información disponible existía un límite natural —la capacidad humana del procesamiento de esos datos— para el uso de esa información.

Sin embargo, las más recientes tecnologías están cambiando esto también. Gracias a los algoritmos, el *big data* y la inteligencia artificial (también llamado *machine learning*), no solamente hay un abaratamiento del coste de acceso a la información —lo que había hasta ahora gracias a *Facebook* y *LinkedIn*— sino que también se está produciendo un abaratamiento sin precedente en el coste del procesamiento

⁴ I. Ajunwa et al., “Limitless worker surveillance”, *California Law Review*, 105 (3), 2017, pp. 102-142; P. Moore et al. (eds), *Humans and Machines at Work: Monitoring, Surveillance and Automation in Contemporary Capitalism*, Palgrave Macmillan, 2018; V. De Stefano, ““Negotiating the Algorithm”: Automation, Artificial intelligence and labour protection”, *Comparative Labor Law & Policy Journal*, 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3178233

⁵ Para conocer respecto a la historia de los métodos de vigilancia y control de los trabajadores desde los principios de la industrialización hasta ahora, ver I. Ajunwa et al., “Limitless worker surveillance”, *cit.*, pp. 107-8.

⁶ M.V. Wolf, J. Sims, y H. Ynag, “Social media utilization in human resource management”, *Web based communities and social Media 2014 conference*, Lisbon, 2014; M. Ouridi et al., “Technology adoption in employee recruitment: The case of social media in Central and Eastern Europe”, *Computers in Human Behavior*, 57, 2016, pp. 240-249.

de dicha información para que sea útil e incluso facilitando la toma de decisiones basadas en esta información (decisiones automatizadas).

De esta forma, actualmente, la introducción de nuevas tecnologías en el procedimiento de evaluación y control de los trabajadores ha modificado estos procesos en tres variables fundamentales: i) la forma de recopilar información; ii) la forma de procesar esa información; iii) la forma de tomar decisiones.

i) Aumento del volumen de información disponible. Tecnologías como la videovigilancia, el GPS, los *wearables* —pulseras que controlan las pulsaciones y el estado de atención y actividad del trabajador—, provocan un aumento de la información disponible.

De la misma forma, los sistemas de reputación online —evaluaciones por parte de los clientes— permiten obtener información sobre el comportamiento del trabajador de forma mucho más barata⁷.

ii) Aumento en la capacidad de procesamiento de esa información: En segundo lugar, existe la necesidad de procesar dicha información. Y, en este terreno, las nuevas tecnologías han supuesto un importante avance en la capacidad para llevar a cabo esa acción. Por ejemplo, en el caso de la videovigilancia hasta ahora era necesario que una persona visualizara las horas de videovigilancia para observar si había habido alguna irregularidad con el trabajador. No obstante, los sistemas de reconocimiento facial y de figuras permiten que de forma automatizada se señale si ha habido alguna irregularidad, informado exactamente del momento en el que ocurre, y abaratando el coste del control de los trabajadores.

En el caso de los *wearables*, tener una persona de recursos humanos controlando las pulsaciones de todos los trabajadores (o el posicionamiento si es a través de GPS), podría ser excesivamente costoso y, por tanto, inservible. Sin embargo, mediante los sistemas automatizados (y algoritmos) es posible, y resulta económicamente muy barato, establecer alarmas que informen al responsable de recursos humanos sobre la existencia de largos periodos de inactividad de un trabajador.

⁷ A. Thierer, C. Koopman, A. Hobson, y C. Kuiper, “How the internet, the sharing economy, and reputational feedback mechanisms solve the “Lemons Problem”, *Mercatus Working paper*, 2015, p. 7. Disponible en <https://www.mercatus.org/system/files/Thierer-Lemons-Problem.pdf>

De esta forma, no es necesario un responsable de recursos humanos que controle la información ni realice tareas de vigilancia, sino que simplemente el responsable recibirá una “alerta” de forma automática cuando se dé la situación que merece observación.

Algunas empresas en EEUU están desarrollando aparatos que llevan micrófono, no con intención de grabar las conversaciones de los trabajadores, sino con el objeto de conocer el estado de ánimo del trabajador dependiendo de su tono de voz. También este aparato permite medir las interacciones del trabajador con compañeros para saber con qué compañeros se interactúa y durante cuánto tiempo⁸.

De la misma forma, en el caso de la reputación online, la recogida de información y evaluación sobre los clientes puede ser excesivamente costosa de analizar y sistematizar, sin embargo, el sistema informatizado de puntuaciones permite categorizar la información y obtener medias y alertas cuando existe un comportamiento del trabajador que se aleja de los estándares aceptables. La reducción de la necesidad de la interacción del responsable de recursos humanos, obviamente, abarata las posibilidades de control.

iii) Capacidad de toma de decisiones automatizadas: Este es el último paso en la obtención de la máxima eficiencia en el control de los trabajadores y consiste en la desaparición de toda intervención humana. De esta forma, la inteligencia artificial se convierte en sustituto del responsable de recursos humanos incluso en la adopción de decisiones.

En definitiva, el abaratamiento de estos tres niveles en la evaluación implica necesariamente un aumento del control de los trabajadores. Es decir, cuanto más barato sea realizar el control, más medidas pondrá en marcha el empresario para proteger sus legítimos intereses empresariales. Actualmente, el art. 20 ET concede al empresario la potestad para elegir las formas de vigilancia y control del trabajador que estime convenientes. Sin embargo, esta previsión se promulgó en un momento en el que la vigilancia y el control estaban limitados por su propia naturaleza —en pocas palabras, porque era cara—.

⁸ *The Rise of Workplace Spying*, WEEK 5 de julio de 2015 <http://theweek.com/articles/564263/rise-workplace-spying>

A diferencia de lo que se cree comúnmente, la tecnología pocas veces permite un control que antes no era, en absoluto, posible. Lo que permite la tecnología es abaratar ese control. En efecto, un mayor número de vigilantes en una tienda, producen el mismo efecto, respecto al control de los trabajadores, que múltiples videocámaras instaladas. Lo que realmente posibilitan las videocámaras es hacer más eficiente y barato el mismo control. Por ello, con la aparición de estas nuevas tecnologías, y el abaratamiento del control, parece necesaria la existencia de nuevos límites o controles del uso de esa tecnología. En este objetivo se enmarca el nuevo art. 64.4 ET introducido en un primer momento por el RDL 9/2021 de 11 de mayo y convalidado en la Ley 12/2021, de 28 de septiembre.

II. EL DEBER DE TRANSPARENCIA ALGORÍTMICA EN EL ÁMBITO LABORAL

1. *Derecho de información de los representantes de los trabajadores sobre los algoritmos: el nuevo art. 64.4 ET*

El impacto social y mediático de las plataformas digitales ha puesto de relieve el algoritmo como herramienta de gestión empresarial con efectos sobre los trabajadores más allá del mundo de las plataformas. Conforme la doctrina viene señalando en los últimos años⁹, cada vez

⁹ A.E. Dembe et al., “The impact of overtime and long work hours on occupational injuries and illnesses: new evidence from the United States”, *Occupational and Environmental Medicine*, 62 (9), 2005, pp. 588-597; J. Domeinski et al., “Human redundancy in automation monitoring: Effects of social loafing and social compensation”, *Proceeding of human factors and ergonomics Society 51st Annual meeting*, 2007, pp. 587-591; W-S. Hung et al., “Managing the risk of overusing mobile phones in the working environment: a study of ubiquitous techno-stress”, *PACIS 2011 Proceedings* Paper 81, 2011; S. Schumacher, “What employees should know about electronic performance monitoring” *ESSAI* 8 (28), 2011, 138-144; EU-OSHA, *Monitoring technology: the 21st century’s pursuit of well-being?*, 2017. Disponible en <https://osha.europa.eu/en/tools-and-publications/publications/monitoring-technology-workplace/view>; EU-OSHA, *Foresight on new and emerging occupational safety and health risks associated with digitalisation by 2025*, 2018. Disponible en <https://osha.europa.eu/en/tools-and-publications/publications/foresight-new-and-emerging-occupational-safety-and-health-risks/view>; D. Derks, y A.B. Bakker, “Smartphone use,

es más habitual que las empresas usen los algoritmos o la inteligencia artificial para tomar decisiones que afectan a los trabajadores (contrataciones y despidos, determinación de cuadrantes de horarios, evaluación de los trabajadores, ascensos, etc.). Unos métodos que muchas veces son desconocidos para los propios trabajadores que consideran que dichas decisiones son tomadas por humanos cuando realmente vienen apoyados por esta tecnología o directamente el responsable humano se limita a comunicar lo que el algoritmo ha decidido.

work-home interference and burnout: a diary study on the role of recovery”, *Applied Psychology*, 63 (3), 2014; O. Pérez Zapata et al., “Digitalización, intensificación del trabajo y salud de los trabajadores españoles”, 2019. Disponible en <https://www.ccoo.es/24c0e370fa4b4d1f3682b1780854af9c000001.pdf>; O. Pérez-Zapata, *Trabajo sin límites, salud insostenible: La intensificación del trabajo del conocimiento* (e-prints, Universidad Complutense de Madrid), 2015; G. Lindsay, “We spent two weeks wearing employee Trackers: Here’s what we learned”, *Fact Coexist*, 2015, <https://www.fastcompany.com/3051324/we-spent-two-weeks-wearing-employee-trackers-heres-what-we-learned>; P. Akhtar, y P. Moore, “The Psycho-Social Impacts of Technological Change in Contemporary Workplaces and Trade Union Responses” *International Journal of Labor Research*, 8 (1-2), 2016, pp. 102-131; I. Ajunwa et al., “Limitless worker surveillance”, *cit.*; J. Horton et al., “Workplace Safety Futures: The impact of emerging technologies and platforms on work health and safety and workers’ compensation over the next 20 years”, CSIRO, Canberra, 2018; J. A. Fernández Avilés, “NTIC y riesgos psicosociales en el trabajo: estado de situación y propuestas de mejora”, *DSL*, 2, 2017; V. Rodríguez-Rico Roldan, “Los retos para la prevención de riesgos laborales ante la tecnificación del trabajo”, en N. Cerejeria Namora et al., *Health at work, ageing and environmental effects on future social security and labour law Systems*, Cambridge Scholars Publishing, 2018; P. Moore, *Work and the GDPR: The future for algorithms and people analytics*. Disponible en <https://staffblogs.le.ac.uk/management/2018/09/20/work-and-the-gdpr-the-future-for-algorithms-and-people-analytics/>, 2017; P. Moore, *The quantified self in precarity. Work, technology and what counts*, New York, Routledge, 2018; P. Moore, *Humans and machines at work: monitoring, surveillance and automation in contemporary capitalism*, London, Palgrave Macmillan, 2018; P. Moore, *Data subjects, digital surveillance, AI and the future of work*, European Parliament Science and Technology Office, Brussels, 2020; UTS, *Heat Stress and On-Demand work: The experience of food delivery and courier cyclists*, Climate Justice Centre, 2019. Disponible en <https://opus.lib.uts.edu.au/bitstream/10453/134736/1/On%20Demand%20cyclists%20UTS%20final.pdf>; A. Todolí Signes, “En cumplimiento de la primera Ley de la robótica: Análisis de los riesgos laborales asociados a un algoritmo artificial / inteligencia dirigiendo el trabajo”, *Labor Law & Issues*, 5, 2, 2019. Disponible en https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3506584

Cabe decir que, el RGPD regula, desde su entrada en vigor en 2018, la obligación de la empresa de informar al trabajador cuando estos métodos de procesamiento automatizado y de creación de perfiles eran usados. Un derecho de alcance individual que, en muchas ocasiones, se incumple dada la incapacidad del trabajador individual, de un lado, de conocer cuándo se está incumpliendo esta obligación, de otro lado, por el propio miedo del trabajador a represalias si decide reclamarlo. Por esta razón, la doctrina ha criticado al RGPD por olvidar la faceta colectiva y sindical del mundo del trabajo y no haber concedido derechos de información y negociación sobre el algoritmo a los representantes de los trabajadores¹⁰.

Un “defecto” que la ley 21/2021 pretende solventar con la introducción de una nueva letra d) en el apartado 4 del artículo 64 (relativo a los derechos de información y consulta del comité de empresa), con la siguiente redacción:

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.

Así, el comité de empresa tiene derecho a conocer, no solamente si existe algún algoritmo que tome decisiones o influya en las condiciones de trabajo, acceso o mantenimiento del empleo, sino también a saber los parámetros, reglas e instrucciones de funcionamiento de dicho algoritmo o sistema de inteligencia artificial. Es importante señalar que esta obligación se incluye para todo tipo de empresas (sean de plataformas o no) y para todo tipo de sectores¹¹. De esta forma, aunque la ley 21/2021 se ha bautizado como “ley rider”, la realidad

¹⁰ A. Todolí Signes, “La gobernanza colectiva de la protección de datos en las relaciones laborales: big data, creación de perfiles, decisiones empresariales automatizadas y los derechos colectivos”, *RDS*, núm. 84, 2018, pp. 69-88; C. Molina Navarrete, *Datos y derechos digitales de las personas trabajadoras en tiempos de (pos)covid19: entre eficiencia de gestión y garantías*, Bomarzo, 2021.

¹¹ C. Molina Navarrete, “La “Ley” de personas repartidoras en plataformas online (“riders”): ¿pequeño paso legal, gran paso para humanizar el precariado digital?”, *Transformaework*, 2021.

es que la mitad de sus preceptos (uno de dos) es una modificación que afecta a todo tipo de empresas, en todos los sectores.

2. *Concepto de algoritmo*

La introducción en el Estatuto de los trabajadores de una regulación sobre los algoritmos plantea una primera pregunta ¿qué es un algoritmo?

En este sentido, la RAE define algoritmo como el “conjunto ordenado y finito de operaciones que permite hallar la solución de un problema”. De esta forma, lo primero que se debe hacer es desligar el concepto de algoritmo del de programa informático (software). En efecto, un algoritmo no es más que un procedimiento metodológico que permite solucionar un problema, sin que sea necesario que para ello se use un programa informático —aunque habitualmente se haga para agilizar y simplificar la aplicación del procedimiento o de las operaciones necesarias. En este sentido, los representantes de los trabajadores tienen derecho a conocer cualquier sistema (conjunto ordenado de operaciones) usado por la empresa con objeto de incidir en las condiciones de trabajo, el acceso y mantenimiento del empleo, incluida la elaboración de perfiles.

El uso del término “algoritmo”, debe ponerse en contexto con la normativa europea, a pesar de que la terminología no encaje perfectamente¹². En la regulación Europea se usa un concepto más restrictivo que el “algoritmo”. Así, el RGPD enmarca los derechos individuales frente a la inteligencia artificial y los algoritmos mediante el concepto de “decisiones automatizadas”. En este sentido, en la regulación europea, para aplicar la normativa especial, la empresa debió establecer un conjunto de reglas que, de forma autónoma y automática —sin intervención humana—, llegan a una solución (por ejemplo, decidir a qué trabajador se contrata de entre distintos candidatos o decidir qué candidatos se descartan del proceso de selección).

¹² En el mismo sentido se pronuncia la “Guía práctica y herramienta sobre la obligación empresarial de información sobre el uso de algoritmos en el ámbito laboral”, 2022, elaborada por el Ministerio de Trabajo y Economía Social.

Así, aunque el nuevo apartado d) del art. 64. ET no incluya expresamente ese requisito de “automaticidad” exigido en Europa, la realidad es que todo algoritmo, para ser considerado tal, debe permitir hallar una solución —una decisión o una respuesta—, sin intervención humana. En este sentido, el conjunto ordenado y finito de operaciones, debe funcionar de forma automática —una operación tras otra— hasta alcanzar la solución buscada en su programación.

III. COMPARATIVA ENTRE LOS DERECHOS DE INFORMACIÓN DE ESTATUTO Y LA REGULACIÓN EUROPEA

La principal diferencia entre los derechos de información del art 64 ET en materia de algoritmos y la regulación del art. 22 RGPD, como ya se ha indicado, es que los primeros son de carácter colectivo y los segundos de carácter individual. Sin embargo, aquí no terminan las diferencias.

En este sentido, el art 64.4 ET otorga derechos de información en favor de los representantes de los trabajadores en muchos más supuestos que lo hace el art. 22 RGPD, pero con efectos menores.

Así, para la activación del art. 22 RGPD se requiere que las decisiones automáticas se tomen *sin intervención humana significativa*¹³. Por el contrario, letra d) en el apartado 4 del artículo 64, solamente exige que dichos algoritmos *afecten* a la toma de decisiones que *puedan incidir* sobre el trabajador. Así, aunque se use el algoritmo como simple apoyo a la toma de decisiones por parte del empresario se aplicará el art. 64.4 d) ET. Dicho en otras palabras, aunque el algoritmo no sea determinante para la decisión final tomada sobre el trabajador, su mero uso, implicará el nacimiento de los derechos de información de los representantes de los trabajadores. Resultando, en definitiva, que la aplicación del art. 64.4 d) ET se dará en muchos más casos que el art. 22 RGPD.

¹³ Como establece el GT29 en *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, Adoptadas el 3 de octubre de 2017, p. 21.

En la otra cara de la moneda, los efectos del art. 22 RGPD son mucho mayores, cuando efectivamente es de aplicación, que los meros derechos de información de carácter colectivo del art. 64.4 d) ET. Así, el art. 22 RGPD establece, como primera garantía existente en caso de decisiones basadas únicamente en el tratamiento automatizado de datos, su *prohibición*. En este sentido, el art. 22 RGPD concede el derecho a no ser objeto de este tipo de decisiones, derecho que el GT29 interpreta como una prohibición —sin necesidad de reclamar activamente el derecho— a los responsables de datos de tomar las decisiones con esta metodología automatizada¹⁴. No obstante, esta es una prohibición relativa dado que existen excepciones, a saber: a) que sea necesaria para la celebración o la ejecución de un contrato; b) que esté autorizada por el Derecho interno de un Estado Miembro; c) que se cuente con el consentimiento explícito del interesado. Ahora bien, dado que, en el contrato de trabajo, el consentimiento del trabajador, por regla general, no será válido¹⁵ —y a falta de una regulación nacional—, solamente se podrá entender válida la decisión automatizada si es *necesaria* para la celebración o ejecución del contrato de trabajo (excepción (a)).

A su vez, en caso de que una de las excepciones a la prohibición de uso de los algoritmos fuera aplicable (necesidad para cumplir con el contrato) el RGPD, en el art. 22.3, establece una serie de garantías que la empresa debe cumplir “para salvaguarda de los derechos y libertades y los intereses legítimos del interesado”, siendo estas garantías, “como mínimo el derecho a obtener intervención humana por

¹⁴ GT29, en *Guidelines on transparency under Regulation 2016/679*, p. 23.

¹⁵ El considerando 43 del RGPD establece que no es posible aceptar la licitud para el tratamiento y procesamiento de datos con base en el consentimiento en una relación donde exista un fuerte desequilibrio de poder entre las partes. Concretamente, el Reglamento hace referencia a la relación entre Administraciones Públicas y ciudadanos, no obstante, esta consideración parece perfectamente aplicable a la relación laboral. De hecho, el Grupo de Trabajo del art 29, en su Guía sobre el consentimiento bajo el RGPD (p. 8) entiende que el consentimiento de un trabajador difícilmente puede ser otorgado cumpliendo dichos requisitos (específicamente de forma “libre”). Por esta razón, el Grupo de Trabajo dictamina que, dada la naturaleza de la relación entre empresario y trabajador, como regla general, no se deberá entender válido el otorgamiento de consentimiento por parte del trabajador, teniendo que ser solamente aceptado como válido de forma excepcional.

parte del responsable, a expresar su punto de vista y a impugnar la decisión”¹⁶.

La doctrina ha interpretado, conforme a este artículo, que existe una obligación del responsable de datos de informar al afectado de las razones que han llevado a tomar esa decisión. Es decir, la empresa, cuando tome decisiones automatizadas, deberá indicarlo e informar al trabajador individual de qué parámetros (y qué ponderación ha asignado a cada uno de ellos) ha utilizado para alcanzar dicha resolución. Esta interpretación es apoyada por el art. 5 RGPD que exige que el procesamiento de datos sea legal, justo y transparente y, también, por los artículos 13.2 f) y 14.2 g) RGPD que exigen que cuando el sujeto es objeto de decisiones automatizadas, incluyendo la elaboración de perfiles, el responsable de datos deberá entregar al sujeto *información significativa sobre la lógica aplicada, así como la importancia y las consecuencias previstas de dicho tratamiento para el interesado*.

De esta manera, se comprueba que la configuración como derecho colectivo dentro del derecho nacional es de mayor alcance: se aplica a un superior número de supuestos que la regulación europea de carácter individual. Sin embargo, esta última, cuando surte efecto, tiene un contenido mayor que no consiste en el mero derecho de información, sino que permite al trabajador individual exigir la intervención humana en la decisión tomada por el algoritmo, así como, el derecho del trabajador a expresar su punto de vista.

IV. CONTENIDO DE LA INFORMACIÓN A ENTREGAR

El art. 64.4 d) ET establece que el comité de empresa tendrá derecho a ser informado de *los parámetros, reglas e instrucciones en las que se basan los algoritmos o sistemas de inteligencia artificial*. Una vez más, esta norma se aparta de la redacción de la normativa europea que establece el derecho individual del trabajador afectado a conocer la información *significativa sobre la lógica aplicada, así co-*

¹⁶ Ver más en A. Todolí Signes, “La gobernanza colectiva de la protección de datos...”, *cit.*, pp. 69-88.

mo la importancia y las consecuencias previstas de dicho tratamiento para el interesado.

Respecto a la normativa europea, la doctrina, de manera unánime, interpreta que el interesado tiene derecho a recibir información sobre el uso de los algoritmos en tres aspectos: i) informar de que el sujeto está envuelto en un proceso automatizado de toma de decisiones, es decir, informar al trabajador de que el proceso de decisión será total o parcialmente automatizado; ii) proveer información significativa sobre la lógica del algoritmo, esto es, entre otras cuestiones, indicar los parámetros evaluados por el algoritmo que toma la decisión y la ponderación de dichos parámetros; iii) informar de las consecuencias del proceso, es decir, qué consecuencias tendrá para el trabajador la decisión, en un sentido u otro, tomada de forma automática¹⁷.

Sin embargo, la doctrina no es unánime a la hora de interpretar el grado de explicación necesaria. Por una parte, una minoría de autores establece que la obligación se limita a exigir una explicación *ex ante* (anterior a la toma de la decisión) y *general* de los datos suministrados

¹⁷ S. Wacher, B. Mittelstadt, y L. Floridi, “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation”, *International Data Privacy Law*, 7, 2017, pp. 79–90; I. Mendoza, y L. A. Bygrave, “The Right Not to Be Subject to Automated Decisions Based on Profiling”, en T. Synodinou et al. (eds.) *EU internet law: regulation and enforcement*, 2017, <https://papers.ssrn.com/abstract=2964855> [<https://perma.cc/XV3T-G98W>]; L. Edwards, y M. Veale, “Slave to the Algorithm? Why a ‘Right to an Explanation’ Is Probably Not the Remedy You Are Looking For”, 16 *Duke Law & Technology Review*, 2017, pp. 18–84; K. Wan, y B. Routledge, “Algorithmic Transparency, a Right to Explanation, and Placing Trust”, *Squarespace*, 2017; G. Malgieri, y G. Comandé, “Why a Right to Legibility of Automated DecisionMaking Exists in the General Data Protection Regulation”, *International data privacy Law*, 7 (4) p. 243, pp. 246–47, 2017; B. Goodman, y S. Flaxman, “EU Regulations on Algorithmic Decision-Making and a “Right to Explanation,” *ARXIV*, 6, 2016, <http://arxiv.org/abs/1606.08813> [<https://perma.cc/SZTR-WG8R>]; A.D. Selbst, y J. Powles, “Meaningful information and the right to explanation”, *International Data Privacy Law*, 233, 2017; S. Wachter et al, “Counterfactual explanations without opening the black box: automated decisions and the GDPR”, *Arxiv*, 2017, p. 1–51, en <https://arxiv.org/ftp/arxiv/papers/1711/1711.00399.pdf>; E. Gil González, “Aproximación al estudio de las decisiones automatizadas en el seno del Reglamento General Europeo de Protección de Datos a la luz de las tecnologías big data y de aprendizaje computacional”, *Revista española de la Transparencia*, 5, 2017, pp. 165–179.

al algoritmo para que este tome la decisión¹⁸. Por su parte, la doctrina mayoritaria establece que el art. 22.3 RGPD, en conjunción con los arts. 13.2 f) y 14.2 g) y el considerando 71 del RGPD, exigen una explicación *ex post* (después de tomar la decisión) y *específica* de cómo y porqué se ha tomado dicha decisión sobre ese concreto trabajador¹⁹.

Respecto al nuevo art. 64.4 d) ET no parece que vaya a repetirse dicho debate. En efecto, dado que los derechos de información de los representantes de los trabajadores son generales sobre el uso del algoritmo y no respecto a una concreta decisión —o uso— individual del algoritmo, parece que los derechos de información se limitarán a una explicación *ex ante* (anterior al uso del algoritmo para supuestos concretos) y general de los parámetros, reglas e instrucciones con los que el algoritmo está diseñado para que este tome decisiones.

Una cuestión sobre la que la normativa nacional guarda silencio es sobre si la empresa deberá informar a los representantes de los trabajadores del uso que se vaya a dar a ese algoritmo y de las consecuencias que pueda tener. Esto es, se plantea la duda de si la empresa estará obligada a informar respecto a si el algoritmo se usará para procesos de selección, evaluación de trabajadores, ascensos, despidos, etc., o cuál será el nivel de intervención del algoritmo en estas materias.

Sobre esta cuestión cabe decir que difícilmente se puede separar una información de otra. La información entregada a los representantes de los trabajadores sobre los parámetros, reglas e instrucciones del algoritmo, con objeto de que pueda ser de alguna utilidad, debe entregarse en su contexto. Lo cual incluye, al menos, para qué se usará ese algoritmo.

¹⁸ S. Wacher, B. Mittelstadt, y L. Floridi, “Why a Right to Explanation of Automated Decision...”, *cit.*; S. Wachter et al., “Counterfactual...”, *cit.*, pp. 1-51, E. Gil González, “Aproximación...”, *cit.*, pp. 165-179.

¹⁹ I. Mendoza, y L.A. Bygrave, “The Right Not to Be Subjected...”, *cit.*; L. Edwards, y M. Veale, “Slave to the Algorithm?...”, *cit.*; K. Wan, y B. Routledge, “Algorithmic Transparency...”, *cit.*; G. Malgieri, y G. Comandé, “Why a Right to Legibility of Automated...”, *cit.*; B. Goodman, y S. Flaxman, “EU Regulations on Algorithmic...”, *cit.*; A.D. Selbst, y J. Powles, “Meaningful information and...”, *cit.*; A. Ginès i Fabrellas, “¿Tienen l@s riders derecho a conocer el algoritmo?, *Do Better*, 2021. Accessible en <https://dobetter.esade.edu/es/riders-derecho-algoritmo>

Ciertamente, cualquier análisis sobre los riesgos que pueda tener el algoritmo —riesgos de discriminación²⁰, riesgos para la salud de los trabajadores²¹, etc.—, partirá de conocer para qué se va a usar —qué decisiones va a tomar el algoritmo— y con qué intensidad se va a usar —¿tomará decisiones sin intervención significativa de un ser humano?, ¿se usará solamente como apoyo a la toma de decisiones?, ¿irá acompañado de auditorías “humanas”?

Lo que no parece que puedan exigir los representantes de los trabajadores con la redacción actual es información sobre todas y cada una de las decisiones tomadas por el algoritmo *a posteriori* (una vez tomadas). Un vacío que dificultará que los representantes de los trabajadores puedan analizar y evaluar el buen funcionamiento del propio algoritmo con objeto de poder garantizar que este no discrimina, no vulnera derechos fundamentales o, sencillamente, que no está tomando decisiones arbitrarias. En cualquier caso, nada obsta para que esto se pacte en la negociación colectiva o se incluya posteriormente en la ley. Adicionalmente, los trabajadores individuales afectados sí podrán reclamarlo vía art. 22 RGPD o los representantes en su nombre como he defendido en otros lugares²².

Tampoco parece que la redacción actual del nuevo apartado d) del art. 64 ET obligue a la empresa a identificar qué información ha suministrado al algoritmo para que pueda tomar sus decisiones. Los algoritmos, con objeto de poder establecer correlaciones en los datos, que le permitan posteriormente tomar decisiones, necesitan “alimentarse” con gran cantidad de información previa, momento en el que el algoritmo “aprenderá”. Tal y como señala la doctrina²³, la calidad de estos datos será clave para asegurar que el algoritmo no discrimina y es que, si se “alimenta” al algoritmo con datos sesgados, el algoritmo entenderá que discriminar es la mejor forma de tomar decisiones. Así pues, de nuevo, sin un conocimiento claro de los datos usados por

²⁰ M. Hardt, “How big data is unfair”, *Medium*, 2014, <https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de>; M.P. Rivas Vallejo, *La aplicación de la inteligencia artificial al trabajo y su impacto discriminatorio*, Aranzadi, 2020.

²¹ A. Todolí Signes, “En cumplimiento de la primera Ley de la robótica...”, *cit.*

²² A. Todolí Signes, “La gobernanza colectiva de la protección de datos...”, *cit.*, pp. 69-88.

²³ M. Hardt, “How big data is unfair”, *cit.*

la empresa para “entrenar” al algoritmo será difícil analizar el buen funcionamiento del mismo.

No obstante lo anterior, cabe señalar que la “Guía práctica y herramienta sobre la obligación empresarial de información sobre el uso de algoritmos en el ámbito laboral” publicada por el Ministerio de Trabajo en 2022 sí establece como obligatoria la entrega de los datos de “entrenamiento”. En este sentido se señala que “los datos de entrenamiento y, en su caso, validación, utilizados y sus características, la referencia a información sobre la lógica del algoritmo (artículos 13.2.f) y 14.2.g) 15.1.h) RGPD) y a “instrucciones” (artículo 64.4.d) ET) debe interpretarse en el sentido que incluye también los datos de entrenamiento y, en su caso, validación del sistema, por cuanto estos también influyen en su lógica o instrucciones del algoritmos. Por tanto, la empresa deberá informar sobre (i) los datos de entrenamiento y, en su caso, validación utilizados, (ii) su calidad, entendida como que sean adecuados, pertinentes, no excesivos, en atención a la finalidad para los que fueron obtenidos, etc., y (iii) el tipo de patrones identificados en los datos de entrenamiento.”

V. DERECHOS DE INFORMACIÓN YA EXISTENTES QUE AFECTAN A LOS ALGORITMOS

Es importante señalar que los derechos de información sobre el algoritmo o sistema de inteligencia artificial no terminan con el nuevo apartado d) del art. 64.4 ET. El resto de derechos de información de los representantes de los trabajadores siguen vigentes y serán aplicables cuando se use un algoritmo para tomar decisiones.

Así, el art. 64.5 ET establece que el comité de empresa “tendrá derecho a ser informado y consultado sobre todas las decisiones de la empresa que pudieran provocar cambios relevantes en cuanto a la organización del trabajo”. Así pues, el uso de un algoritmo que afecte a la organización del trabajo dará derecho al comité, no solo a ser informado sobre el mismo y su funcionamiento, sino también obligará a la empresa a consultar a los representantes antes de su puesta en funcionamiento.

De la misma manera, conforme al art. 64.5 f) ET, si el algoritmo tiene incidencia sobre los sistemas de organización y control del tra-

bajo —incluyendo poder disciplinario—, los sistemas de rendimiento, de tiempo de trabajo o de primas o incentivos, el comité de empresa tendrá derecho a ser informado y a emitir un informe previa su puesta en funcionamiento.

En ambos casos, para que se cumplan los derechos de consulta y emisión de informe es necesario que la información entregada sea apropiada y adecuada conforme el art. 64.6 ET. Así pues, al menos en estos casos, sí parece que el comité de empresa pueda solicitar información, una vez puesto en marcha el algoritmo, sobre las decisiones que está tomando y su funcionamiento con objeto de valorar que no está discriminando y, en general, que no está vulnerando ningún derecho de los trabajadores. Esta información respecto a las consecuencias de las decisiones tomadas por el algoritmo —incluido los perfiles— debería entregarse en un formato adecuado para su tratamiento por parte de los representantes. De la misma forma, en estos casos, parece que sí sería exigible que los representantes de los trabajadores conocieran qué datos han sido suministrados (o lo van a ser) al algoritmo para su desarrollo.

VI. CONCLUSIONES: TRASCENDENCIA DEL NUEVO DERECHO DE INFORMACIÓN

Cabe decir que el derecho de información de los representantes de los trabajadores contenido en el art. 64.4 d), probablemente, pudiera haberse extraído del resto de derechos de información ya existentes en el Estatuto de los trabajadores. Sin embargo, se debe aplaudir que el legislador no haya esperado a que el Tribunal Supremo se hubiera pronunciado sobre esta cuestión y que, por el contrario, demuestre verdadera iniciativa propia regulando una cuestión actual, adelantándose, así, a la conflictividad.

En la práctica, uno de los mayores problemas que puede surgir en materia de algoritmos es el desconocimiento total que pueden tener los representantes de los trabajadores y los propios trabajadores de su existencia. No es solo que la novedad de su uso haga difícil que estos estén informados, sino que, además, la intangibilidad del algoritmo puede hacer imposible a la parte social el conocimiento de su mera existencia. De esta forma, parece muy relevante que la norma, de for-

ma clara, exponga la obligación de informar, no solo a los propios trabajadores (art. 22 RGPD), sino también a los representantes de los trabajadores (nuevo art. 64.4 d)) ET de su existencia.

La doctrina ha señalado en multitud de ocasiones los perjuicios que puede tener para el trabajador el uso de estos sistemas como sustituto de los supervisores humanos —discriminación, deshumanización, vulneración de derechos fundamentales, problemas de salud—²⁴. Por ello, cualquier medida que traiga transparencia sobre su uso debe ser elogiada.

No obstante, este —el conocimiento de su existencia— no deja de ser solamente el primer paso. Dado que los algoritmos pueden ser determinantes para los sistemas de trabajo, las condiciones de trabajo, la contratación, los ascensos y los despidos, parece obvio que el siguiente paso será negociar colectivamente su funcionamiento²⁵. En el fondo, esto no será otra cosa que cumplir con la función clásica de la negociación colectiva: negociar colectivamente las condiciones de trabajo.

²⁴ P. Moore, *Work and the GDPR: The future for algorithms...*, cit.; P. Moore, *The quantified self in precarity...*, cit.; P. Moore, *Humans and machines at work...*, cit.; P. Moore, *Data subjects, digital surveillance, AI and...*, cit.; UTS, *Heat Stress and On-Demand work...*, cit.; A. Todolí Signes, “En cumplimiento de la primera Ley de la robótica...”, cit.; M.P. Rivas Vallejo, *La aplicación de la inteligencia artificial*, cit.

²⁵ V. De Stefano, “Negotiating the Algorithm...”, cit.; A. Todolí Signes, “La gobernanza colectiva de la protección de datos...”, cit., pp. 69-88. Así se pronuncia también E. Rojo Torrecilla, en “Y llegó la norma que declara la relación laboral de “las personas dedicadas al reparto en el ámbito de las plataformas digitales”. Primeras notas y comentarios al RDL 9/2021 de 11 de mayo”, en los siguientes términos “abriéndose así un amplio espacio para, más allá de la mera información, empezar a debatir y negociar sobre “la gestión del algoritmo”.

Apuesta por Tirant Online, la base de datos jurídica de la editorial más prestigiosa de España.*



www.tirantonline.com

Suscríbete a nuestro servicio de base de datos jurídica y tendrás acceso a todos los documentos de Legislación, Doctrina, Jurisprudencia, Formularios, Esquemas, Consultas o Voces, y a muchas herramientas útiles para el jurista:

- * Biblioteca Virtual
- * Herramientas Salariales
- * Calculadoras de tasas y pensiones
- * Tirant TV
- * Personalización
- * Foros y Consultoría
- * Revistas Jurídicas
- * Gestión de despachos
- * Biblioteca GPS
- * Ayudas y subvenciones
- * Novedades

* Según ranking del CSIC

 96 369 17 28

 96 369 41 51

 atencionalcliente@tirantonline.com

 www.tirantonline.com