Jorge Castellanos Claramunt
Adrián Palma Ortigosa
Directores

Garantías ante las decisiones y perfiles automatizados en el sector público

CONSTITUCIONAL

Atelier
LIBROS JURÍDICOS

Garantías ante las decisiones y perfiles automatizados en el sector público

GARANTÍAS ANTE LAS DECISIONES Y PERFILES AUTOMATIZADOS EN EL SECTOR PÚBLICO

Jorge Castellanos Claramunt
Adrián Palma Ortigosa
Directores

Autores

Jorge Castellanos Claramunt
María Teresa García-Berrio Hernández
José Miguel Iturmendi Rubia
María Dolores Montero Caro
Adrián Palma Ortigosa
Francisca Ramón Fernández
Marco Emilio Sánchez Acevedo
Pere Simón Castellano
Gabriele Vestri

Atelier
LIBROS JURÍDICOS

Colección: Atelier Constitucional
Directores:
Joan Manuel Trayter
(Catedrático de Derecho administrativo)

Belén Noguera de la Muela
(Catedrática de Derecho administrativo)

Proyecto Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas [2023-2026] (PID2022-136439OB-I00), financiado por MCIN/AEI/10.13039/501100011033/ y «FEDER Una manera de hacer Europa».

© 2025 Atelier
Santa Dorotea 8, 08004 Barcelona
e-mail: editorial@atelierlibros.es
www.atelierlibros.es
Tel.: 93 295 45 60

I.S.B.N.: 979-13-87867-91-1
Depósito legal: B 21827-2025

Diseño y composición: Addenda, Pau Claris 92, 08010 Barcelona
www.addenda.es

Impresión: Safekat

Índice

Presentación 9

Utilización de sistemas de Inteligencia Artificial generativa y las decisiones automatizadas: algunos aspectos clave tras el reglamento (UE) 2016/679 y el reglamento (UE) 2024/1689. 17
Francisca Ramón Fernández

Inteligencia Artificial y algoritmos en la gestión de pandemias: lecciones para el derecho público a partir de la Covid-19 39
Gabriele Vestri

Decisiones automatizadas y derecho a explicación: supervisión y transparencia en la administración pública 61
Jorge Castellanos Claramunt

La inteligencia artificial en la evaluación de solicitudes de asilo: un análisis crítico del caso alemán 83
José Miguel Iturmendi Rubia

Hacia una transformación post-burocrática en el uso responsable de la inteligencia artificial en la administración pública europea: desafíos, riesgos y nuevos horizontes para una gobernanza ética de la IA 107
María Teresa García-Berrio Hernández

Las decisiones automatizadas relevantes en el reglamento general de protección de datos: un análisis a las últimas resoluciones judiciales del TJUE 131
Adrián Palma Ortigosa

La interacción entre la ley de servicios digitales, la carta de derechos fundamentales de la ue y los derechos fundamentales reconocidos en las constituciones española y portuguesa 149
María Dolores Montero Caro

Especificaciones de seguridad, modelos verificables y control jurídico: hacia una ia confiable en las administraciones públicas 165
Pere Simón Castellano

Transparencia algorítmica y seguridad digital en la administración pública: construcción de un régimen operativo de compatibilización en los sistemas de inteligencia artificial del estado colombiano como referente iberoamericano 191
Marco Emilio Sánchez Acevedo

PRESENTACIÓN

La presente obra colectiva, *Garantías ante las decisiones y perfiles automatizados en el sector público*, ha sido financiada por el Ministerio de Ciencia e Innovación dentro del programa estatal de generación de conocimiento, y reúne a un nutrido grupo de especialistas en Derecho constitucional, administrativo, civil, penal, internacional y en ciencias políticas, con el propósito común de abordar uno de los desafíos jurídicos más relevantes de nuestro tiempo, tal es la inserción de la inteligencia artificial y los sistemas automatizados de decisión en el ámbito público, y las consecuencias que ello proyecta sobre los derechos fundamentales y las garantías del Estado democrático de Derecho.

La acelerada digitalización de las Administraciones públicas, impulsada por las estrategias europeas y nacionales de transformación digital —desde la Agenda España Digital 2025 hasta el Plan de Digitalización de las Administraciones Públicas 2021-2025— ha multiplicado los espacios en los que las decisiones algorítmicas inciden en la vida de los ciudadanos. A este respecto, la incorporación de tecnologías de aprendizaje automático, análisis predictivo y tratamiento masivo de datos promete una mejora sustancial en la eficiencia de la gestión pública, pero al mismo tiempo plantea interrogantes esenciales sobre la transparencia, la explicabilidad, la responsabilidad y la no discriminación en las actuaciones administrativas. Estas cuestiones, de hondo calado constitucional, se sitúan en el corazón del proyecto DERGORITMOS y constituyen la materia vertebral del presente volumen.

La investigación parte de un presupuesto epistemológico claro: la inteligencia artificial no es una realidad futura, sino un presente irreversible. Su desarrollo y utilización, tanto en el sector privado como en el público, resultan inevitables y, en muchos casos, beneficiosos para el interés general. Sin embargo, esta aceptación pragmática de la IA exige una contrapartida igualmente firme en el plano jurídico: la construcción de un marco normativo que garantice que su progreso se alinee con los principios democráticos, la dignidad de la persona y los derechos fundamentales. No se trata, por tanto, de oponer el Derecho a la tecnología, sino de asegurar una convivencia regulada entre innovación y garantía, entre la eficacia de los algoritmos y la tutela de los derechos.

El proyecto DERGORITMOS se sitúa en la intersección de tres grandes ejes regulatorios europeos: el Reglamento General de Protección de Datos (RGPD), el reciente Reglamento Europeo de Inteligencia Artificial (RIA) y la Carta de Derechos Digitales, impulsada en España como marco de referencia para una digitalización humanista. La interacción de estos instrumentos configura un entramado jurídico complejo, en el que se superponen obligaciones de transparencia, trazabilidad, explicabilidad y supervisión humana. Frente a la dispersión normativa y la todavía incipiente sistematización de garantías, el proyecto busca identificar modelos coherentes de cumplimiento normativo proactivo y responsabilidad por diseño, que permitan anticipar los riesgos y prevenir las vulneraciones antes de que se produzcan.

En este contexto, la presente obra se concibe como un espacio de diálogo interdisciplinar, donde convergen las perspectivas jurídicas, éticas y técnicas que demanda el estudio de la inteligencia artificial en el sector público. El propósito no es solo analizar las normas vigentes, sino también reflexionar críticamente sobre su adecuación a los desafíos que plantea la automatización administrativa. Las aportaciones que la integran abordan cuestiones tan diversas como la aplicabilidad del RGPD a los sistemas de decisión automatizada, la naturaleza jurídica del derecho a obtener una explicación, los límites constitucionales de la delegación de potestades decisorias en algoritmos, la compatibilidad de los procesos de perfilado con el principio de igualdad, o la necesidad de mecanismos efectivos de supervisión humana y rendición de cuentas.

El enfoque del proyecto —y, por extensión, de este libro— se aparta de los análisis puramente prohibitivos o alarmistas que a menudo acompañan a la inteligencia artificial. Sin desconocer los riesgos inherentes a su uso —sesgos estructurales, opacidad algorítmica, pérdida de control humano, vulneraciones de la privacidad o discriminaciones indirectas—, la aproximación que aquí se adopta es esencialmente garantista y constructiva. Se parte de la convicción de que el Derecho, lejos de ser un freno al progreso tecnológico, constituye la herramienta indispensable para encauzarlo conforme a los valores del constitucionalismo contemporáneo. La finalidad última no es limitar la inteligencia artificial, sino dotarla de legitimidad democrática y jurídica.

La obra se inscribe, además, en una trayectoria consolidada de investigación que el equipo viene desarrollando desde hace más de un lustro en torno al impacto de la digitalización en los poderes públicos. El carácter transversal y multidimensional del proyecto se refleja también en la composición del equipo de autores, que reúne a profesores e investigadores de distintas universidades españolas e internacionales, especialistas en Derecho constitucional, administrativo, civil, procesal, penal y filosófico, además de expertos en ética tecnológica y protección de datos. Esta diversidad disciplinar permite ofrecer una visión integral del fenómeno algorítmico, desde la dogmática de los derechos fundamentales hasta las técnicas de auditoría, evaluación de impacto y cumplimiento normativo. El resultado es un mosaico de perspectivas complementarias que, sin renunciar al rigor jurídico, aspira a servir como referencia para el legislador, la administración y la comunidad académica.

Entre los temas transversales que articulan la obra destacan especialmente dos. En primer lugar, la no discriminación algorítmica y el análisis de los sesgos que pueden derivarse de la utilización de datos de entrenamiento insuficientes, desactualizados o representativos de desigualdades sociales preexistentes. Los algoritmos, lejos de ser neutrales, pueden reproducir o amplificar patrones de discriminación indirecta, generando lo que algunos autores han denominado «discriminación por error». La identificación, prevención y corrección de estos sesgos constituye una de las principales líneas de trabajo del proyecto, que combina herramientas jurídicas y técnicas para garantizar decisiones justas y equitativas.

En segundo lugar, la obra presta atención prioritaria a la explicabilidad, trazabilidad y transparencia de los sistemas de decisión automatizada. Estos principios no solo responden a exigencias de control administrativo y rendición de cuentas, sino que se erigen en auténticas garantías del ciudadano frente a decisiones opacas o ininteligibles. El derecho a comprender las decisiones que le afectan —y a exigir la intervención humana en caso de discrepancia— se perfila como un nuevo eje de legitimidad democrática en la era digital. La transparencia algorítmica, entendida en su dimensión jurídica y técnica, se convierte así en un presupuesto para el ejercicio efectivo de otros derechos fundamentales, como la tutela judicial efectiva, la igualdad o la protección de datos personales.

Por lo indicado *supra*, este libro pretende ser más que una mera recopilación de estudios especializados. Es, ante todo, una aportación colectiva al proceso de construcción de un Derecho público de la inteligencia artificial, capaz de conjugar innovación y garantías, progreso y derechos. La inteligencia artificial, como en su día ocurrió con la imprenta o con internet, transforma las estructuras de poder, comunicación y conocimiento, por lo que el reto del jurista contemporáneo consiste en asegurar que esa transformación se produzca dentro del marco del Estado de Derecho y bajo el signo de la dignidad humana.

Los trabajos que se presentan a continuación, firmados por destacados expertos nacionales e internacionales, ofrecen respuestas concretas a estos desafíos desde perspectivas complementarias. Pero antes de adentrarse en el detalle de cada capítulo, conviene tener presente la idea rectora que inspira todo el volumen: la confianza en que el Derecho puede —y debe— acompañar a la inteligencia artificial, garantizando que la automatización y el algoritmo se conviertan en instrumentos al servicio de la libertad, la igualdad y la justicia, y no en amenazas para ellas.

El recorrido comienza con la contribución de la profesora **Francisca Ramón Fernández**, catedrática de Derecho civil en la Universitat Politècnica de València, que ofrece un análisis exhaustivo sobre la utilización de los sistemas de inteligencia artificial generativa y las decisiones automatizadas a la luz del Reglamento (UE) 2016/679 y del reciente Reglamento (UE) 2024/1689. Su aportación examina los riesgos derivados de los sesgos discriminatorios y la posible vulneración de derechos fundamentales, destacando la necesidad de garantizar una supervisión humana significativa en todos aquellos procesos catalogados como de alto riesgo. Con un enfoque crítico y normativo, la autora subraya que la

regulación europea busca equilibrar innovación y protección de la dignidad humana, imponiendo obligaciones de transparencia, trazabilidad y rendición de cuentas tanto a proveedores como a administraciones públicas.

Por su parte, **Gabriele Vestri**, presidente del Observatorio Sector Público e Inteligencia Artificial, centra su capítulo en las lecciones que dejó la gestión de la pandemia de COVID-19 respecto al uso de la inteligencia artificial y los algoritmos en el sector público. El autor muestra cómo estas herramientas se aplicaron en ámbitos tan diversos como el rastreo de contactos, la vigilancia de cuarentenas, la telemedicina o la asignación de recursos sanitarios, siempre bajo la tensión entre eficacia sanitaria y respeto a los derechos fundamentales. Su trabajo destaca las implicaciones constitucionales y administrativas de estas prácticas, señalando la necesidad de proporcionalidad, transparencia y supervisión humana para evitar vulneraciones de la privacidad, la igualdad o la dignidad personal. La reflexión concluye que la experiencia pandémica no solo evidenció el potencial transformador de la IA en la salud pública, sino también la urgencia de marcos normativos sólidos que aseguren su implementación conforme a los valores democráticos y al Estado de Derecho.

El profesor **Jorge Castellanos Claramunt**, titular de Derecho Constitucional en la Universitat de València, ofrece un estudio exhaustivo sobre las decisiones automatizadas en la administración pública y el derecho a obtener una explicación comprensible de las mismas. Su trabajo analiza en detalle la interacción entre el Reglamento General de Protección de Datos y el nuevo Reglamento Europeo de Inteligencia Artificial, subrayando cómo ambos instrumentos configuran un marco robusto para garantizar la transparencia y la supervisión de los algoritmos en el sector público. A partir de un enfoque jurídico-constitucional, el autor expone los riesgos que el uso indiscriminado de sistemas automatizados plantea para derechos fundamentales como la privacidad, la igualdad o la tutela judicial efectiva. Castellanos reivindica la necesidad de una supervisión humana significativa y de mecanismos efectivos de explicabilidad que permitan a la ciudadanía comprender, cuestionar e impugnar las decisiones algorítmicas. Su aportación resalta que solo a través de este doble eje —control humano y derecho a explicación— puede asegurarse que la digitalización administrativa se mantenga alineada con los valores democráticos y con el Estado de derecho.

Por su parte, el profesor **José Miguel Iturmendi Rubia**, docente de Filosofía del Derecho en CUNEF Universidad, aborda de manera crítica el uso de la inteligencia artificial en los procedimientos de asilo en Alemania. El capítulo examina dos mecanismos especialmente controvertidos —la extracción de datos de dispositivos electrónicos de los solicitantes y el análisis lingüístico automatizado para inferir su país de origen— mostrando cómo estas prácticas, introducidas bajo la lógica de la eficiencia, generan serias dudas jurídicas y éticas. El autor subraya la falta de transparencia algorítmica, los riesgos de decisiones discriminatorias y la amenaza que supone la automatización para principios como la proporcionalidad, la motivación individualizada o el acceso a recursos efectivos. A partir de un enfoque multidisciplinar, Iturmendi conecta el caso alemán con el marco europeo de control fronterizo y con la regulación emer-

gente sobre IA, advirtiendo de la necesidad de salvaguardias sólidas, transparencia y control judicial. Su contribución pone de relieve que, en contextos tan sensibles como el derecho de asilo, la innovación tecnológica debe subordinarse a la primacía de los derechos humanos y a los valores democráticos que sustentan el Estado de derecho.

A continuación, la profesora **María Teresa García-Berrio Hernández**, titular de Filosofía del Derecho en la Universidad Complutense de Madrid, examina en su contribución el impacto de la inteligencia artificial en la administración pública europea desde una perspectiva post-burocrática. Su capítulo traza la evolución desde el modelo weberiano de burocracia clásica hasta las nuevas formas de gobernanza inteligente, centradas en la creación de valor público y en la satisfacción ciudadana. A través del análisis de experiencias pioneras en países como Estonia, Dinamarca, Países Bajos y España, la autora muestra cómo la IA se ha convertido en un motor de innovación para optimizar procesos, personalizar servicios y favorecer la transparencia, sin dejar de advertir sobre los riesgos de sesgos, opacidad y afectación de derechos fundamentales. En un enfoque que combina reflexión filosófica y análisis normativo, García-Berrio subraya la necesidad de articular un marco ético y jurídico sólido —asado en los principios de autonomía, beneficencia, no maleficencia y justicia— que garantice un uso confiable de la IA en el sector público y refuerce la confianza ciudadana en las instituciones democráticas.

El siguiente trabajo lleva la firma del profesor **Adrián Palma Ortigosa**, ayudante doctor de Derecho Administrativo en la Universitat de València, que centra su capítulo en el análisis del artículo 22 del Reglamento General de Protección de Datos, que regula las decisiones plenamente automatizadas con efectos jurídicos o significativamente similares. Su estudio recorre el origen histórico de esta norma —desde la pionera ley francesa de 1978 hasta la reciente jurisprudencia del Tribunal de Justicia de la Unión Europea— y muestra cómo, en los últimos años, la creciente expansión de la inteligencia artificial ha convertido este precepto en un eje fundamental de garantía para los ciudadanos. A través de un examen detallado de las bases de legitimación y de las garantías previstas en favor de los interesados, así como de los últimos pronunciamientos judiciales, Palma subraya la necesidad de una supervisión humana significativa y de un auténtico derecho de explicación para hacer efectivo el control sobre los algoritmos. El capítulo concluye alertando de que la irrupción de modelos de inteligencia artificial generativa desafía los marcos normativos tradicionales, planteando la urgencia de repensar las herramientas jurídicas disponibles para asegurar la protección de derechos en la nueva era de la automatización.

La profesora **María Dolores Montero Caro**, docente de Derecho Constitucional en la Universidad de Córdoba, ofrece una reflexión jurídica profunda sobre la relación entre la Ley de Servicios Digitales (DSA) de la Unión Europea y los derechos fundamentales reconocidos en la Carta de Derechos Fundamentales de la UE y en las Constituciones española y portuguesa. La autora sitúa la DSA como un hito del nuevo constitucionalismo digital europeo, que pretende equilibrar la libertad, la privacidad y la seguridad en el entorno digital. A través

de un análisis comparado, examina cómo los marcos constitucionales de España y Portugal integran los principios de libertad de expresión, protección de datos y proporcionalidad frente a los desafíos de la digitalización, la desinformación y la inteligencia artificial. Montero advierte sobre los riesgos que los algoritmos y la manipulación informativa suponen para la democracia, subrayando la necesidad de una soberanía digital europea que preserve la dignidad humana, la transparencia y la rendición de cuentas frente al poder de las grandes plataformas tecnológicas. En definitiva, su contribución reivindica un humanismo digital basado en el Derecho, en el que la innovación tecnológica se subordine a los valores constitucionales y al fortalecimiento de la democracia en la era digital.

Por su parte, el profesor **Pere Simón Castellano**, lleva a cabo el trabajo titulado *«Especificaciones de seguridad, modelos verificables y control jurídico: hacia una IA confiable en las Administraciones Públicas»*, que ofrece una contribución clave al debate sobre la gobernanza algorítmica en el sector público. Desde una perspectiva técnico-jurídica, el autor propone un marco integrado para garantizar la confiabilidad de la inteligencia artificial en la Administración, articulando tres pilares fundamentales: las especificaciones de seguridad que traducen valores y obligaciones legales en requisitos técnicos verificables; los modelos de IA verificables, que permiten demostrar matemáticamente la corrección, equidad y robustez de los sistemas; y el control jurídico e institucional, que asegura la supervisión humana, el control democrático y la rendición de cuentas. Su análisis conecta el Reglamento Europeo de Inteligencia Artificial (RIA) con normas internacionales como la ISO/IEC 42001:2023 y destaca la convergencia entre estándares técnicos y marcos normativos como vía para construir una IA pública transparente, auditable y conforme al Estado de Derecho.

Por último, **Marco Emilio Sánchez Acevedo**, profesor de la Universidad Católica de Colombia presenta un capítulo en el que examina el desafío de armonizar la transparencia algorítmica con la seguridad digital en el uso de sistemas de inteligencia artificial y decisiones automatizadas en la administración pública colombiana. A partir de la Sentencia T-067 de 2025 de la Corte Constitucional y la Directiva Conjunta 007 de 2025, el autor propone un marco operativo que equilibra el derecho ciudadano a comprender el funcionamiento de los algoritmos con la necesidad de proteger la infraestructura digital y los datos personales del Estado. Se destacan mecanismos como la explicabilidad y el lenguaje claro, los análisis de impacto y auditorías algorítmicas, los canales de revisión y objeción, la ciberseguridad por diseño y la aplicación del test de daño para determinar el alcance de la divulgación de información. El trabajo concluye que transparencia y seguridad no son principios opuestos, sino complementarios, y que su integración es condición esencial para una gobernanza algorítmica democrática, responsable y confiable en el contexto iberoamericano.

En último lugar, queremos expresar nuestro más sincero agradecimiento al Ministerio de Ciencia, Innovación y Universidades y a la Agencia Estatal de Investigación, por la financiación del proyecto *Derechos y garantías públicas*

frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas [2023-2026] (PID2022-136439OB-I00), financiado por MCIN/ AEI/10.13039/501100011033 y cofinanciado por el Fondo Europeo de Desarrollo Regional (FEDER) «Una manera de hacer Europa», del que es fruto esta obra colectiva.

Utilización de sistemas de Inteligencia Artificial generativa y las decisiones automatizadas: algunos aspectos clave tras el reglamento (UE) 2016/679 y el reglamento (UE) 2024/1689[1]

Francisca Ramón Fernández
Catedrática de Derecho civil
Universitat Politècnica de València

SUMARIO: I. INTRODUCCIÓN. II. INTELIGENCIA ARTIFICIAL Y DECISIONES AUTOMATIZADAS: ALGUNAS CONSIDERACIONES. III. DECISIONES AUTOMATIZADAS Y SESGOS DISCRIMINATORIOS. IV. CONCLUSIONES. *BIBLIOGRAFÍA*.

1. Esta investigación se ha realizado en el marco del Grupo de Investigación de Excelencia Generalitat Valenciana «Algorithmical Law» (Proyecto Prometeu 2021/009, 2021-2024), y del proyecto de I+D+i *Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas* [2023-2026] (PID2022-136439OB-I00), financiado por MCIN/AEI/10.13039/501100011033/ y «FEDER Una manera de hacer Europa».

I. INTRODUCCIÓN

La utilización de sistemas de inteligencia artificial generativa y las decisiones automatizadas está presente en distintos ámbitos como es la salud[2], administración[3], contratación[4], judicial[5], tributario[6], y laboral[7], entre otros[8].

2. M. Pérez Sarabia, «La prohibición a la toma de decisiones automatizadas, clave esencial para los derechos de los pacientes en el ámbito de la medicina», en C. Villegas Delgado y P. Martín Ríos (ed.), *El derecho en la encrucijada tecnológica: Estudios sobre derechos fundamentales, nuevas tecnologías e inteligencia artificial*, Tirant lo Blanch, Valencia, 2022, pp. 357 y sigs.

3. D. U. Galetta, «Derechos y garantías concretas respecto del uso por los poderes públicos de decisiones automatizadas e inteligencia artificial: La importancia de las garantías en el procedimiento administrativo», en M. Bauzá Reilly (coord.), L. Cotino Hueso (dir.), *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, Thomson Reuters Aranzadi, Cizur Menor, 2022, pp. 171 y sigs.; E. Gamero Casado, «Sistemas automatizados de toma de decisiones en el Derecho Administrativo Español», *Revista General de Derecho Administrativo*, núm. 63, 2023; A. Garriga Domínguez, «Decisiones automatizadas basadas en algoritmos y protección de datos personales», en J. A. Viguri Cordero (coord.), *Las cláusulas específicas del Reglamento General de Protección de Datos en el ordenamiento jurídico español: cuestiones clave de orden nacional y europeo*, Tirant lo Blanch, Valencia, 2021, pp. 333 y sigs.

4. Véase: J. Águila Martínez y X. Dorado Ferrer, «El modelo de madurez en analítica de datos y el régimen jurídico de las decisiones automatizadas en el sector privado», en R. Oliver Cuello (dir.), *Las tecnologías de la información en la actividad empresarial. Aspectos legales y fiscales*, Thomson Reuters Aranzadi, Cizur Menor, 2022; Cherñavsky, N. «Inteligencia artificial y «big data» jurídica. Decisiones automatizadas. Modelos de aplicación en general y en el ámbito jurídico», en C. Ch. Sueiro (coord.), M. Alfredo Riquert (dir.), *Sistema penal e informática, volumen 7. Ciberdelitos, evidencia digital, TICs*, Hammurabi, Argentina, 2024, pp. 33 y sigs. B. D., «La toma de las decisiones automatizadas y el derecho fundamental a la protección de datos de carácter personal», *Quincena fiscal*, núm. 15-16, 2022; «La teoría de las garantías adecuadas en materia de protección de datos y sus implicaciones respecto de la toma de decisiones automatizadas en la Administración tributaria», en B. D. Olivares Olivares (coord.), *La inteligencia artificial en la relación entre los obligados y la administración tributaria: retos ante la gestión tecnológica*, La Ley, Madrid, 2022, pp. 240 y sigs.; A. G. Orofino, «Decisiones automatizadas, transformación de la administración y prestación de servicios públicos digitales», en *La digitalización en los servicios públicos: garantías de acceso, gestión de datos, automatización de decisiones y seguridad*, Marcial Pons, Madrid, 2023, pp. 139 y sigs.

5. Véase: H. R. Granero, «Derechos y garantías concretas frente al uso de inteligencia artificial y decisiones automatizadas, especialmente en el ámbito judicial y de aplicación de la ley», en M. Bauzá Reilly (coord.), L. Cotino Hueso (dir.), *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, Thomson Reuters Aranzadi, Cizur Menor, 2022, pp. 107 y sigs.

6. Pérez Bernabeu, B. «Decisiones automatizadas en la Administración tributaria», en B. D. Olivares Olivares (coord.), *La inteligencia artificial en la relación entre los obligados y la administración tributaria: retos ante la gestión tecnológica*, La Ley, Madrid, 2022, pp. 146 y sigs.; «Los contribuyentes ante las decisiones automatizadas de la administración tributaria», en A. M. Pita Gandal y L. A. Malvárez Pascual (coord.), *La digitalización en los procedimientos tributarios y el intercambio automático de información*, Thomson Reuters Aranzadi, Cizur Menor, 2023, pp. 235 y sigs.

7. Véase: E. Álvarez Fernández, «Sobre las decisiones automatizadas en el ámbito laboral y el derecho a la privacidad de los trabajadores», *Diario La Ley*, núm. 10587, 2024; Pérez Daudí, V. «El precedente judicial: la previsibilidad de la sentencia y la decisión automatizada del conflicto», *Revista iberoamericana de Derecho Procesal. RIDP*, núm. 2, 2020, pp. 141 y sigs.; «La previsibilidad de la sentencia y la decisión automatizada del conflicto», en M. De Prada Rodríguez, S. Calaza López y J. C. Muinelo Cobo (dir.), *El impacto de la oportunidad sobre los principios procesales clásicos: Estudios y diálogos*, Iustel, Madrid, 2021, pp. 395 y sigs. Disponible en: https://dialnet.

Para su desarrollo vamos a atender a la metodología consistente en el análisis de la legislación aplicable como es el 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 y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos)[9], la Ley Orgánica 3/2018, de 5 de diciembre, de protección de datos personales y garantía de los derechos digitales[10], el Real Decreto 203/2021, de 30 de marzo, por el que se aprueba el Reglamento de actuación y funcionamiento del sector público por medios electrónicos[11], y Ley Orgánica 7/2021, de 26 de mayo, de protección de datos personales tratados para fines de prevención, detección, investigación y enjuiciamiento de sanciones penales[12] y el Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) núm. 300/2008, (UE) núm. 167/2013, (UE) núm. 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial).[13] También analizaremos la aplicación del Reglamento (UE) 2018/1725 del Parlamento Europeo y del Consejo, de 23 de octubre de 2018, relativo a la protección de

unirioja.es/descarga/libro/845425.pdf (Consultado el 25 de marzo de 2025); «El precedente judicial. La previsibilidad de la sentencia y la decisión automatizada del conflicto», en M. J. Ariza Colmenarejo (dir.), *Revisión del Sistema de Fuentes y su repercusión en el Derecho Procesal*, Dykinson, Madrid, 2021, pp. 199 y sigs.; Ramón Fernández, F. «La utilización de la inteligencia artificial para la verificación de la identidad y el control de presencia mediante sistemas biométricos en el entorno laboral: algunas cuestiones», *La regulación de la inteligencia artificial y el derecho del trabajo. Retos en materia de dirección algorítmica del trabajo*, Aranzadi La Ley, Cizur Madrid, 2025, pp. 93 y sigs.; Rivas Vallejo, M. P. «Decisiones automatizadas y discriminación en el trabajo», *Revista General de Derecho del Trabajo y de la Seguridad Social*, núm. 66, 2023; Rodríguez Cardo, I. A. «Decisiones automatizadas y discriminación algorítmica en la relación laboral: ¿hacia un Derecho del Trabajo de dos velocidades?», *Revista española de derecho del trabajo*, núm. 253, 2022, pp. 135 y sigs.; Rodríguez Escanciano, S. y Álvarez Cuesta, H. «La toma de decisiones automatizadas en el marco de la relación laboral: otra vuelta de tuerca al poder de dirección y vigilancia empresarial», en L. Gamarra Vilchez (coord.), J. E. López Ahumada (dir.), *La gobernanza de los derechos digitales de las personas trabajadoras*, Cinca, Madrid, 2023, pp. 109 y sigs.

8. J. A. Moreno Martínez, «Las garantías en la protección de datos personales en las ciudades inteligentes «smart cities» e implantación de decisiones automatizadas con uso de la inteligencia artificial», *Revista de derecho privado*, núm. 108, 2024, pp. 61 y sigs.; Sobrino García, I. «Las decisiones automatizadas en el sector público: conflictos entre la protección de datos y la inteligencia artificial», en P. R. Bonorino Ramírez y R. Fernández Acevedo (dir.), *Nuevas normatividades: inteligencia artificial, derecho y género*, Thomson Reuters Aranzadi, Cizur Menor, 2021, pp. 17 y sigs.; Viguri Cordero, J. A. «La transparencia como reto en la toma de decisiones automatizadas en los procesos migratorios», en L. S. Heredia Sánchez (coord.), A. Ortega Giménez (dir.), *Protección de datos, transparencia, asociacionismo y voluntariado: buenas prácticas de actuación con el colectivo migrante*, Aranzadi, Cizur Menor, 2023, pp. 113 y sigs.

9. DOUE núm. 119, de 4 de mayo de 2016.

10. BOE núm. 294, de 6 de diciembre de 2018.

11. BOE núm. 77, de 31 de marzo de 2021.

12. BOE núm. 126, de 27 de mayo de 2021.

13. DOUE núm. 1689, de 12 de julio de 2024.

las personas físicas en lo que respecta al tratamiento de datos personales por las instituciones, órganos y organismos de la Unión, y a la libre circulación de esos datos, y por el que se derogan el Reglamento (CE) núm. 45/2001 y la Decisión núm. 1247/2002/CE.[14]

Prestaremos atención a la postura de la doctrina que se ha pronunciado en relación con las decisiones automatizadas con la finalidad de obtener una visión conjunta sobre dicho aspecto. Posteriormente, obtendremos unas conclusiones de interés.

II. INTELIGENCIA ARTIFICIAL Y DECISIONES AUTOMATIZADAS: ALGUNAS CONSIDERACIONES

Tras la aprobación del Reglamento (UE) 2024/1689 la doctrina ha prestado atención al análisis de los elementos que las Administraciones públicas en España tienen que tener en cuenta cuando automaticen el funcionamiento y la prestación de servicios públicos utilizando sistemas de inteligencia artificial y los recursos que deben disponer para realizarlo de forma adecuada.[15]

Previamente a la normativa comunitaria mencionada, en España, se ha publicado el Real Decreto 817/2023, de 8 de noviembre, que establece un entorno controlado de pruebas para el ensayo del cumplimiento de la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial[16] que define el sistema de inteligencia artificial (artículo 3) como un sistema diseñado para funcionar con un cierto nivel de autonomía y que, basándose en datos de entradas proporcionadas por máquinas o por personas, infiere cómo lograr un conjunto de objetivos establecidos utilizando estrategias de aprendizaje automático o basadas en la lógica y el conocimiento, y genera información de salida, como contenidos (sistemas de inteligencia artificial generativos), predicciones, recomendaciones o decisiones, que influyan en los entornos con los que interactúa.

Esta norma se anticipó a la aprobación del actual Reglamento (UE) 2024/1689[17], en cuyo artículo 26. 11, se refiere a las obligaciones de los responsables del despliegue de sistemas de inteligencia artificial de alto riesgo, y en concreto las obligaciones de transparencia de los proveedores y responsables

14. DOUE núm. 295, de 21 de noviembre de 2018.

15. Sigo la exposición de A. Cerrillo i Martínez, «Cerrillo i Martínez, A. «El impacto del Reglamento de Inteligencia Artificial en las Administraciones Públicas», *Revista Jurídica de les Illes Balears*, núm. 26, 2024, p. 75. Disponible en: https://revistajuridicaib.tirant.com/index.php/rjib/article/view/6/6 (Consultado el 27 de marzo de 2025).

16. BOE núm. 268, de 9 de noviembre de 2023. También resulta de interés Orden TDF/619/2024, de 18 de junio, por la que se crea y regula el Consejo Asesor Internacional en Inteligencia Artificial (BOE núm. 150, de 21 de junio de 2024).

17. Véase: AA. VV. *Comentarios al Reglamento Europeo de Inteligencia Artificial*, M. Barrio Andrés (dir.), La Ley, Madrid, 2024; AA.VV. *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, L. Cotino Hueso y P. Simó Castellanos (coord.), Aranzadi La Ley, Cizur Menor, 2024.

del despliegue de determinados sistemas de inteligencia artificial, e indica que sin perjuicio de lo indicado en el artículo 50 de la mencionada norma, los responsables del despliegue de los sistemas de inteligencia artificial de alto riesgo a los que se refiere el anexo III que tomen decisiones o ayuden a tomar decisiones relacionadas con personas físicas informarán a las personas físicas de que están expuestas a la utilización de los sistemas de inteligencia artificial de alto riesgo.

En el caso de los sistemas de inteligencia artificial de alto riesgo que se utilicen a los efectos de la garantía los responsables del despliegue de los sistemas de inteligencia artificial de alto riesgo a que se refiere el anexo III que tomen decisiones o ayuden a tomar decisiones relacionadas con personas físicas informarán a las personas físicas de que están expuestas a la utilización de los sistemas de inteligencia de alto riesgo. En el caso de los sistemas de inteligencia artificial de alto riesgo que se utilicen a los efectos de la garantía del cumplimiento del Derecho, se aplicará el artículo 13 de la Directiva (UE) 2016/680 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales por parte de las autoridades competentes para fines de prevención, investigación, detección o enjuiciamiento de infracciones penales o de ejecución de sanciones penales, y a la libre circulación de dichos datos y por la que se deroga la Decisión Marco 2008/977/JAI del Consejo.[18]

Respecto a las decisiones automatizadas y la infracción de los derechos fundamentales en concreto la protección de los datos personales, este derecho se garantiza en el Reglamento (UE) 2016/679 y (UE) 2018/1725 y la Directiva (UE) 2016/680, además de tenerse en cuenta la Directiva 2002/58/CE del Parlamento Europeo y del Consejo, de 12 de julio de 2002, relativa al tratamiento de los datos personales y a la protección de la intimidad en el sector de las comunicaciones electrónicas[19] que protege la vida privada y la confidencialidad de las comunicaciones, también estableciendo condiciones para cualquier almacenamiento de datos personales y no personales en los equipos terminales, y el acceso desde estos.

Estos actos legislativos de la Unión constituyen la base para un tratamiento de datos sostenible y responsable, también cuando los conjuntos de datos son personales y no personales.

El Reglamento (UE) 2024/1689 no pretende afectar a la aplicación de la normativa en materia de datos personales, incluidas las funciones y competencias de las autoridades de supervisión independientes competentes para vigilar el cumplimiento de dichos instrumentos. Tampoco afecta a las obligaciones de los proveedores y los responsables del despliegue de sistemas de inteligencia artificial en su papel de responsables o encargados del tratamiento de datos derivadas del Derecho de la Unión o nacional en materia de protección de da-

18. DOUE L 119/89, de 4 de mayo de 2016.
19. DO L 201, de 31 de julio de 2002.

tos personales en la medida en que el diseño, el desarrollo o el uso de sistemas de inteligencia artificial impliquen el tratamiento de datos personales.

Sigue indicando el citado Reglamento (UE) 2024/1689 que conviene aclarar que los interesados siguen disfrutando de todos los derechos y garantías que les confiere dicho Derecho de la Unión, incluidos los derechos relacionados con las decisiones individuales totalmente automatizadas, como la elaboración de perfiles. Unas normas armonizadas para la introducción en el mercado, la puesta en servicio y la utilización de sistemas de inteligencia artificial establecidas en virtud de este Reglamento deben facilitar la aplicación efectiva y permitir el ejercicio de los derechos y otras vías de recurso de los interesados garantizados por el Derecho de la Unión en materia de protección de datos personales, así como de otros derechos fundamentales.

Pueden existir casos específicos en los que los sistemas de inteligencia artificial referidos en ámbitos predefinidos especificados en el Reglamento no entrañen un riesgo considerable de causar un perjuicio a los intereses jurídicos amparados por dichos ámbitos, dado que no influyen sustancialmente en la toma de decisiones o no perjudican dichos intereses sustancialmente. A efectos del presente Reglamento, por sistema de inteligencia artificial que no influye sustancialmente en el resultado de la toma de decisiones debe entenderse un sistema de inteligencia artificial que no afecta al fondo, ni por consiguiente al resultado, de la toma de decisiones, ya sea humana o automatizada.

Un sistema de inteligencia artificial que no influye sustancialmente en el resultado de la toma de decisiones podría incluir situaciones en las que se cumplen una o varias de las siguientes condiciones.

La primera de dichas condiciones debe ser que el sistema de inteligencia artificial esté destinado a realizar una tarea de procedimiento delimitada, como un sistema de inteligencia artificial que transforme datos no estructurados en datos estructurados, un sistema de inteligencia artificial que clasifique en categorías los documentos recibidos o un sistema de inteligencia artificial que se utilice para detectar duplicados entre un gran número de aplicaciones. La naturaleza de esas tareas es tan restringida y limitada que solo presentan riesgos limitados que no aumentan por la utilización de un sistema de inteligencia artificial en un contexto que un anexo al presente Reglamento recoja como uso de alto riesgo.

La segunda condición debe ser que la tarea realizada por el sistema de inteligencia artificial esté destinada a mejorar el resultado de una actividad previa llevada a cabo por un ser humano, que pudiera ser pertinente a efectos de las utilizaciones de alto riesgo enumeradas en un anexo del presente Reglamento. Teniendo en cuenta esas características, el sistema de inteligencia artificial solo añade un nivel adicional a la actividad humana, entrañando por consiguiente un riesgo menor. Esa condición se aplicaría, por ejemplo, a los sistemas de inteligencia artificial destinados a mejorar el lenguaje utilizado en documentos ya redactados, por ejemplo, en lo referente al empleo de un tono profesional o de un registro lingüístico académico o a la adaptación del texto a una determinada comunicación de marca.

La tercera condición debe ser que el sistema de inteligencia artificial esté destinado a detectar patrones de toma de decisiones o desviaciones respecto de patrones de toma de decisiones anteriores. El riesgo sería menor debido a que el sistema de inteligencia artificial se utiliza tras una valoración humana previamente realizada y no pretende sustituirla o influir en ella sin una revisión adecuada por parte de un ser humano. Por ejemplo, entre los sistemas de inteligencia artificial de este tipo, se incluyen aquellos que pueden utilizarse para comprobar a posteriori si un profesor puede haberse desviado de su patrón de calificación determinado, a fin de llamar la atención sobre posibles incoherencias o anomalías. La cuarta condición debe ser que el sistema de IA esté destinado a realizar una tarea que solo sea preparatoria de cara a una evaluación pertinente a efectos de los sistemas de IA enumerados en el anexo del presente Reglamento, con lo que la posible repercusión de los resultados de salida del sistema sería muy escasa en términos de representar un riesgo para la subsiguiente evaluación. Esa condición comprende, entre otras cosas, soluciones inteligentes para la gestión de archivos, lo que incluye funciones diversas tales como la indexación, la búsqueda, el tratamiento de texto y del habla o la vinculación de datos a otras fuentes de datos, o bien los sistemas de IA utilizados para la traducción de los documentos iniciales. En cualquier caso, debe considerarse que los sistemas de IA utilizados en casos de alto riesgo enumerados en un anexo del presente Reglamento presentan un riesgo significativo de menoscabar la salud y la seguridad o los derechos fundamentales si el sistema de IA conlleva la elaboración de perfiles en el sentido del artículo 4, punto 4, del Reglamento (UE) 2016/679, del artículo 3, punto 4, de la Directiva (UE) 2016/680 o del artículo 3, punto 5, del Reglamento (UE) 2018/1725. Para garantizar la trazabilidad y la transparencia, los proveedores que, basándose en las condiciones antes citadas, consideren que un sistema de IA no es de alto riesgo, deben elaborar la documentación de la evaluación previamente a la introducción en el mercado o la entrada en servicio de dicho sistema de IA y facilitarla a las autoridades nacionales competentes cuando estas lo soliciten. Dichos proveedores deben tener la obligación de registrar el sistema en la base de datos de la UE creada en virtud del presente Reglamento. Con el fin de proporcionar orientaciones adicionales sobre la aplicación práctica de las condiciones con arreglo a las cuales los sistemas de IA enumerados en un anexo del presente Reglamento no se consideran, con carácter excepcional, de alto riesgo, la Comisión debe, previa consulta al Consejo de IA, proporcionar directrices que especifiquen dicha aplicación práctica, completadas por una lista exhaustiva de ejemplos prácticos de casos de uso de sistemas de IA que sean de alto riesgo y de casos de uso que no lo sean.

El artículo 14 del Reglamento (UE) 2024/1689 se refiere a la supervisión humana y establece que los sistemas de inteligencia artificial de alto riesgo se diseñarán y desarrollarán de modo que puedan ser vigilados de manera efectiva por personas físicas durante el período que estén en uso, lo que incluye dotarlos de herramientas de interfaz humano-máquina adecuadas.

El objetivo de la supervisión humana será prevenir o reducir al mínimo los riesgos para la salud, la seguridad o los derechos fundamentales que pueden

surgir cuando se utiliza un sistema de inteligencia artificial de alto riesgo conforme a su finalidad prevista o cuando se le da un uso indebido razonablemente previsible, en particular cuando dichos riesgos persistan a pesar de la aplicación de otros requisitos establecidos en la presente sección.

Las medidas de supervisión serán proporcionales a los riesgos, al nivel de autonomía y al contexto de uso del sistema de inteligencia artificial de alto riesgo, y se garantizarán bien mediante uno de los siguientes tipos de medidas, bien mediante ambos:

a) las medidas que el proveedor defina y que integre, cuando sea técnicamente viable, en el sistema de inteligencia artificial de alto riesgo antes de su introducción en el mercado o su puesta en servicio;
b) las medidas que el proveedor defina antes de la introducción del sistema de inteligencia artificial de alto riesgo en el mercado o de su puesta en servicio y que sean adecuadas para que las ponga en práctica el responsable del despliegue.

A los efectos de la puesta en práctica de los dispuesto en los apartados 1, 2 y 3, el sistema de inteligencia artificial de alto riesgo se ofrecerá al responsable del despliegue de tal modo que las personas físicas a quienes se encomiende la supervisión humana pueda, según proceda y de manera proporcionada a:

a) entender adecuadamente las capacidades y limitaciones pertinentes del sistema de inteligencia artificial de alto riesgo y poder vigilar debidamente su funcionamiento, por ejemplo, con vistas a detectar y resolver anomalías, problemas de funcionamiento y comportamientos inesperados;
b) ser conscientes de la posible tendencia a confiar automáticamente o en exceso en los resultados de salida generados por un sistema de inteligencia artificial de alto riesgo («sesgo de automatización»), en particular con aquellos sistemas que se utilizan para aportar información o recomendaciones con el fin de que personas físicas adopten una decisión;
c) interpretar correctamente los resultados de salida del sistema de inteligencia artificial de alto riesgo, teniendo en cuenta, por ejemplo, los métodos y herramientas de interpretación disponibles;
d) decidir, en cualquier situación concreta, no utilizar el sistema de inteligencia artificial de alto riesgo o descartar, invalidar o revertir los resultados de salida que este genere;
e) intervenir en el funcionamiento del sistema de inteligencia artificial de alto riesgo o interrumpir el sistema pulsando un botón de parada o mediante un procedimiento similar que permita que el sistema se detenga de forma segura.

En el caso de los sistemas de inteligencia artificial de alto riesgo mencionados en el anexo III, punto 1, letra a), las medidas a que se refiere el apartado 3 de este artículo 14 garantizarán, además, que el responsable del despliegue

no actúe ni toma ninguna decisión basándose en la identificación generada por el sistema, salvo si al menos dos personas físicas con la competencia, formación y autoridad necesarias han verificado y confirmado por separado dicha identificación.

El requisito de la verificación por parte de al menos dos personas físicas por separado no se aplicará a los sistemas de inteligencia artificial de alto riesgo utilizados con fines de garantía del cumplimiento del Derecho, de migración, de control fronterizo o de asilo cuando el Derecho nacional o de la Unión considere que la aplicación de este requisito es desproporcionada.

El Anteproyecto de Ley para el buen uso y la gobernanza de la inteligencia artificial menciona que los sistemas de inteligencia artificial pueden llegar a funcionar con un elevado grado de autonomía y autoaprendizaje, y que una vez hayan sido entregados pueden inferir los sistemas por sí mismos recomendaciones, decisiones, predicciones o contenidos, sin que sea preciso conocer y programar previamente qué datos y qué valoraciones originaron esos resultados ofrecidos por el sistema.

Se remite al artículo 26.11 del Reglamento (UE) 2024/1689 relativo a la obligación de informar a las personas físicas de su exposición a la utilización del sistema cuando se trate de un sistema de alto riesgo de los descritos en el Anexo III de dicho Reglamento que participe en decisiones relacionadas con estas personas.

III. DECISIONES AUTOMATIZADAS Y SESGOS DISCRIMINATORIOS

La aplicación de la inteligencia artificial a través de decisiones automatizadas puede provocar sesgos discriminatorios[20] en los sujetos y la consiguiente infracción de los derechos fundamentales.[21]

El Reglamento (UE) 2016/679 [22] establece que las decisiones automatizadas y la elaboración de perfiles sobre la base de categorías particulares de datos

20. J. L. Bustelo Gracia, «Sesgos de género y raciales en la IA: implicaciones éticas y legales del reconocimiento facial», *Derecho y Economía de la Integración*, núm. 13, 2024, pp. 11 y sigs. Disponible en: https://revistas.colex.es/index.php/derechoeconomiaintegracion/article/view/280/468 (Consultado el 24 de marzo de 2025).

A. J. Coaquira Flores, «Los sesgos algorítmicos en la toma de decisiones automatizas: Retos y oportunidades para el sistema jurídico peruano», *Informática y Derecho: Revista Iberoamericana de Derecho Informático*, núm. 15, 2, 2024, pp. 159 y sigs. Disponible en: https://dialnet.unirioja.es/descarga/articulo/9887893.pdf (Consultado el 24 de marzo de 2025).

21. S. Duro Carrión, «Aspectos fundamentales de la igualdad de trato y no discriminación en la educación, medios de comunicación social y publicidad, internet y redes sociales, inteligencia artificial y mecanismos de toma de decisión automatizados», en A. V. Sempere Navarro y M. B. García Gil (dir.), *Una visión transversal del derecho a la igualdad: Ley 15/2022, de 12 de julio*, Sepín, Madrid, 2023, pp. 83 y sigs.

22. Mas ampliamente: A. Palma Ortigosa, «Decisiones automatizadas en el RGPD. El uso de algoritmos en el contexto de la protección de datos», *Revista General de Derecho Administrativo*,

personales únicamente deben permitirse en condiciones específicas, y que se plasma en el artículo 22 al indicar que 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.

Las Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679 adoptadas el 3 de octubre de 2017 revisadas por última vez y adoptadas el 6 de febrero de 2018[23] indican que el responsable del tratamiento no puede obviar las disposiciones del artículo 22 inventándose una participación humana. Por ejemplo, si alguien aplica de forma rutinaria perfiles generados automáticamente a personas sin que ello tenga influencia real alguna en el resultado, esto seguiría siendo una decisión basada únicamente en el tratamiento automatizado.

Para ser considerada como participación humana, el responsable del tratamiento debe garantizar que cualquier supervisión de la decisión sea significativa, en vez de ser únicamente un gesto simbólico.

Debe llevarse a cabo por parte de una persona autorizada y competente para modificar la decisión.

Como parte del análisis, debe tener en cuenta todos los datos pertinentes.

Como parte de la Evaluación de Impacto relativa a la Protección de Datos (EIPD), el responsable del tratamiento debe identificar y registrar el grado de participación humana en el proceso de toma de decisiones y en qué punto se produce esta

Como señala la Agencia Española de Protección de Datos[24] muchas decisiones automatizadas en realidad involucran cierto grado de intervención humana, sin embargo, para considerarse como tal, tiene que ser activa y no solo un gesto simbólico, es decir, tiene que tener un grado determinado de relevancia y capacidad.

---

núm. 50, 2019; *Decisiones automatizadas y protección de datos: Especial atención a los sistemas de inteligencia artificial*, Dykinson, Madrid, 2022; Roig i Batalla, A. *Las garantías frente a las decisiones automatizadas: del Reglamento general de Protección de Datos a la gobernanza algorítmica*, Bosch, Barcelona, 2020.

23. *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679 adoptadas el 3 de octubre de 2017 revisadas por última vez y adoptadas el 6 de febrero de 2018*, p. 23. Disponible en: https://www.aepd.es/documento/wp251rev01-es.pdf (Consultado el 1 de abril de 2025).

24. Agencia Española de Protección de Datos, *Evaluación de la intervención humana en las decisiones automatizadas*, 4 de marzo de 2024. Disponible en: https://www.aepd.es/prensa-y-comunicacion/blog/evaluacion-de-la-intervencion-humana-en-las-decisiones-automatizadas (Consultado el 21 de marzo de 2025), y añade que: «La STJUE de 7 diciembre 2023, asunto C-634/21, «SCHUFA», establece que también es decisión automatizada en sentido art.22 del RGPD, cuando es la generación automática de un valor a partir de datos personales es transmitido por el responsable del tratamiento a un tercero, también responsable del tratamiento, y este tercero, de un modo determinante, basa una decisión sobre la persona en dicho valor». Véase: L. Cotino Hueso, «Holanda: «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», *La Ley privacidad*, núm. 4, 2020.

La Ley 12/2021, de 28 de septiembre, por la que se modifica el texto refundido de la Ley del Estatuto de los Trabajadores, aprobado por el Real Decreto Legislativo 2/2015, de 23 de octubre, para garantizar los derechos laborales de las personas dedicadas al reparto en el ámbito de plataformas digitales[25] introduce una nueva letra d) en el artículo 64.4 en cuanto a los derechos de información y consulta y competencias que expresa lo siguiente que el comité de empresa, con la periodicidad que proceda en cada caso, tendrá derecho a 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[26].

La Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación[27], en su artículo 23 sobre inteligencia artificial y mecanismos de toma de decisión automatizados, se refiere a la Estrategia Nacional de Inteligencia Artificial, la Carta de Derechos Digitales[28] y de las iniciativas europeas respecto a la inteligencia artificial, las administraciones públicas favorecerán la puesta en marcha de mecanismos para que los algoritmos involucrados en la toma de decisiones que se utilicen en las administraciones públicas tengan en cuenta criterios de minimización de sesgos, transparencia y rendición de cuentas, siempre que sea factible técnicamente. En estos mecanismos se incluirán su diseño y datos de entrenamiento, y abordarán su potencial impacto discriminatorio.[29]

---

25. BOE núm. 233, de 29 de septiembre de 2021.

26. A. Todolí Signes, «La gobernanza colectiva de la protección de datos: algoritmos, decisiones automatizadas y discriminación», en *El futuro del trabajo: cien años de la OIT: comunicaciones,* Ministerio de Trabajo, Migraciones y Seguridad Social, Subdirección General de Información Administrativa y Publicaciones, Madrid, 2019, pp. 1637 y sigs.; «La reputación digital de los trabajadores: Perfiles y decisiones automatizadas», en M. Bauzá Reilly (coord.), L. Cotino Hueso (dir.), *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, Thomson Reuters Aranzadi, Cizur Menor, 2022, pp. 301 y sigs.; P. Vargas Martínez, «¿Ser o no ser (objeto de decisiones automatizadas)? Evolución legislativa y jurisprudencia del concepto a nivel transnacional», en R. Santillán Santa Cruz (coord.), M. J. Sánchez Cano y J. Martínez Calvo (dir.), *El derecho privado en el marco de los objetivos de desarrollo sostenible: una panorámica global*, Colex, A Coruña, 2024; R. Vela Díaz, «Digitalización y nuevos trámites automatizados: las decisiones algorítmicas impregnan la actuación de la Administración Laboral y de Seguridad Social», *Trabajo y derecho: nueva revista de actualidad y relaciones laborales*, núm. 83, 2021.

27. BOE núm. 167, de 13 de julio de 2022. Se complementa con Ley Orgánica 6/2022, de 12 de julio, complementaria de la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación, de modificación de la Ley Orgánica 10/1995, de 23 de noviembre del Código Penal (BOE núm. 167, de 13 de julio de 2022).

28. Gobierno de España, *Carta de Derechos Digitales*, 2021. Disponible en: https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/140721-Carta_Derechos_Digitales_RedEs.pdf (Consultado el 21 de marzo de 2025).

29. Se puede consultar: Sánchez Hernández, J. «Posthumanismo, tecnología y evolución generacional de los derechos humanos: hacia la protección contra la discriminación algorítmica y el uso transparente y responsable de la IA», *Revista general de derecho constitucional*, núm. 40, 2024; Soriano Arnanz, A. «Decisiones automatizadas y discriminación: aproximación y propuestas generales», *Revista General de Derecho Administrativo*, núm. 56, 2021.

Para ello, se promoverá la realización de evaluaciones de impacto que determinen el posible sesgo discriminatorio.

También incide el citado precepto que las administraciones públicas, en el marco de sus competencias en el ámbito de los algoritmos involucrados en procesos de toma de decisiones, priorizarán la transparencia en el diseño y la implementación y la capacidad de interpretación de las decisiones adoptadas por los mismos.

Las administraciones públicas y las empresas promoverán el uso de una inteligencia artificial ética, confiable y respetuosa con los derechos fundamentales, siguiendo especialmente las recomendaciones de la Unión Europea en este sentido.

Además se indica que se promoverá un sello de calidad de los algoritmos.

La Estrategia Nacional de Inteligencia Artificial 2024[30] menciona como sistemas de inteligencia artificial de alto riesgo (HRAIS) los que pueden conducir a un riesgo significativo para la salud, la seguridad o los derechos fundamentales como es la contratación, promoción, evaluación, acceso a créditos y seguros de vida y salud, la identificación biométrica (salvo la mera identificación del usuario final y las prácticas prohibidas), las infraestructuras críticas según la Directiva (UE) 2022/2557 del Parlamento Europeo y del Consejo de 14 de diciembre de 2022 relativa a la resiliencia de las entidades críticas y por la que se deroga la Directiva 2008/114/CE del Consejo[31], otros productos ya regulados por normas armonizadas UE, como pueden ser los dispositivos médicos, ascensores, vehículos autónomos, entre otros).

En el ámbito de la salud, el Reglamento (UE) 2025/327 indica que debe prohibirse cualquier intento de utilizar datos de salud electrónicos para medidas perjudiciales para las personas físicas, tales como aumentar las primas de seguro, desarrollar actividades perjudiciales para las personas físicas relacionadas con el empleo, las pensiones o el sector bancario, incluidas las hipotecas sobre bienes inmuebles, anunciar productos o tratamientos y automatizar la toma de decisiones individuales.

La Ley 2/2025, de 2 de abril, para el desarrollo e impulso de la inteligencia artificial en Galicia[32] define, en su artículo 4, lo que se considera como actuación administrativa semiautomatizada. Es aquella previsión o recomendación formulada por un sistema de inteligencia artificial en el marco de un procedimiento administrativo o de prestación de un servicio público que será utilizada por un empleado público en el marco de un procedimiento administrativo o de prestación de un servicio público.

---

30. Ministerio para la Transformación Digital y de la Función Pública, *Estrategia Nacional de Inteligencia Artificial 2024*. Disponible en: https://portal.mineco.gob.es/es-es/digitalizacionIA/Documents/Estrategia_IA_2024.pdf (Consultado el 21 de marzo de 2025).

31. DOUE núm. 33, de 27 de diciembre de 2022.

32. DOG núm. 66, de 4 de abril de 2025. El Anteproyecto se puede consultar en: https://ficheiros-web.xunta.gal/transparencia/normativa-tramitacion/facenda-administracion-publica/FAC-lei-intelixencia-artificial-cas.pdf (Consultado el 26 de marzo de 2025).

El artículo 12 regula la reserva de humanidad y de revisión humana. Se indica que en el ámbito de la administración y del sector público en el ámbito de la Comunidad Autónoma gallega podrán emplear sistemas de inteligencia artificial o modelos de inteligencia artificial de uso general tanto en su actividad material o técnica como en la adopción de actos administrativos formalizados, tanto de trámite como resolutorios de acuerdo con lo indicado en el precepto, y de conformidad con lo establecido en la legislación estatal de aplicación.

En el caso de uso de sistemas de inteligencia artificial o modelos de inteligencia artificial de uso general que sirvan de apoyo o fundamento para la adopción de actos o decisión administrativas, se adoptarán las garantías necesarias a los efectos de mitigar cualquiera sesgos por parte del órgano competente resolutorio. En ningún caso, tales actuaciones en las que se empleen sistemas de inteligencia artificial o modelos de inteligencia artificial de uso general constituirán de por sí decisiones o actos administrativos sin validación por el titular del órgano competente.

En los casos de uso de sistemas de inteligencia artificial o modelos de inteligencia artificial de uso general que sirvan para la adopción de actos administrativos formalizados, tanto de trámite como resolutorios, de manera automatizada sin intervención humana directa de acuerdo con lo que indica el artículo 76 de la Ley 4/2019, de 17 de julio, de administración digital de Galicia[33], deberán tratarse de actos administrativos que no requieran de una valoración subjetiva de las circunstancias concurrentes o de una interpretación jurídica.

33. BOE núm. 229, de 24 de septiembre de 2019. El precepto referido establece respecto a la actuación administrativa automatizada, que la Administración general y las entidades públicas instrumentales del sector público autonómico promoverán el fomento de actuaciones administrativas automatizadas cuando se trate de actos o actuaciones respecto a los cuales los criterios de análisis o decisión puedan integrarse en un programa que realice la actuación automatizada.

La firma electrónica de las actuaciones administrativas automatizadas podrá realizarse mediante los siguientes sistemas de firma electrónica, de acuerdo con el artículo 42 de la Ley 40/20215, de 1 de octubre, de régimen jurídico del sector público (BOE núm. 236, de 2 de octubre de 2015): a) Sello electrónico de administración pública, órgano, organismo público o entidad de derecho público, basado en un certificado electrónico reconocido o cualificado que reúna los requisitos exigidos por la legislación de firma electrónica; b) Código seguro de verificación vinculado al sector público autonómico, permitiéndose en todo caso la comprobación de la integridad del documento mediante el acceso a la sede electrónica de la Xunta de Galicia.

En el caso de firma electrónica con código seguro de verificación, deberá asegurarse la autenticidad e integridad del documento durante toda su vigencia. A este fin, y al objeto de favorecer la interoperabilidad, se podrá superponer al documento firmado un sello electrónico basado en un certificado electrónico reconocido o cualificado.

Las actuaciones administrativas automatizadas deberán declararse mediante una resolución conjunta del órgano competente para la definición de las especificaciones, programación, mantenimiento, supervisión y control de calidad y, en su caso, auditoría del sistema de información y de su código fuente, así como del órgano responsable a efectos de impugnación. En esta resolución se especificará la identificación de tales órganos y los sistemas de firma utilizados, en su caso, para la actuación administrativa automatizadas.

Se publicará en la sede electrónica de la Xunta de Galicia y en el «Diario Oficial de Galicia» el texto íntegro de las resoluciones indicadas en el apartado anterior.

En caso contrario, no podrán realizarse actuaciones administrativas automatizadas a través del uso de sistema de inteligencia artificial o modelos de inteligencia artificial de uso general excepto que se cumplan todos los siguientes requisitos:

a) El órgano competente de la actuaciones administrativa automatizada aprobará previamente y serán incorporadas en la resolución conjunta a la que se refiere el artículo 76.4 de la Ley 4/2019, las instrucciones administrativas que permitan concretar los requisitos necesarios para definir de forma detallada e inequívoca los casos comunes a los que resulte de aplicación.
b) El órgano competente en materia de tecnologías de la información y de la comunicación, innovación y desarrollo tecnológico preparará el diseño tecnológico del sistema de comunicación, innovación y desarrollo tecnológico preparará el diseño. tecnológico del sistema de inteligencia artificial en el que se basará la actuación administrativa automatizada, que respete la norma correspondiente reguladora del procedimiento y las instrucciones administrativas indicadas en el apartado anterior, que no permita la alteración no supervisada del funcionamiento del sistema o modelo, y que proporcione información sencilla y fácil de entender sobre su funcionamiento para permitir a los afectados comprender y cuestionar el resultado.

En la regulación de los procedimientos administrativos para la adopción de decisiones administrativas automatizadas en las que se empleen sistemas de inteligencia artificial o modelos de inteligencia artificial de uso general se preverá el momento, modo y alcance de la intervención de personas físicas para garantizar el cumplimiento de los principios y derechos recogidos en la presente ley.

En todo caso, en los casos en los que las decisiones, previsiones o recomendaciones generadas por sistemas de inteligencia artificial o modelos de inteligencia artificial de uso general tengan un impacto irreversible o de difícil reversión, o impliquen actuaciones que puedan generar riesgos para la vida o integridad física o psicosocial de los individuos, será necesaria una validación de una persona física en el proceso decisorio, así como una decisión humana final.

Sin perjuicio de los correspondientes recursos administrativos o acciones judiciales, se reconocerá el derecho a presentar sugerencias o quejas relativas al funcionamiento de los propios sistemas de inteligencia artificial o modelos de inteligencia artificial de uso general empleados por la Administración autonómica.

El artículo 23 establece respecto a la transparencia y supervisión humana que todas las personas en las que en su relación con la Administración General de la Comunidad Autónoma de Galicia y su sector público intervengan sistemas de inteligencia artificial o modelos de inteligencia artificial de uso general tendrán derecho a:

— Recibir la debida información clara y comprensible con posibilidad del uso de iconos o símbolos fácilmente reconocibles sobre: el carácter auto-

matizado de las interacciones y de las decisiones, en particular, a saber si está interactuando con un sistema de inteligencia artificial o modelo de inteligencia artificial de uso general; la configuración general del sistema, tipos de decisiones, recomendaciones o predicciones que se pretende hacer y las consecuencias de su uso para las personas afectadas; la racionalidad y la lógica del sistema de inteligencia o modelo de inteligencia artificial de uso general; la identificación de la titularidad del sistema de inteligencia artificial o modelo de inteligencia artificial de uso general; el grado de contribución o participación del sistema de inteligencia artificial o modelo de inteligencia artificial de uso general y del empleado público en el proceso de toma de decisión, predicción o recomendación; las categorías de datos personales utilizados por los sistemas de inteligencia artificial y modelos de inteligencia artificial de uso general y su origen o fuentes; las medidas de seguridad, de no discriminación y de fiabilidad adoptadas; el modo de ejercitar el derecho a la transparencia y supervisión humana, así como otros derechos que le asistan.

— Recibir una explicación de la decisión, recomendación o previsión tomada por los sistemas de inteligencia artificial o modelos de inteligencia artificial de uso general. Las personas afectadas por decisiones, previsiones o recomendaciones efectuadas por sistemas de inteligencia artificial y modelos de inteligencia artificial de uso general, de no serles ofrecida de oficio, podrán exigir una explicación con la información debida respecto de los factores, criterios y procedimientos que incidan en dichas decisiones, previsiones o recomendaciones, en particular sobre la ponderación de los criterios para la adopción de la decisión en su caso particular, el nivel de intervención humana en la adopción de la decisión que le afecta y los mecanismos por medio de los que puede reclamar contra la decisión adoptada.

  En caso de que la decisión administrativa final se separe de la propuesta de decisión, de la previsión o de las recomendaciones efectuadas por sistemas de inteligencia artificial y modelos de inteligencia artificial de uso general, se ofrecerá una explicación debidamente motivada de las razones que justifican tal separación.

— Exigir la intervención de un empleado público en el proceso de adopción de una decisión, previsión o recomendación por parte de un sistema de inteligencia artificial, cuando la misma pueda producir efectos relevantes o que impacten de manera significativa en sus intereses.

El diseño de los sistemas de inteligencia artificial en el artículo 43 establece que cuando el sistema de inteligencia artificial tenga por finalidad la adopción de actos administrativos automatizados, el diseño del sistema de inteligencia artificial recogerá los requisitos exigidos por el artículo 41.2 de la Ley 40/2015 y el artículo 76 de la Ley 4/2019.

El alcance del uso de sistemas de inteligencia artificial y modelos de inteligencia artificial de uso general en la toma de decisiones según el artículo 50

determina que la Administración General de la Comunidad Autónoma de Galicia y su sector público podrá adoptar actos administrativos, de forma automatizada y sin la intervención directa de un empleado público, mediante sistemas de inteligencia artificial, en el marco de un procedimiento administrativo, cuando ejerzan potestades regladas y, excepcionalmente, potestades discrecionales, en caso de que la margen de discrecionalidad se agotara previamente en el diseño del sistema de inteligencia artificial.

La Administración General de la Comunidad Autónoma de Galicia y su sector público podrá apoyar su toma de decisiones en las previsiones o recomendaciones emitidas por sistemas de inteligencia artificial o modelos de inteligencia artificial de uso general.

IV. CONCLUSIONES

Actualmente, la regulación a nivel europeo del uso de la inteligencia artificial en la Unión Europeo está constituida por el Reglamento (UE) 2024/1689. Dicha norma representa un paso decisivo en la construcción de un marco jurídico armonizado sobre la inteligencia artificial en la Unión Europea. Su aprobación complementa el Reglamento 2016/679 y otros instrumentos legislativos europeos, configurando un entramado regulador que busca equilibrar la innovación tecnológica con la protección de los derechos fundamentales. Este Reglamento introduce principios, obligaciones y mecanismos de control específicos para los sistemas de inteligencia artificial, en especial aquellos catalogados como de alto riesgo, lo que fortalece la seguridad jurídica en su aplicación por entidades públicas y privadas. A su vez, la normativa incorpora criterios de proporcionalidad y especificidad técnica que reflejan una comprensión más matizada del papel que juega la inteligencia artificial en los diferentes contextos sociales, jurídicos y administrativos.

El uso creciente de sistemas de inteligencia artificial, especialmente en procesos de toma de decisiones automatizadas, plantea riesgos significativos que han sido reconocidos por el legislador europeo. Entre ellos se encuentran los sesgos discriminatorios, la opacidad algorítmica, la falta de rendición de cuentas y la posible vulneración de derechos fundamentales como el derecho a la intimidad, la igualdad de trato y el acceso a una tutela judicial efectiva. En este sentido, el Reglamento (UE) 2024/1689 insiste en la necesidad de preservar la dignidad humana y garantizar el control sobre los procesos algorítmicos que pueden incidir directamente en la vida de las personas. Estos riesgos no solo son teóricos, sino que se han materializado en distintos ámbitos como el laboral, el sanitario, el financiero y el administrativo, por lo que se justifica la imposición de obligaciones específicas para mitigar su impacto.

Una de las principales garantías introducidas por el nuevo Reglamento (UE) 2024/1689 es la obligación de incorporar mecanismos de supervisión humana significativa sobre los sistemas de inteligencia artificial de alto riesgo. Esta supervisión no puede limitarse a una intervención simbólica o meramente formal;

debe ser efectiva, deliberada y con capacidad real para modificar o revertir decisiones automatizadas. Esta exigencia busca evitar una delegación completa de funciones decisorias en sistemas opacos. La participación humana se erige como un requisito indispensable para validar la legitimidad de las decisiones algorítmicas, especialmente cuando estas afectan derechos subjetivos o intereses esenciales de las personas, asegurando que exista siempre una instancia de control consciente y responsable.

El diseño y desarrollo de los sistemas de inteligencia artificial debe responder a criterios técnicos robustos, pero también a principios éticos y jurídicos. La normativa europea establece que dichos sistemas deben ser comprensibles, trazables y auditables. Esto implica que sus resultados deben poder ser explicados a los afectados de manera clara y accesible, permitiendo así el ejercicio efectivo de derechos como el de oposición, corrección o impugnación. La transparencia no es un valor abstracto, sino una condición práctica para el control democrático de la tecnología, especialmente en sectores donde la automatización puede generar desigualdades o reforzar dinámicas de exclusión. El Reglamento (UE) 2024/1689 prevé además que se elaboren evaluaciones de impacto, documentación técnica y protocolos de uso que garanticen una rendición de cuentas continua.

Junto con la legislación europea, las normas nacionales y autonómicas —como la Ley Orgánica 3/2018 o la Ley 2/2025 de Galicia— desempeñan un papel crucial en la operacionalización de los principios del Reglamento (UE) 2024/1689. Estas disposiciones adaptan los estándares europeos a los contextos institucionales y organizativos específicos, incorporando criterios sobre el uso interno de los sistemas de inteligencia artificial, el control administrativo de las decisiones automatizadas, y la necesidad de proporcionar explicaciones accesibles a los ciudadanos. El desarrollo legislativo español ha avanzado en establecer derechos específicos frente al uso de algoritmos en el ámbito laboral, sanitario y de los servicios públicos, evidenciando una preocupación creciente por la protección de los derechos fundamentales en un entorno de digitalización acelerada.

La creciente complejidad de los sistemas de inteligencia artificial requiere no solo marcos normativos sólidos, sino también estructuras de gobernanza que integren diferentes perspectivas técnicas, jurídicas y sociales. La gobernanza algorítmica debe ser dinámica y capaz de adaptarse a la evolución tecnológica constante, incluyendo mecanismos de consulta con expertos, órganos supervisores, sociedad civil y titulares de derechos. Además, debe promover una cultura institucional orientada a la evaluación continua de riesgos, la minimización del impacto discriminatorio, y la incorporación de principios de justicia algorítmica en todas las fases del ciclo de vida de los sistemas de inteligencia artificial. La transparencia, la interoperabilidad y la rendición de cuentas deben formar parte del núcleo de dicha gobernanza.

Finalmente, resaltar la importancia de consolidar una inteligencia artificial ética, confiable y legalmente segura. Los sistemas de IA no deben estar al servicio exclusivo de la eficiencia o del interés económico, sino que deben orientarse hacia la protección y promoción de los derechos fundamentales, la inclu-

sión social y la justicia distributiva. La normativa europea y española converge en esta dirección, fomentando el desarrollo de tecnologías que respeten la dignidad humana, eviten prácticas discriminatorias y refuercen los valores democráticos. Asimismo, se aboga por la creación de instrumentos como sellos de calidad algorítmica, evaluaciones ex ante y ex post, y canales efectivos de reclamación, como garantías efectivas para los usuarios y afectados por decisiones automatizadas.

BIBLIOGRAFÍA

AA. VV. *Comentarios al Reglamento Europeo de Inteligencia Artificial*, M. Barrio Andrés (dir.), La Ley, Madrid, 2024.

AA.VV. *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, L. Cotino Hueso y P. Simó Castellanos (coord.), Aranzadi La Ley, Cizur Menor, 2024.

Agencia Española de Protección de Datos, *Evaluación de la intervención humana en las decisiones automatizadas*, 4 de marzo de 2024. Disponible en: https://www.aepd.es/prensa-y-comunicacion/blog/evaluacion-de-la-intervencion-humana-en-las-decisiones-automatizadas (Consultado el 21 de marzo de 2025).

Águila Martínez, J. y Dorado Ferrer, X. «El modelo de madurez en analítica de datos y el régimen jurídico de las decisiones automatizadas en el sector privado», en R. Oliver Cuello (dir.), *Las tecnologías de la información en la actividad empresarial. Aspectos legales y fiscales*, Thomson Reuters Aranzadi, Cizur Menor, 2022.

Álvarez Fernández, E. «Sobre las decisiones automatizadas en el ámbito laboral y el derecho a la privacidad de los trabajadores», *Diario La Ley*, núm. 10587, 2024.

Bustelo Gracia, J. L. «Sesgos de género y raciales en la IA: implicaciones éticas y legales del reconocimiento facial», *Derecho y Economía de la Integración*, núm. 13, 2024, pp. 11-31. Disponible en: https://revistas.colex.es/index.php/derechoeconomiaintegracion/article/view/280/468 (Consultado el 24 de marzo de 2025).

Cerrillo i Martínez, A. «El impacto del Reglamento de Inteligencia Artificial en las Administraciones Públicas», *Revista Jurídica de les Illes Balears*, núm. 26, 2024, pp. 73-105. Disponible en: https://revistajuridicaib.tirant.com/index.php/rjib/article/view/6/6 (Consultado el 27 de marzo de 2025).

Cherñavsky, N. «Inteligencia artificial y «big data» jurídica. Decisiones automatizadas. Modelos de aplicación en general y en el ámbito jurídico», en C. Ch. Sueiro (coord.), M. Alfredo Riquert (dir.), *Sistema penal e informática, volumen 7. Ciberdelitos, evidencia digital, TICs*, Hammurabi, Argentina, 2024, pp. 33-54.

Coaquira Flores, A. J. «Los sesgos algorítmicos en la toma de decisiones automatizas: Retos y oportunidades para el sistema jurídico peruano»,

Informática y Derecho: Revista Iberoamericana de Derecho Informático, núm. 15, 2, 2024, pp. 159-170. Disponible en: https://dialnet.unirioja.es/descarga/articulo/9887893.pdf (Consultado el 24 de marzo de 2025).

COTINO HUESO, L. «Holanda: «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», *La Ley privacidad*, núm. 4, 2020.

Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679 adoptadas el 3 de octubre de 2017 revisadas por última vez y adoptadas el 6 de febrero de 2018. Disponible en: https://www.aepd.es/documento/wp251rev01-es.pdf (Consultado el 1 de abril de 2025).

DURO CARRIÓN, S. «Aspectos fundamentales de la igualdad de trato y no discriminación en la educación, medios de comunicación social y publicidad, internet y redes sociales, inteligencia artificial y mecanismos de toma de decisión automatizados», en A. V. Sempere Navarro y M. B. García Gil (dir.), *Una visión transversal del derecho a la igualdad: Ley 15/2022, de 12 de julio*, Sepín, Madrid, 2023, pp. 83-94.

Gobierno de España, *Carta de Derechos Digitales*, 2021. Disponible en: https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/140721-Carta_Derechos_Digitales_RedEs.pdf (Consultado el 21 de marzo de 2025).

GALETTA, D. U. «Derechos y garantías concretas respecto del uso por los poderes públicos de decisiones automatizadas e inteligencia artificial: La importancia de las garantías en el procedimiento administrativo», en M. Bauzá Reilly (coord.), L. Cotino Hueso (dir.), *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, Thomson Reuters Aranzadi, Cizur Menor, 2022, pp. 171-192.

GAMERO CASADO, E. «Sistemas automatizados de toma de decisiones en el Derecho Administrativo Español», *Revista General de Derecho Administrativo*, núm. 63, 2023.

GARRIGA DOMÍNGUEZ, A. «Decisiones automatizadas basadas en algoritmos y protección de datos personales», en J. A. Viguri Cordero (coord.), *Las cláusulas específicas del Reglamento General de Protección de Datos en el ordenamiento jurídico español: cuestiones clave de orden nacional y europeo*, Tirant lo Blanch, Valencia, 2021, pp. 333-360.

GRANERO, H. R. «Derechos y garantías concretas frente al uso de inteligencia artificial y decisiones automatizadas, especialmente en el ámbito judicial y de aplicación de la ley», en M. Bauzá Reilly (coord.), L. Cotino Hueso (dir.), *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, Thomson Reuters Aranzadi, Cizur Menor, 2022, pp. 107-137.

Ministerio para la Transformación Digital y de la Función Pública, *Estrategia Nacional de Inteligencia Artificial 2024*. Disponible en: https://portal.mineco.gob.es/es-es/digitalizacionIA/Documents/Estrategia_IA_2024.pdf (Consultado el 21 de marzo de 2025).

Moreno Martínez, J. A. «Las garantías en la protección de datos personales en las ciudades inteligentes «smart cities» e implantación de decisiones automatizadas con uso de la inteligencia artificial», *Revista de derecho privado*, núm. 108, 2024, pp. 61-89.

Olivares Olivares, B. D. «La toma de las decisiones automatizadas y el derecho fundamental a la protección de datos de carácter personal», *Quincena fiscal*, núm. 15-16, 2022.

Olivares Olivares, B. D. «La teoría de las garantías adecuadas en materia de protección de datos y sus implicaciones respecto de la toma de decisiones automatizadas en la Administración tributaria», en B. D. Olivares Olivares (coord.), *La inteligencia artificial en la relación entre los obligados y la administración tributaria: retos ante la gestión tecnológica*, La Ley, Madrid, 2022, pp. 240-263.

Orofino, A. G. «Decisiones automatizadas, transformación de la administración y prestación de servicios públicos digitales», en *La digitalización en los servicios públicos: garantías de acceso, gestión de datos, automatización de decisiones y seguridad*, Marcial Pons, Madrid, 2023, pp. 139-162.

Palma Ortigosa, A. «Decisiones automatizadas en el RGPD. El uso de algoritmos en el contexto de la protección de datos», *Revista General de Derecho Administrativo*, núm. 50, 2019.

Palma Ortigosa, A. *Decisiones automatizadas y protección de datos: Especial atención a los sistemas de inteligencia artificial*, Dykinson, Madrid, 2022.

Pérez Bernabeu, B. «Decisiones automatizadas en la Administración tributaria», en B. D. Olivares Olivares (coord.), *La inteligencia artificial en la relación entre los obligados y la administración tributaria: retos ante la gestión tecnológica*, La Ley, Madrid, 2022, pp. 146-162.

Pérez Bernabeu, B. «Los contribuyentes ante las decisiones automatizadas de la administración tributaria», en A. M. Pita Gandal y L. A. Malvárez Pascual (coord.), *La digitalización en los procedimientos tributarios y el intercambio automático de información*, Thomson Reuters Aranzadi, Cizur Menor, 2023, pp. 235-254.

Pérez Daudí, V. «El precedente judicial: la previsibilidad de la sentencia y la decisión automatizada del conflicto», *Revista iberoamericana de Derecho Procesal. RIDP*, núm. 2, 2020, pp. 141-184.

Pérez Daudí, V. «El precedente judicial. La previsibilidad de la sentencia y la decisión automatizada del conflicto», en M. J. Ariza Colmenarejo (dir.), *Revisión del Sistema de Fuentes y su repercusión en el Derecho Procesal*, Dykinson, Madrid, 2021, pp. 199-222.

Pérez Sarabia, M. «La prohibición a la toma de decisiones automatizadas, clave esencial para los derechos de los pacientes en el ámbito de la medicina», en C. Villegas Delgado y P. Martín Ríos (ed.), *El derecho en la encrucijada tecnológica: Estudios sobre derechos fundamentales, nuevas tecnologías e inteligencia artificial*, Tirant lo Blanch, Valencia, 2022, pp. 357-376.

Ramón Fernández, F. «La utilización de la inteligencia artificial para la verificación de la identidad y el control de presencia mediante sistemas

biométricos en el entorno laboral: algunas cuestiones», *La regulación de la inteligencia artificial y el derecho del trabajo. Retos en materia de dirección algorítmica del trabajo*, Aranzadi La Ley, Cizur Madrid, 2025, pp. 93-124.

Rivas Vallejo, M. P. «Decisiones automatizadas y discriminación en el trabajo», *Revista General de Derecho del Trabajo y de la Seguridad Social*, núm. 66, 2023.

Rodríguez Cardo, I. A. «Decisiones automatizadas y discriminación algorítmica en la relación laboral: ¿hacia un Derecho del Trabajo de dos velocidades?», *Revista española de derecho del trabajo*, núm. 253, 2022, pp. 135-188.

Rodríguez Escanciano, S. y Álvarez Cuesta, H. «La toma de decisiones automatizadas en el marco de la relación laboral: otra vuelta de tuerca al poder de dirección y vigilancia empresarial», en L. Gamarra Vilchez (coord.), J. E. López Ahumada (dir.), *La gobernanza de los derechos digitales de las personas trabajadoras*, Cinca, Madrid, 2023, pp. 109-144.

Roig i Batalla, A. *Las garantías frente a las decisiones automatizadas: del Reglamento general de Protección de Datos a la gobernanza algorítmica*, Bosch, Barcelona, 2020.

Sánchez Hernández, J. «Posthumanismo, tecnología y evolución generacional de los derechos humanos: hacia la protección contra la discriminación algorítmica y el uso transparente y responsable de la IA», *Revista general de derecho constitucional*, núm. 40, 2024.

Sobrino García, I. «Las decisiones automatizadas en el sector público: conflictos entre la protección de datos y la inteligencia artificial», en P. R. Bonorino Ramírez y R. Fernández Acevedo (dir.), *Nuevas normatividades: inteligencia artificial, derecho y género*, Thomson Reuters Aranzadi, Cizur Menor, 2021, pp. 17-34.

Soriano Arnanz, A. «Decisiones automatizadas y discriminación: aproximación y propuestas generales», *Revista General de Derecho Administrativo*, núm. 56, 2021.

Todolí Signes, A. «La gobernanza colectiva de la protección de datos: algoritmos, decisiones automatizadas y discriminación», en *El futuro del trabajo: cien años de la OIT: comunicaciones,* Ministerio de Trabajo, Migraciones y Seguridad Social, Subdirección General de Información Administrativa y Publicaciones, Madrid, 2019, pp. 1637-1656.

Todolí Signes, A. «La reputación digital de los trabajadores: Perfiles y decisiones automatizadas», en M. Bauzá Reilly (coord.), L. Cotino Hueso (dir.), *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, Thomson Reuters Aranzadi, Cizur Menor, 2022, pp. 301-315.

Vargas Martínez, P. «¿Ser o no ser (objeto de decisiones automatizadas)? Evolución legislativa y jurisprudencia del concepto a nivel transnacional», en R. Santillán Santa Cruz (coord.), M. J. Sánchez Cano y J. Martínez Calvo (dir.), *El derecho privado en el marco de los objetivos de desarrollo sostenible: una panorámica global*, Colex, A Coruña, 2024.

VELA DÍAZ, R. «Digitalización y nuevos trámites automatizados: las decisiones algorítmicas impregnan la actuación de la Administración Laboral y de Seguridad Social», *Trabajo y derecho: nueva revista de actualidad y relaciones laborales*, núm. 83, 2021.

VIGURI CORDERO, J. A. «La transparencia como reto en la toma de decisiones automatizadas en los procesos migratorios», en L. S. Heredia Sánchez (coord.), A. Ortega Giménez (dir.), *Protección de datos, transparencia, asociacionismo y voluntariado: buenas prácticas de actuación con el colectivo migrante*, Aranzadi, Cizur Menor, 2023, pp. 113-114.

Inteligencia Artificial y algoritmos en la gestión de pandemias: lecciones para el derecho público a partir de la Covid-19[1]

Gabriele Vestri
Profesor Ayudante Doctor de Derecho Procesal (acred. Titular)
Universidad de Sevilla
Presidente del Observatorio Sector Público e Inteligencia Artificial
www.ospia.org

SUMARIO: I. INTELIGENCIA ARTIFICIAL EN LA GESTIÓN DE LA PANDEMIA: UNA INTRODUCCIÓN GENERAL. II. CUESTIONES CONSTITUCIONALES DEL USO DE IA EN PANDEMIAS. III. IMPLICACIONES DE DERECHO ADMINISTRATIVO EN LA IMPLEMENTACIÓN DE IA. IV. EN CONCLUSIÓN: LECCIONES APRENDIDAS Y PROPUESTAS PARA FUTUROS MARCOS NORMATIVOS. VI. BIBLIOGRAFÍA.

1. El texto se produce como actividad investigadora del Proyecto de Investigación I+D+i «Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas» (PID2022- 136439OB-I00). Entidad financiadora: Ministerio de Ciencia e Innovación, Gobierno de España. Investigador principal: Jorge Castellanos Claramunt.

Asimismo, este trabajo se enmarca dentro de las actividades desarrolladas como miembro del Grupo de Investigación «BrAIn» del Observatorio Sector Público e IA.

En este trabajo se utiliza la forma neutra de ciudadano, interesado, usuario, etcétera, en singular o en plural, para referirse tanto a hombres como a mujeres, es decir: cuando se escribe ciudadano, interesado, usuarios, etcétera, puede ser ciudadana, interesada, usuaria, etcétera. El control gramatical, sintáctico y lingüístico de este trabajo ha sido efectuado con sistemas de inteligencia artificial generativa.

I. INTELIGENCIA ARTIFICIAL EN LA GESTIÓN DE LA PANDEMIA: UNA INTRODUCCIÓN GENERAL

La pandemia de COVID-19 planteó un desafío sin precedentes para los sistemas sanitarios y las autoridades públicas. En ese contexto crítico, la inteligencia artificial (IA) surgió como una aliada estratégica en la gestión de la crisis. Desde los inicios del brote epidémico, se depositaron grandes expectativas en que la ciencia de datos y la IA pudieran emplearse para apoyar la lucha contra el coronavirus, contribuyendo a «llenar las lagunas» que la ciencia tradicional aún no alcanzaba a resolver[2]. Los gobiernos de todo el mundo —incluyendo España— incorporaron con rapidez diversas aplicaciones y sistemas de IA como parte de sus medidas de respuesta sanitaria, aprovechando la capacidad de estas herramientas para procesar enormes volúmenes de información en tiempo real y así optimizar la toma de decisiones. Ahora bien, el despliegue de dichas soluciones tecnológicas tuvo que alinearse con exigencias jurídico-legales y éticas. Incluso en situaciones excepcionales debían respetarse los marcos normativos vigentes en materia de Derechos fundamentales y protección de datos.

Una de las herramientas más difundidas por el sector público durante la pandemia fueron las aplicaciones móviles de rastreo de contactos. Estas apps, generalmente desarrolladas por autoridades sanitarias, sirvieron para identificar y notificar a personas que habían estado en proximidad de un caso positivo de COVID-19, con el fin de *romper* las cadenas de contagio de forma más eficiente que los métodos tradicionales. Por ejemplo, la aplicación española Radar COVID[3] se diseñó siguiendo estándares técnicos estrictos de privacidad avalados por la Comisión Europea: ningún usuario puede ser identificado ni localizado, ya que no se recopila dato personal alguno y todo el proceso de detección de contactos se ejecuta de forma local en el teléfono del usuario. Además, tanto la instalación de la app como la comunicación voluntaria de un diagnóstico positivo por parte del ciudadano se mantenía siempre bajo control del individuo y nunca imponiéndose de manera coercitiva.

De esta manera, Radar COVID y aplicaciones similares lograron alertar anónimamente a miles de personas sobre posibles exposiciones de riesgo, complementando los procedimientos tradicionales de rastreo epidemiológico[4]. Muchos países adoptaron aplicaciones análogas. Incluso gigantes tecnológicos como Google y Apple proporcionaron, en abril de 2020, una interfaz estándar (API) de notificación de exposiciones para facilitar la interoperabilidad, requiriendo,

2. Sobre IA como herramienta de ayuda y predicción en la pandemia véase: R. Benjamins. «Hacia una IA sostenible: una perspectiva 360 incluyendo negocio, sociedad, ética y cambio climático», en W., Arellano Toledo (dir.), *Derecho, ética e inteligencia artificial*, Tirant lo Blanch, Valencia, 2023, p. 24.

3. Radar COVID finalizó su actividad el 9/10/2022.

4. Véase: I. Cerrato, N., González Alarcón, «¿Ya tenemos suficientes apps?, Blog «Abierto al público», 1 de septiembre de 2020. En: https://shorturl.at/sQFTW (short url) [Consultado el 18 de julio de 2025].

eso sí, que cualquier app que la utilizase fuera respaldada por una entidad pública de salud. En América Latina, Uruguay fue pionero al integrar dicha API en su aplicación nacional, cumpliendo las condiciones de validación oficial exigidas por las tiendas de aplicaciones[5].

Este tipo de iniciativas públicas ilustran, de forma somera, cómo la IA y las tecnologías móviles, cuando utilizadas con salvaguardas adecuadas, pueden reforzar significativamente la capacidad de las autoridades para vigilar la propagación de virus protegiendo al mismo tiempo la privacidad individual.

Más controvertido fue el uso de sistemas digitales de geolocalización para monitorear el cumplimiento de cuarentenas obligatorias y restricciones de movimiento. En ciertos países, particularmente de Asia, los gobiernos hicieron un uso intensivo de la tecnología móvil para asegurar que las personas confinadas respetaran las órdenes sanitarias. Corea del Sur, por ejemplo, implementó un sistema por el cual se enviaban alertas automáticas a las autoridades sanitarias cuando una persona bajo aislamiento incumplía la cuarentena y se desplazaba a lugares concurridos (como transporte público o centros comerciales)[6]. En Taiwán, a las personas infectadas se les proporcionaba un teléfono móvil equipado con GPS para seguir en todo momento su ubicación, de modo que la policía pudiera verificar que no se alejaran de su domicilio durante el periodo de aislamiento obligatorio[7]. Singapur combinó herramientas de geolocalización con llamadas telefónicas aleatorias y visitas domiciliarias para controlar a quienes tenían órdenes de permanecer en casa, en una política de vigilancia estricta aceptada socialmente en ese país[8]. Incluso en algunas democracias occidentales se estudiaron medidas similares: Italia llegó a desarrollar una aplicación piloto capaz de rastrear el itinerario de una persona contagiada y advertir a sus contactos, asegurando supuestamente el anonimato de los datos; mientras que Israel autorizó temporalmente la localización individual de teléfonos móviles para alertar a ciudadanos expuestos.

Si bien estos métodos de vigilancia digital se consideraron eficaces para frenar la difusión del virus, causaron ciertos dilemas jurídicos y éticos.

En Europa, por ejemplo, se enfatizó que las normas de protección de datos personales (como el Reglamento General de Protección de Datos de la UE) mantienen plena vigencia durante la crisis sanitaria, de modo que cualquier uso gubernamental de datos de geolocalización o sistemas de reconocimiento biométrico debe contar con base legal, salvaguardias de privacidad y supervisión independiente.

La experiencia internacional evidenció así la necesidad de equilibrar la eficacia epidemiológica de estas tecnologías con la protección de los Derechos fundamentales.

---

5. *Ídem*.

6. Véase, Consejo de Europa, «La IA y el control del coronavirus Covid-19» en https://rb.gy/29in9w (short url). [Fecha consulta: 14 de julio de 2025].

7. *Ídem*.

8. *Ídem*.

Otra aplicación de la IA desplegada por el sector público fue la creación de asistentes virtuales inteligentes para informar y orientar a la ciudadanía. Mediante *chatbots* basados en procesamiento de lenguaje natural, distintos gobiernos y organismos públicos—en el caso español, por ejemplo, diputaciones, ayuntamientos y Universidades— lograron atender millones de consultas relacionadas con la COVID-19 de forma automática, uniforme y disponible las 24 horas[9]. Asimismo, el Gobierno implementó en abril de 2020 el Hispabot-Covid19 (desarrollado por la Secretaría de Estado de Digitalización e IA en colaboración con el Ministerio de Sanidad), un *chatbot* oficial en WhatsApp diseñado para responder preguntas frecuentes sobre el coronavirus con información veraz y actualizada[10]. El sistema había sido entrenado con cientos de preguntas potenciales (sobre síntomas, medidas de prevención, protocolos de actuación, etc.) y podía interpretarlas incluso si eran formuladas de diversas maneras, garantizando siempre una contestación basada en las fuentes oficiales.

En conjunto, estos asistentes virtuales representaron una herramienta comunicativa innovadora que reforzó la transparencia y el alcance de los mensajes de salud pública durante la crisis, demostrando cómo la IA puede ayudar a mejorar la interacción entre el gobierno y la ciudadanía en momentos de alta demanda informativa.

La pandemia obligó asimismo a acelerar la transformación digital de los servicios médicos, impulsando la telemedicina—en algunos casos apoyada en IA— como alternativa segura a las consultas presenciales[11]. Ante las medidas de distanciamiento social y la necesidad de proteger tanto a pacientes como a personal sanitario, numerosos países adoptaron plataformas de atención médica remota para diagnósticos iniciales, seguimiento de casos leves y consultas rutinarias[12].

La IA y la digitalización —más en general— tuvieron un papel importante integrándose en varias de estas soluciones. Se emplearon algoritmos de *triaje virtual* para evaluar síntomas reportados por los pacientes y determinar si requerían atención especializada. También se habilitaron sistemas de diagnóstico asistido por IA (por ejemplo, para interpretar radiografías o tomografías enviadas electrónicamente), que apoyaron a los médicos en la detección de neumonía,

9. Véase: G. Vestri. «Inteligencia artificial y chatbots en el proceso de transición digital del sector público» en M.D. Cervilla Garzón, M.A. Blandino Garrido (dirs.), y A. Nieto Cruz (coord.), *Declaración de voluntad en un entorno virtual*, Thomson Reuters-Aranzadi, Pamplona. España, 2021, págs., 461-473. Véase también: G. Bonales et al. «Chatbot como herramienta comunicativa durante la crisis sanitaria COVID-19 en España», *ComHumanitas. Revista Científica de Comunicación*, vol. 11, núm. 3, 2020, pp. 1-22.

10. Véase: https://covid19.gob.es/hispabot-covid19 [Consultado el 18 de julio de 2025].

11. Sobre la introducción de las e-consultas en el periodo de la pandemia, véase: Pavón de Paz, I., et. al. La e-consulta como herramienta para la relación entre Atención Primaria y Endocrinología. Impacto de la epidemia por COVID-19 en su uso. *Journal of Healthcare Quality Research*, Vol. 37, núm. 3 (mayo-junio), 2022, p. 157. DOI: 10.1016/j.jhqr.2021.10.006.

12. Se recomienda la lectura de: L. Cotino Hueso. «Inteligencia artificial, big data y aplicaciones contra la COVID-19: privacidad y protección de datos». *IDP. Internet, Derecho y Política*, núm. 31, 2020, pp. 1-17.

típica del virus, u otras complicaciones de COVID-19. Gracias a la telemedicina y a estas aplicaciones inteligentes, fue posible mantener la continuidad de la atención sanitaria incluso en periodos de confinamiento estricto.

No obstante, su implementación generalizada planteó retos de mucha importancia, como por ejemplo garantizar la calidad del servicio a distancia, la seguridad de los datos clínicos transmitidos y la igualdad de acceso tecnológico para toda la población, cuestiones que los legisladores y autoridades sanitarias tuvieron que considerar cuidadosamente al fomentar estas modalidades asistenciales de emergencia.

En paralelo, los equipos de respuesta a la pandemia aprovecharon el poder de la IA para predecir la evolución del virus y anticiparse a los acontecimientos. En un entorno muy cambiante, resultó decisivo contar con proyecciones fiables sobre la expansión geográfica y temporal de los contagios, la tasa de reproducción del virus o el impacto de las intervenciones (confinamientos, toques de queda, etc.). Para ello, se desarrollaron modelos epidemiológicos basados en *machine learning*, alimentados con datos masivos que iban desde históricos sanitarios y demográficos hasta patrones de movilidad ciudadana o incluso tendencias en búsquedas de internet.

Los sistemas de IA aplicados a la modelización de la pandemia se convirtieron así en un insumo ventajoso para la toma de decisiones basada en evidencia, si bien su fiabilidad dependía de la calidad de los datos y de su adecuada interpretación por expertos humanos antes de traducirlos en medidas concretas.

Por último, la IA también fue aprovechada para optimizar la gestión de los recursos sanitarios durante la pandemia, un periodo marcado por la escasez y la presión asistencial. Mediante algoritmos de optimización y análisis predictivo, los gestores públicos pudieron asignar de forma más racional insumos críticos (como camas de cuidados intensivos, ventiladores mecánicos, pruebas diagnósticas y equipos de protección personal) allí donde más se necesitaban en cada momento. En el ámbito hospitalario, por ejemplo, se implementaron modelos predictivos de ocupación que permitieron anticipar con varios días de margen el aumento de ingresos de pacientes COVID-19, ayudando a redirigir personal y abrir unidades adicionales antes de que se desbordaran los servicios. En definitiva, la experiencia del COVID-19 dejó patente que la IA puede ser una poderosa herramienta al servicio de la salud pública y el bienestar común, pero su implementación por los poderes públicos ha de integrarse en un marco jurídico sólido que conjugue la eficacia sanitaria con el respeto a las libertades individuales.

Muy involucrado en esta pandemia, se vio el Derecho público y especialmente el Derecho constitucional y administrativo que de alguna manera fueron las disciplinas que tuvieron que justificar las bases legales para las actuaciones de los poderes públicos. En este trabajo interesa verificar especialmente como las dos disciplinas jurídicas mencionadas reaccionaron a la pandemia. Asimismo, interesa analizar lo ocurrido jurídicamente para futuras situaciones de crisis. Todo ello, a la luz de que —como se señaló en otro trabajo— durante la pandemia se tuvo la sensación de que nos vimos obligados a enfrentarnos a un

sistema jurídico adormecido, el cual, en ocasiones, se vio forzado a generar respuestas ágiles para resolver un problema real con la urgencia que exigía el momento[13].

II. CUESTIONES CONSTITUCIONALES DEL USO DE IA EN PANDEMIAS

La utilización de herramientas de IA durante la gestión de la pandemia de COVID-19 ha planteado importantes retos constitucionales en España. Diversos Derechos fundamentales se han visto involucrados en este contexto, y las respuestas gubernamentales han tenido que equilibrar la eficacia sanitaria con el respeto a las garantías constitucionales. En particular, han entrado en tensión derechos como la privacidad y protección de datos personales, la libertad de circulación y de reunión, la libertad de expresión, así como el principio de igualdad y no discriminación. La Constitución Española sigue tutelando estos derechos incluso en situaciones de crisis, de modo que cualquier restricción excepcional debe estar jurídicamente fundada y resultar necesaria y proporcionada al objetivo de salud pública perseguido. Así, la emergencia sanitaria no suspende el Estado de Derecho: la supremacía constitucional exige que la gestión de la pandemia mediante IA respete el núcleo de los Derechos fundamentales. En este escenario y como señala Martínez García que la pandemia «nos ha abierto los ojos y nos ha hecho ver cuánta estructura de Estado hace falta para sobrevivir a una crisis sanitaria de gran magnitud»[14].

Uno de los derechos más afectados por la implementación de sistemas de IA en la pandemia ha sido el derecho a la privacidad y a la protección de datos personales (art. 18 CE). Las autoridades recopilaron datos masivos de salud y geolocalización, por ejemplo, como señalado, a través de aplicaciones móviles de rastreo de contactos y análisis de movilidad. Este seguimiento digital planteó riesgos de injerencia en la intimidad de los ciudadanos, requiriendo un cuidadoso encaje con la normativa de protección de datos. La Agencia Española de Protección de Datos enfatizó que ni siquiera una situación de emergencia «puede suponer una suspensión del derecho fundamental a la protección de datos personales», si bien la normativa de privacidad tampoco debe emplearse para obstaculizar medidas sanitarias justificadas. En otras palabras, la lucha contra la pandemia mediante IA debía encontrar un justo equilibrio: requerir ciertas intromisiones en la esfera privada, pero siempre respetando los principios de finalidad, minimización y temporalidad propios del marco legal de datos personales. Así, cualquier procesamiento masivo de datos personales con herra-

13. G. Vestri. «Catarsis o plasticidad jurídica: el Derecho público contra las cuerdas», en A. del Campo (Compilador), *Pensar la pandemia. Más allá de la sanidad y la economía*, Dykinson, Madrid, 2021, págs. 79-93.

14. E. Martínez García. «Retos de la función jurisdiccional para un mundo interdependiente y ecodependiente». *Teoría & Derecho*, núm. 37 (Derecho e inteligencia artificial), 2024, p. 249.

mientas de IA —por útil que fuese para la salud pública— debía realizarse conforme a la legislación vigente y bajo supervisión de autoridades independientes, sin que la urgencia justificase zonas de sombra legales.

Junto a la intimidad, las medidas tecnológicas incidieron en la libertad de circulación y reunión (arts. 19 y 21 CE). Durante el confinamiento, España declaró el estado de alarma para restringir drásticamente la movilidad de la población —un fenómeno sin precedentes en la etapa constitucional, por lo menos no de la envergadura que se dio durante la pandemia—, y exploró a la vez herramientas digitales para vigilar el cumplimiento de cuarentenas y distanciamiento social. El empleo de vigilancia digital, como drones policiales o aplicaciones móviles de geolocalización, podía suponer una restricción de facto al derecho de circulación (deambulación) y al derecho de reunión, al permitir monitorizar desplazamientos y disolver concentraciones no autorizadas[15]. Estas restricciones, aunque motivadas por razones imperiosas de salud pública, debían encuadrarse cuidadosamente en el marco constitucional de derechos. La Constitución no permite la suspensión general de Derechos fundamentales salvo bajo estados de excepción o sitio (art. 55 CE), que no llegaron a declararse durante la COVID-19. Por tanto, las limitaciones a la libre circulación impuestas bajo el estado de alarma debían interpretarse como limitaciones provisionales y justificadas, nunca como una abolición absoluta del derecho. El Tribunal Constitucional, al examinar posteriormente estas medidas, subrayó que incluso en una pandemia los límites a derechos como la circulación han de ser razonables y concretos según las circunstancias, no equivalentes a una privación total de libertades. En Alemania, por ejemplo, el Tribunal Constitucional Federal llegó a considerar ilegítima una prohibición general de manifestaciones, instando a ponderar caso por caso. En España, el Gobierno optó por gestionar la crisis mediante restricciones puntuales y controladas de derechos (bajo el principio de «normalidad» jurídica), sin derogar formalmente ninguna libertad fundamental. Esta elección implicó, no obstante, que las medidas tecnológicas de vigilancia y control —como el rastreo digital de movimientos ciudadanos— debían implementarse observando estrictamente el principio de proporcionalidad. Solo se podían justificar constitucionalmente si resultaban indispensables para evitar contagios y si respetaban el contenido esencial de derechos como la libre circulación. De esta manera, la tecnología no podía convertirse en excusa para que el estado pudiera vigilar de forma ilimitada. Cualquier cuarentena automatizada o control digital debía ampararse en una habilitación legal y pasar un escrutinio de necesidad, adecuación y proporcionalidad respecto del fin sanitario. Las personas conservaron, aun en confinamiento, su derecho a conocer las razones y alcances de tales restricciones, así como a impugnarlas ante la justicia ordinaria si consideraban que vulneraban sus derechos. En este sentido, señala

15. Véase: C. Márquez Carrasco, C.; J.A. Ortega Ramírez. «La COVID-19 y los desafíos de la vigilancia digital para los derechos humanos: a propósito de la app DataCOVID prevista en la Orden Ministerial SND/29/2020, de 27 de marzo», *Revista Bioética y Derecho*, núm. 50, 2020, pp. 205-220.

Porta Frutos: «La llegada del año 2020, con la aparición del Covid-19 y las consecuencias propias de una situación pandémica global —estado de alarma con restricción de movilidad incluidos— propiciaron que tanto justiciables como órganos jurisdiccionales buscaran alternativas tecnológicas que permitieran hacer más flexible y accesible la práctica procesal. En este contexto, prácticas que anteriormente eran reservadas para supuestos puntuales como la celebración de vistas de forma telemática se convirtieron en recursos utilizados diariamente por las partes. Indudablemente, este tipo de medidas permitieron que el reinicio de la actividad procesal a partir de junio de 2020 se realizase de una forma ciertamente razonable. Sobre todo, habida cuenta de la principal consecuencia de la pandemia a nivel procesal: el completo colapso de una jurisdicción civil ya de por sí absolutamente tensionada»[16].

Otro frente sensible fue la libertad de expresión e información (art. 20 CE), a propósito del uso de algoritmos para combatir la desinformación durante la pandemia. La propagación de noticias falsas o bulos sobre la COVID-19 llevó a muchas autoridades a buscar mecanismos de censura algorítmica o moderación automatizada de contenidos en internet, con la intención de frenar rumores peligrosos para la salud pública. En España, el Gobierno aprobó un Procedimiento de actuación contra la desinformación (Orden PCM/1030/2020) alineado con el Plan de Acción de la UE, que preveía coordinar medios humanos y técnicos para detectar y responder a campañas de desinformación. Si bien dicha estrategia insistía en privilegiar las contra-narrativas y la comunicación pública transparente, también suscitó preocupación en cuanto a posibles excesos. Organizaciones como *Article 19* advirtieron que considerar la desinformación un asunto de seguridad nacional podría conllevar un involucramiento excesivo de los servicios de seguridad en la vigilancia de contenidos online[17]. El despliegue de herramientas automatizadas para filtrar o bloquear información con el pretexto sanitario debía, por ende, manejarse con elevado cuidado desde la óptica constitucional. La libertad de expresión goza de una tutela reforzada en nuestra Constitución, prohibiéndose la censura previa y permitiendo solo restricciones a posteriori y motivadas (por ejemplo, por protección de la salud o el orden público, y siempre bajo control judicial). Incluso frente a informaciones erróneas, el Estado no puede erigirse en árbitro único de la verdad, ni tampoco delegar dicha función a algoritmos opacos. Cualquier limitación al flujo de información ha de perseguir fines legítimos muy concretos (evitar un peligro real para la salud pública, por ejemplo) y ser necesaria y proporcionada a ese fin. En la práctica, esto significa que una medida algorítmica que suprima determinados contenidos solo será constitucional si demuestra que ese discurso falso implica un riesgo cierto (p. ej., incitación a conductas gravemente perjudiciales para la

16. C. Porta Frutos. «Inteligencia artificial y proceso civil. Uso de los sistemas de IA en la adopción de medidas cautelares: riesgos, beneficios y perspectivas de futuro». *Revista Aranzadi de Derecho y nuevas tecnologías*, núm. 67 (enero-abril), 2025, pp. (Legalteca).

17. Véase: Article 19, «España: La acción del gobierno contra la desinformación debería contar con todas las partes interesadas» en https://shorturl.at/ZHLsb (short url). [Fecha consulta: 14 de julio de 2025].

salud) y si la intervención es la mínima indispensable para neutralizarlo. Medidas generales o indiscriminadas de filtro digital podrían equivaler a censura prohibida. Por ello, se ha abogado por enfoques alternativos como la verificación independiente de los hechos, la promoción de información veraz y la educación mediática, antes que por bloqueos automatizados. En definitiva, el empleo de IA contra la desinformación durante la pandemia debía encauzarse dentro del respeto al art. 20 CE: sin socavar el debate público abierto, con base legal clara y con garantías de revisión, de modo que el remedio tecnológico no terminase lesionando más la libertad de expresión que el propio problema de los bulos.

Por último, la adopción acelerada de sistemas de IA en la emergencia planteó interrogantes sobre la igualdad y la no discriminación (art. 14 CE). Los algoritmos empleados en distintos ámbitos —desde modelos predictivos de propagación hasta herramientas de triaje clínico automatizado— podían incorporar sesgos que generasen impactos desiguales en ciertos grupos poblacionales[18]. Existe el riesgo de que la IA, al procesar datos históricos o incompletos, reproduzca y amplifique desigualdades preexistentes. Por ejemplo, un algoritmo de asignación de recursos sanitarios podría, sin una meticulosa calibración, terminar privilegiando a unos pacientes sobre otros en base a factores cuestionables (edad, código postal, nivel socioeconómico) o pasar por alto a colectivos infrarrepresentados en los datos de entrenamiento. Este fenómeno de discriminación algorítmica quebranta el principio de igualdad efectiva, que obliga a los poderes públicos a evitar decisiones arbitrarias o arbitrariamente desiguales. El Consejo de Europa ha advertido expresamente que en el uso de tecnologías de vigilancia pandémica deben tenerse en cuenta los posibles sesgos, pues «pueden causar una discriminación importante» entre individuos[19]. Pensemos en aplicaciones móviles que solo sean accesibles para quienes poseen teléfonos inteligentes de última generación, dejando fuera (y por tanto sin alertas de exposición) a poblaciones de mayor edad o bajos ingresos; o sistemas automáticos de diagnóstico que funcionen peor con datos de ciertos grupos étnicos por estar calibrados principalmente con datos de otros grupos. Tales disparidades contradicen la exigencia constitucional de no discriminación. Por ello, durante la pandemia se insistió en que la IA aplicada a políticas públicas sanitarias debía ser auditada y supervisada para detectar y corregir sesgos (tanto en los datos como en los modelos). Además, el principio de dignidad humana (art. 10.1 CE), intrínsecamente ligado a la igualdad, impide que se trate a las personas como meros números o casos estadísticos. Incluso en la urgencia por salvar vidas, cada individuo conserva su dignidad y valor intrínseco, lo que significa que no puede quedar enteramente supeditado a una decisión mecanizada sin intervención humana. En la práctica, esto impone que los sistemas algorítmicos de apoyo a decisiones críticas (por ejemplo, a quién ingresar en una UCI cuando hay esca-

18. En este sentido véase: L. Wynants, et. al. «Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal», *BMJ (Clinical research ed.)*, 369, m1328. https://doi.org/10.1136/bmj.m1328.

19. *Op. cit.*

sez de camas) no operen de manera autónoma. Debe haber siempre un control humano que evalúe las recomendaciones de la máquina, para evitar resultados injustos o inhumanos. En palabras de Ariel Guersenzvaig y de David Casacuberta, «en caso de utilizarse, estos algoritmos no deben ser vinculantes. Deben ser siempre los profesionales sanitarios, y en ningún caso los algoritmos, quienes decidan quiénes deben o pueden ser tratados y quiénes no. Cuando se trata de decisiones cruciales para el bienestar humano, y especialmente en casos de vida o muerte, no podemos dejar que «el sistema» sea el que decida»[20]. Así se garantiza que criterios como la compasión, la individualización de cada caso y la equidad prevalezcan sobre una fría optimización numérica. En otros términos, la IA no puede ser excusa para apartarse de la igualdad y la dignidad; al contrario, su implementación debe venir acompañada de salvaguardias adicionales para que no introduzca nuevos sesgos ni menoscabe el trato humanitario debido a todas las personas.

Ahora bien, la gestión de la pandemia con apoyo de IA debe someterse a los grandes principios constitucionales que rigen la actuación de los poderes públicos. En primer lugar, el principio de jerarquía normativa (art. 9.3 CE) implica que la introducción de algoritmos y tecnologías por parte de la Administración no puede, obviamente, contravenir las normas de superior rango. La Constitución y las leyes definen el marco dentro del cual han de operar estas herramientas. Por mucho que una solución tecnológica resulte eficiente, no puede desplegarse al margen de la legalidad vigente. Así, por ejemplo, la implementación de sistemas masivos de reconocimiento facial para vigilar cuarentenas chocaría frontalmente con algunos Derechos fundamentales (honor, intimidad, propia imagen) y con la normativa de protección de datos personales —que prohíbe el tratamiento de categorías sensibles de datos biométricos salvo excepciones muy tasadas—. Hoy, además, habría que valorar la herramienta según el catálogo de sistemas de alto riesgo—como mínimo— establecido por el RIA[21].

Un principio fundamental en la materia es el de necesidad y proporcionalidad, que rige cualquier limitación de Derechos fundamentales. La proporcionalidad obliga a ponderar cuidadosamente el beneficio sanitario aportado por una herramienta de IA frente al grado de restricción de derechos que conlleva. Se compone, clásicamente, de tres sub-principios: idoneidad (la medida tecnológica debe ser apta para lograr el fin de protección de la salud), necesidad en sentido estricto (que no exista otra medida igualmente eficaz pero menos lesiva de los derechos) y proporcionalidad *stricto sensu* (un balance equilibrado entre ventajas e inconvenientes, de forma que el sacrificio al derecho no sea excesivo en comparación con el beneficio esperado). Aplicado al caso de la IA en la

20. A. Guersenzvaig, D. Casacuberta. «Los peligros del algoritmo en tiempos del coronavirus» en https://elpais.com/tecnologia/2020-03-30/los-peligros-del-algoritmo-en-tiempos-del-coronavirus.html [Fecha consulta: 14 de julio de 2025].

21. Por ejemplo, el art. 5e) del RIA, «Prácticas de IA prohibidas» señala: «la introducción en el mercado, la puesta en servicio para este fin específico o el uso de sistemas de IA que creen o amplíen bases de datos de reconocimiento facial mediante la extracción no selectiva de imágenes faciales de internet o de circuitos cerrados de televisión».

pandemia, este principio exigía, por ejemplo, evaluar si realmente un algoritmo dado mejora de forma sustancial la gestión sanitaria. Es decir, es importante probar si sin esa herramienta tecnológica, los objetivos de salud no se pueden conseguir o se conseguirían solo parcialmente y por ende si se ha calibrado que los efectos negativos sobre la privacidad, la libertad u otros derechos son lo más reducidos posible. Si la respuesta a alguna de estas preguntas fuese negativa, la medida debería atenuarse o desistirse.

De hecho, restricciones severas a la privacidad que no se demuestren esenciales para salvar vidas o sostener funciones críticas no pueden reputarse necesarias, y por tanto serían, en abstracto, inconstitucionales. La proporcionalidad, entonces, actuó como criterio guía para evitar excesos tecnológicos. Cabe destacar que este principio enlaza con la obligación de la Administración de buscar soluciones menos intrusivas siempre que sea posible. Así, antes de instaurar una vigilancia digital masiva, se debió valorar si medidas tradicionales (rastreo manual de contactos, campañas informativas, recomendaciones voluntarias), podían lograr resultados similares. Solo cuando quedó claro que la IA aportaba un valor añadido indispensable (por ejemplo, agilizar notificaciones de exposición a contagio en tiempo real) se justificó su adopción amplia, y aun así bajo condicionantes. En definitiva, el principio de necesidad y proporcionalidad sirvió de dique para que, en nombre de la eficacia contra el virus, no se implementaran herramientas algorítmicas y de IA más invasivas o restrictivas de lo estrictamente requerido por la situación epidemiológica.

Vinculado a lo anterior está el principio de dignidad humana (art. 10 CE), que debe permear cualquier actuación pública, incluida la toma de decisiones automatizadas. La dignidad es inviolable y núcleo de los Derechos fundamentales; implica reconocer a cada persona como un fin en sí misma y no un mero medio o estadística. Aplicado al uso de IA en la pandemia, el principio de dignidad exigía evitar tratos degradantes o deshumanizadores que pudieran derivarse de la automatización. Por ejemplo, se debatió éticamente sobre los algoritmos de triaje, mencionados con anterioridad, que priorizaban la atención de unos pacientes COVID-19 sobre otros[22]. Si tales algoritmos funcionaran sin supervisión humana y exclusivamente en base a criterios como la expectativa de vida o la productividad económica, podrían lesionar la dignidad de aquellos relegados, tratándolos implícitamente como menos valiosos. Para conjurar ese peligro, se recalcó que la decisión última en ámbitos sensibles debe recaer en profesionales humanos, capaces de considerar las circunstancias individuales y los valores éticos, y no en un cálculo ciego de una máquina. El propio personal médico, en numerosas guías, afirmó que los algoritmos solo tendrían un carácter orientativo y nunca vinculante, preservando así la compasión y el juicio clínico caso por caso. Hoy, esta dinámica se llamaría supervisión humana. Asi-

22. Véase: E.A. Toache, M.A. Rosales. «Preocupaciones éticas en el uso de inteligencia artificial, transparencia y derecho de acceso a la información. El caso de los chatbots en el gobierno de México, en el contexto de la Covid-19». *Estudios en Derecho a la Información*, 2023, pp. 85-111 DOI: 10.22201/iij.25940082e.2023.15.17472

mismo, la dignidad exige que ninguna persona sea sometida a decisiones puramente mecánicas en cuestiones que afectan profundamente a su vida, integridad o libertad. Esto conecta con el derecho a la autonomía personal: por ejemplo, si se utilizaban aplicaciones para vigilar el cumplimiento de cuarentenas mediante notificaciones automáticas, debía garantizarse que el ciudadano tuviera oportunidad de explicar o justificar supuestas infracciones (evitando un tratamiento como infractor sin audiencia). Igualmente, cualquier sistema de puntuación de ciudadanos (*social scoring)* para evaluar su riesgo sanitario habría chocado con la dignidad, al reducir a las personas a un número estigmatizante[23]. Afortunadamente, no se implantó nada semejante en nuestro entorno constitucional, justamente porque el valor de la dignidad actuó como límite infranqueable. Incluso en las restricciones necesarias, se procuró un trato respetuoso: por ejemplo, las sanciones por incumplir medidas anti-COVID siguieron garantizando el debido proceso y la consideración de circunstancias personales, en lugar de aplicarse automáticamente vía IA, por ejemplo.

Finalmente, el deber de transparencia y tutela judicial efectiva complementa todos los anteriores principios. La Constitución exige que cualquier restricción de derechos, aun en estado de alarma, sea conocida, motivada y revisable. Esto implica que los ciudadanos tengan acceso a la información sobre las herramientas de IA utilizadas y las razones de su adopción. Durante la pandemia, se planteó la obligación gubernamental de explicar el fundamento científico-técnico de medidas como las aplicaciones de rastreo o los algoritmos de reparto de recursos sanitarios. La transparencia y la explicabilidad contribuyen a la legitimidad: conocer cómo funciona un algoritmo de seguimiento digital se convierte en crucial para que la ciudadanía lo aceptara y colaborara. De hecho, en España se publicaron informes sobre el funcionamiento de Radar COVID y se sometió a código abierto su software, permitiendo así cierta auditoría independiente. Además, la información clara permite que quien se sienta agraviado por una medida pueda ejercer su derecho de defensa. Incluso bajo emergencia, los afectados por decisiones automatizadas deben conservar vías para recurrir o reclamar. Así lo establece, por ejemplo, el RGPD en cuanto al derecho a no ser sometido a decisiones completamente automatizadas sin posibilidad de intervención humana (art. 22 RGPD), y así lo exige el art. 24 CE sobre la tutela judicial efectiva. En la práctica, esto significó que, si alguien consideraba que una

23. Sobre la puntuación social, se pronuncia el actual RIA. El art. 5c) del RIA, «Prácticas prohibidas» prohíbe: «la introducción en el mercado, la puesta en servicio o la utilización de sistemas de IA para evaluar o clasificar a personas físicas o a colectivos de personas durante un período determinado de tiempo atendiendo a su comportamiento social o a características personales o de su personalidad conocidas, inferidas o predichas, de forma que la puntuación ciudadana resultante provoque una o varias de las situaciones siguientes: i) un trato perjudicial o desfavorable hacia determinadas personas físicas o colectivos de personas en contextos sociales que no guarden relación con los contextos donde se generaron o recabaron los datos originalmente, ii) un trato perjudicial o desfavorable hacia determinadas personas físicas o colectivos de personas que sea injustificado o desproporcionado con respecto a su comportamiento social o la gravedad de este».

cuarentena electrónica o un pase sanitario digital vulneraba sus derechos injustificadamente, podía acudir a los tribunales para obtener control jurisdiccional. Los jueces, a su vez, pudieron demandar a la Administración la documentación técnica pertinente para evaluar la legitimidad de la medida impugnada. El rendimiento de cuentas públicas fue igualmente esencial. Las autoridades debieron justificar periódicamente la continuidad de las medidas tecnológicas y demostrar resultados (por ejemplo, la eficacia real de Radar COVID en detección de casos fue objeto de escrutinio parlamentario y social, lo que llevó incluso a cuestionar si mantenía proporción coste-beneficio). Como ha señalado la doctrina, las herramientas de seguimiento digital, por valiosas que sean, «deben estar sometidas a información, transparencia, rendición de cuentas públicas y control judicial»[24]. Este estándar garantiza que ni el secreto ni la opacidad amparen posibles vulneraciones de derechos. En resumidas cuentas, la transparencia funcionó y debe funcionar como antídoto contra la arbitrariedad y la desconfianza, mientras que la supervisión judicial aseguró un remedio frente a posibles excesos o errores de la IA implementada. Ambos elementos —publicidad y control jurisdiccional— reafirman que, aun en la excepcionalidad, el Estado de Derecho permanece operativo y protector de los ciudadanos. Precisamente en este sentido se pronuncia Tejadura Tejada: «El fin no justifica los medios. La lucha contra el COVID-19 requiere adoptar medidas limitadoras de derechos fundamentales. En los peores escenarios puede resultar necesario incluso recurrir a la suspensión de algunos. Ahora bien, esas medidas deben adoptarse respetando los principios y valores esenciales del Estado de Derecho (reserva de ley y seguridad jurídica). Corresponde siempre e inexcusablemente al legislador la potestad de regular y limitar los derechos»[25].

III. IMPLICACIONES DE DERECHO ADMINISTRATIVO EN LA IMPLEMENTACIÓN DE IA

En el contexto de la pandemia es otrosí necesario analizar cómo las decisiones automatizadas se ajustaron al principio de legalidad y a las garantías del procedimiento administrativo, especialmente cuando se impusieron sanciones por incumplir restricciones apoyadas en tecnologías de vigilancia inteligente. Estas decisiones automatizadas deben respetar las garantías procedimentales tradicionales: la competencia del órgano que emite el acto, la existencia de una base legal habilitante, un procedimiento reglado y el derecho del ciudadano a ser oído antes de una decisión desfavorable. La legislación vigente en España ya prevé la figura de la actuación administrativa automatizada. El artículo 41.1

24. C. Márquez Carrasco, J.A. Ortega Ramírez, *op cit.,* pág. 219.

25. J. Tejadura Tejada. «El Estado de Derecho frente al COVID reserva de ley y derechos fundamentales», *Revista Vasca de Administración Pública (RVAP). Administrazio Publikoaren Euskal Aldizkaria*, núm. 120 (Mayo-agosto), 2021 (Ejemplar dedicado a: En recuerdo de Pablo Pérez Tremps), pág. 172.

de la Ley 40/2015 (LRJSP) la define como «cualquier acto o actuación realizada íntegramente a través de medios electrónicos por una Administración Pública en el marco de un procedimiento administrativo y en la que no haya intervenido directamente un empleado público». Ahora bien, la ley exige que se determine previamente qué órgano es responsable del sistema automatizado y será considerado emisor del acto a efectos de impugnación. Esto significa que, aunque la decisión la tome un algoritmo, jurídicamente se atribuye a una autoridad concreta. De esta forma se asegura la competencia y responsabilidad. En este mismo sentido, el acto automatizado es válido solo si un órgano habilitado aprueba su uso y asume el control y la respuesta ante eventuales recursos. No obstante, surgen dudas sobre si durante la pandemia todas las sanciones automatizadas respetaron estos requisitos. Por ejemplo, si cámaras inteligentes generaron propuestas de sanción sin un trámite de audiencia previo, se habría comprometido el derecho del ciudadano a ser escuchado. Además, la obligación de motivar los actos administrativos exige explicar las razones de la decisión, algo complejo cuando el «razonamiento» lo realiza una IA. En otras palabras, la legalidad de estos actos automatizados exige una cuidadosa adaptación procedimental: aprobación formal previa del algoritmo, notificación con indicación de recursos y garantías de audiencia y motivación, para que el procedimiento sea debido incluso en entorno tecnológico. La falta de estos cuidados podría acarrear la nulidad de decisiones automatizadas por violar principios básicos del procedimiento administrativo.

La transparencia algorítmica se ha revelado como condición indispensable del derecho a una buena administración en la era digital. Los ciudadanos tienen derecho a comprender las decisiones automatizadas que les afectan, lo que supone acceso a información sobre la lógica, criterios y fuentes de datos de los algoritmos públicos. Este derecho a entender las decisiones se vincula directamente con el artículo 41 de la Carta de Derechos Fundamentales de la UE y con la exigencia constitucional española de publicidad de las actuaciones administrativas (art. 105.b CE). Sin transparencia, la IA administrativa tiene el riesgo de entrar en una «caja negra» que mina la confianza pública. Durante la pandemia, por ejemplo, se adoptaron algoritmos para clasificar riesgos sanitarios o asignar recursos médicos. Si sus reglas permanecen ocultas, los afectados no pueden saber si han sido tratados equitativamente. La doctrina ha recalcado que la opacidad algorítmica atenta contra el derecho a buena administración, en especial en su vertiente de conocimiento de la motivación de las decisiones[26]. El problema es que a menudo concurren factores que dificultan la transparencia: la complejidad técnica de los algoritmos de aprendizaje automático, los secretos comerciales invocados por sus desarrolladores o incluso la preocupación de las

26. Véase, entre otros, G. Vestri. «La inteligencia artificial ante al desafío de la transparencia algorítmica. Una aproximación desde la perspectiva jurídico-administrativa». *Revista Aragonesa de Administración Pública*, núm. 56, 2021, págs. 368-398. 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, pp. 1-52.

autoridades de que revelar detalles permita a los ciudadanos eludir el algoritmo. Pese a estas resistencias, la tendencia normativa va hacia mayor apertura. Diversas leyes de transparencia y proyectos normativos recientes buscan imponer la publicidad activa de los algoritmos utilizados en la Administración y la explicabilidad de sus decisiones automatizadas. Por ejemplo, algunas comunidades autónomas han dado pasos pioneros: la Comunidad Valenciana, en su ley de transparencia, obliga a publicar los sistemas de decisión automatizada que emplea, desarrollando un registro público de algoritmos en colaboración con universidades; Cataluña ha llegado a difundir fichas técnicas detalladas de varios algoritmos públicos, incluyendo su entrenamiento, posibles sesgos identificados y la existencia de protocolos de revisión humana. Estas iniciativas de registro algorítmico pretenden que cualquier persona pueda saber qué sistemas automáticos operan en la gestión pública y bajo qué parámetros. La Agencia Española de Protección de Datos emitió en 2021 unas directrices sobre requisitos de auditoría de tratamientos con IA, reconociendo la necesidad de verificar externamente la calidad, equidad y legalidad de los algoritmos públicos. En definitiva, garantizar la transparencia algorítmica—mediante divulgación proactiva, evaluaciones técnicas y acceso público a la información esencial de cada sistema— es parte integral del derecho a buena administración y condición para la legitimidad de la Administración digital.

Ligado a lo anterior surge el principio de supervisión humana o «reserva de humanidad» en la toma de decisiones públicas automatizadas. La idea, respaldada por amplia doctrina, es que ciertas decisiones de la Administración no deben dejarse enteramente a las máquinas, por muy eficientes que sean[27]. En particular, las decisiones discrecionales—aquellas que implican valorar circunstancias individuales, ponderar principios o mostrarse empáticas con la situación del ciudadano— requieren un juicio humano insustituible. La IA y los algoritmos pueden ayudar en tareas regladas o repetitivas, pero cuando el poder público ha de sopesar valores sociales o aplicar criterios flexibles, la «reserva de humanidad» actúa como salvaguarda de la dignidad y justicia material en el acto administrativo. Durante la pandemia se vieron ejemplos ilustrativos: decisiones como priorizar quién recibe ciertos cuidados médicos o a quién sancionar con mayor severidad por incumplir normas podían teóricamente automatizarse, pero una aplicación ciega de algoritmos habría ignorado matices humanos (por ejemplo, las razones de una persona para saltarse un confinamiento, o factores compasivos en la atención sanitaria). Por ello, se ha propuesto establecer por ley la obligación de supervisión humana efectiva sobre los sistemas de IA administrativos, especialmente cuando estén en juego Derechos fundamentales o valoraciones extrajurídicas importantes. La Ley 40/2015 ya apuntaba en esa dirección al indicar que todo sistema automatizado debe tener una supervisión y un control de calidad humanos definidos antes de entrar en funcionamiento. Asimismo, el RIA, en el artículo 14, establece la supervisión humana como regla cardinal en el uso de herramientas de IA. Es importante señalar que embargo,

27. Véase nuevamente J. Ponce Solé, *op. cit.*

esta exigencia no debe interpretarse de forma minimalista. La práctica ha demostrado que muchas Administraciones implementan IA sin involucrar adecuadamente a juristas u otros expertos en el equipo de control. La «reserva de humanidad» no se limita a tener a un funcionario al final de la cadena que apruebe o rechace lo que propone el algoritmo; implica diseñar todo el proceso de toma de decisiones automatizado con la posibilidad de intervención y corrección humana en cada fase crítica. Además, en ámbitos sensibles podría declararse una prohibición de automatización plena: las potestades sancionadoras discrecionales o decisiones que afecten gravemente a derechos (como conceder o denegar libertades) no deberían adoptarse exclusivamente por IA, sino que requieren la aprobación final de una persona investida de autoridad. Este equilibrio busca aprovechar la eficiencia de la IA para procesar datos masivos, pero sin desistir del control humano que garantiza la legitimidad democrática y la equidad en el ejercicio de potestades públicas.

Otra cuestión crucial es la responsabilidad y el control jurídico de las decisiones automatizadas. Si un algoritmo empleado por la Administración comete un error que perjudica a un ciudadano—por ejemplo, clasificar erróneamente a alguien como contagioso, denegar indebidamente una prestación social o calcular mal una sanción—, es importante determinar quién responde. En principio, rige el marco general de responsabilidad patrimonial de las Administraciones públicas, es decir, la Administración debe indemnizar los daños causados por el funcionamiento normal o anormal de sus servicios, salvo causas de fuerza mayor. Un acto administrativo automatizado se considera emitido por la Administración responsable, así que, si el error proviene de un fallo del sistema de IA, la institución pública que lo utilizó debe asumir la responsabilidad frente al ciudadano afectado. Esto plantea a su vez la necesidad de depurar internamente responsabilidades: la Administración podrá repercutir contra el proveedor tecnológico si hubo defectos en el software, pero esto no puede ser oponible al perjudicado, quien debe tener siempre una vía de reparación expedita. Un problema práctico es la opacidad técnica: muchos algoritmos son complejos o secretos, de modo que el ciudadano perjudicado puede encontrarse con enormes dificultades para demostrar el error o ilegalidad del sistema, similar a lo que ocurría en la primera revolución industrial con las máquinas opacas. La cita clásica de Josserand—«las cosas inanimadas [...] se tornan más numerosas, mucho más terribles y también mucho más obscuras»— resuena hoy como algo que puede llegar a ser común[28]. Los procesos algorítmicos oscuros dificultan identificar la génesis del daño y la culpa administrativa en cada caso. Para contrarrestar esta situación, es clave reforzar el control administrativo y judicial de los actos apoyados por IA. En sede administrativa, los ciudadanos deben contar con

28. En 1910, hablando del desarrollo de la industria, el jurista francés Louis Josserand señalaba que a medida que la industria progresa y se transforma, los accidentes causados por objetos inanimados se vuelven más frecuentes, más graves y también más difíciles de esclarecer. Ello hace evidente la injusticia del sistema tradicional de responsabilidad: en la gran mayoría de los casos, las víctimas no pueden reconstruir el origen del accidente, averiguar sus causas ni demostrar la culpa del empresario, del conductor o de quien corresponda.

procedimientos de recurso efectivos: impugnaciones que obliguen a la Administración a revisar la decisión automatizada, a explicar su lógica y, si procede, a anularla o corregirla. La LRJSP ya exige que toda notificación de un acto automatizado indique los recursos procedentes y plazos, al igual que un acto tradicional. Sin embargo, en la práctica el ciudadano puede necesitar pruebas técnicas para desacreditar la decisión algorítmica. Aquí el papel del juez es determinante. Los tribunales deben poder escrutar el algoritmo cuando esté en entredicho un derecho, incluso llegando a ordenar la exhibición del código fuente o de informes periciales independientes que analicen su funcionamiento. Aunque choque con secretos industriales, el código fuente de un algoritmo debería tratarse como un elemento análogo al expediente administrativo y, por ende, someterlo al principio de transparencia procesal en un juicio. Solo así el juez podrá valorar si la decisión impugnada fue adoptada con arreglo a Derecho o si el algoritmo incurrió en un vicio (sesgo prohibido, error de programación, criterios contrarios a la ley, etc.). Estamos ante un nuevo reto para la administración de justicia: incorporar peritajes tecnológicos y criterios de auditoría algorítmica en la valoración de la legalidad de los actos. En todo caso, garantizar vías efectivas de impugnación es parte del Derecho fundamental a la tutela judicial efectiva, que debe preservarse también en la era digital. Los ciudadanos no pueden quedar indefensos frente a «máquinas burocráticas»; siempre ha de haber un rostro humano y un tribunal detrás de la decisión, capaces de responder por ella y de revocarla si es injusta[29].

Finalmente, la implementación de IA en la Administración púbica y de gestión exige extremar la protección de datos personales y la seguridad de la información. Durante la pandemia, se recopilaron y procesaron datos sensibles de salud a una escala inédita—desde historiales clínicos hasta geolocalizaciones de contactos— para entrenar y alimentar sistemas de seguimiento epidemiológico y decisión automatizada en salud pública. La legislación de datos impone obligaciones estrictas en estos casos, que la Administración debe cumplir incluso en situaciones de urgencia. Principios como la minimización de datos, la finalidad específica y la limitación del plazo de conservación son legalmente exigibles: solo se pueden recoger los datos necesarios para la finalidad sanitaria declarada (p. ej., trazabilidad de contagios), no usarlos para otros propósitos incompatibles (como vigilancia generalizada) y eliminarlos cuando dejen de ser pertinentes. La AEPD ha recordado que el RGPD permite excepciones por razón de interés público en salud, pero no una suspensión total de los derechos. Cualquier tratamiento masivo durante la pandemia debía evaluarse cuidadosamente y someterse a medidas técnicas y organizativas apropiadas. De hecho, la AEPD publicó en 2020 directrices específicas para adecuar los sistemas de IA a la normativa, insistiendo en la necesidad de evaluaciones de impacto (DPIA)

29. Sobre la relación entre conocimiento del código fuente y garantía de la tutela judicial efectiva, véase: G. Vestri. «El acceso a la información algorítmica a partir del caso bono social vs. Fundación ciudadana Civio». *Revista General de Derecho Administrativo*, núm. 61, octubre-2022, págs. 1-22.

previas en proyectos que combinen IA y datos sensibles[30]. Estas evaluaciones de impacto ayudan a identificar riesgos para la privacidad y mitigarles antes de desplegar la tecnología. En aplicaciones móviles de rastreo de contactos o en algoritmos de predicción de brotes, debían incorporarse mecanismos de anonimización o seudonimización y definirse claramente quién accede a la información. Además de la protección de datos, la ciberseguridad de los sistemas de IA se convirtió en una preocupación primordial. La rápida digitalización de servicios públicos durante la emergencia evidenció vulnerabilidades: algunos sistemas sanitarios o plataformas de certificación COVID sufrieron intentos de acceso no autorizado y ciberataques, buscando explotar lagunas en la seguridad. Un fallo de seguridad en un algoritmo público podría acarrear consecuencias graves, desde la filtración masiva de datos personales de salud hasta la manipulación maliciosa de los resultados que produce la IA. Por ello, como ha subrayado Almonacid Lamelas, es importante que las Administraciones integren la gestión de riesgos tecnológicos en su funcionamiento ordinario[31]. En España, el Esquema Nacional de Seguridad ya obliga a las entidades públicas a adoptar medidas de seguridad de la información acordes a estándares internacionales (control de accesos, cifrado de datos sensibles, auditorías periódicas, planes de respuesta a incidentes, etc.). Estos estándares deben aplicarse con máximo rigor cuando se trata de sistemas de IA críticos. Asimismo, es recomendable diversificar proveedores tecnológicos y evitar dependencias opacas, en el software. Así, la Administración digital debe ser tan segura como eficiente: la confiabilidad de los sistemas de IA gubernamentales se gana demostrando que protegen los datos de los ciudadanos y resisten ataques, asegurando la integridad, disponibilidad y confidencialidad de la información pública.

IV. EN CONCLUSIÓN: LECCIONES APRENDIDAS Y PROPUESTAS PARA FUTUROS MARCOS NORMATIVOS

La experiencia de la pandemia de COVID-19 ha dejado lecciones para futuros marcos normativos y actuaciones sobre el uso de IA en emergencias sanitarias. Esta crisis evidenció la necesidad de una regulación específica de la IA en contextos de salud pública. Frente a la improvisación normativa observada en la crisis del COVID-19, es conveniente establecer protocolos legales anticipados que permitan un uso eficaz de la IA sin mermar las garantías jurídicas. Tales protocolos podrían traducirse en reformas que delimiten la habilitación para el tratamiento masivo de datos personales en emergencias, definan la duración y el ámbito de las aplicaciones de rastreo digital, y prevean controles parlamen-

30. Véase E. Gamero Casado. «Sistemas automatizados de toma de decisiones en el Derecho Administrativo Español. *Revista General de Derecho Administrativo*, núm. 63, 2023, pp. 1-18.

31. V. Almonacid Lamelas. 10 riesgos de la implantación de IA en la Administración y cómo gestionarlos en No solo AYTOS, https://shorturl.at/J0wug, (short url), 1 de febrero de 2025, [Fecha consulta: 15 de julio de 2025].

tarios o judiciales incluso bajo estados de excepción. Se debate si las leyes de estados de alarma o de salud pública deben actualizarse para incluir expresamente estas herramientas, garantizando bases legales acotadas a la situación de crisis.

Por otro lado, es crucial integrar las garantías constitucionales en la transformación digital del sector público. Es imperativo incorporar salvaguardas de Derechos fundamentales en los sistemas de IA gubernamentales desde su diseño. Principios como la «privacidad desde el diseño» y la «ética desde el diseño» deben guiar los desarrollos tecnológicos del Estado. En la práctica, esto implica que las aplicaciones sanitarias que manejan datos sensibles incorporen mecanismos robustos de anonimización-seudonimización y seguridad desde su concepción, y que antes de desplegar algoritmos se realicen evaluaciones de impacto en Derechos fundamentales. Esto es, por ejemplo, lo que requiere el Convenio marco del Consejo de Europa sobre IA. Asimismo, es esencial reforzar la capacitación del funcionariado y una cultura institucional de respeto a los derechos en entornos digitales. Las lecciones del COVID-19 evidencian que se puede aprovechar la IA para maximizar la eficacia sanitaria sin renunciar a la protección de los derechos, siempre que medie un adecuado diseño normativo y técnico.

La respuesta a una pandemia trasciende fronteras nacionales, y lo mismo ocurre con muchas soluciones tecnológicas. De ahí la importancia de la cooperación internacional y la adopción de estándares globales. Se aboga por crear marcos de colaboración transnacional para desarrollar y emplear herramientas de IA compatibles con los valores democráticos y los derechos humanos. Diversas iniciativas ya sirven de referencia, como las recomendaciones de la Organización Mundial de la Salud (OMS) sobre digitalización en salud, las directrices del Comité de Ministros del Consejo de Europa para un uso responsable de la IA y el mismo el Convenio marco del Consejo de Europa sobre IA y el RIA, por lo menos en el ámbito europeo. La convergencia en estándares comunes contribuiría a un uso seguro y ético de la IA en las crisis sanitarias. Se propone crear canales de coordinación internacionales para compartir datos epidemiológicos de forma ética y eficiente, evitando vacíos legales explotables en emergencias. Por ejemplo, la armonización de criterios de privacidad en aplicaciones de rastreo entre países reforzaría la confianza pública y la efectividad de estas herramientas. En síntesis, ninguna jurisdicción actúa en el vacío; los estándares compartidos fortalecen tanto la eficacia de la respuesta a las pandemias como la protección de los derechos fundamentales.

Finalmente, para enfrentar futuras crisis sanitarias, se requiere fortalecer las instituciones y las políticas públicas a nivel interno. Esto incluye dotar a las agencias de salud pública de unidades especializadas en tecnología e inteligencia artificial, capaces de asesorar y liderar la adopción de estas herramientas durante emergencias. Debe fomentarse la colaboración interdisciplinaria—integrando a epidemiólogos, ingenieros, juristas y otros expertos—en la elaboración de planes de respuesta digital que contemplen el uso seguro y efectivo de la IA. La participación de la sociedad civil en la evaluación y control de estas

herramientas es crucial para garantizar la transparencia y legitimidad de las medidas adoptadas. Además, es fundamental robustecer los organismos de supervisión existentes, como las agencias de protección de datos o los comités de ética, dotándolos de mayores recursos y facultades para monitorear las aplicaciones de IA en emergencias. De este modo, las administraciones podrán reaccionar con rapidez ante nuevas pandemias, pero dentro de un esquema delineado que asegure la rendición de cuentas y el respeto del Estado de Derecho aun en circunstancias excepcionales.

El éxito de la IA en la gestión de pandemias depende de la formación continua de quienes la utilizan y, desde luego, de una cultura institucional comprometida con los Derechos Humanos. Una Administración pública que comprenda tanto el potencial como las limitaciones de la IA estará mejor preparada para afrontar futuras emergencias sanitarias con eficacia y legitimidad.

BIBLIOGRAFÍA

Almonacid Lamelas, V., 10 riesgos de la implantación de IA en la Administración y cómo gestionarlos en No solo AYTOS, https://shorturl.at/J0wug, (short url), 1 de febrero de 2025, [Fecha consulta: 15 de julio de 2025].

Article 19, «España: La acción del gobierno contra la desinformación debería contar con todas las partes interesadas» en https://shorturl.at/ZHLsb (short url). [Fecha consulta: 14 de julio de 2025].

Benjamins, R. «Hacia una IA sostenible: una perspectiva 360 incluyendo negocio, sociedad, ética y cambio climático» en W. Arellano Toledo, W. (dir.), *Derecho, ética e inteligencia artificial*, Tirant lo Blanch, Valencia, 2023.

Bonales, G. et al. «Chatbot como herramienta comunicativa durante la crisis sanitaria COVID-19 en España». *ComHumanitas. Revista Científica de Comunicación*, vol. 11, núm. 3, 2020.

Cerrato, I., González Alarcón, N. «¿Ya tenemos suficientes apps?, Blog Abierto al público, 1 de septiembre de 2020. En: https://shorturl.at/sQFTW (short url) [Consultado el 18 de julio de 2025].

Consejo de Europa, «La IA y el control del coronavirus Covid-19» en https://rb.gy/29in9w (short url). [Fecha consulta: 14 de julio de 2025].

Cotino Hueso, L. «Inteligencia artificial, big data y aplicaciones contra la COVID-19: privacidad y protección de datos». *IDP. Internet, Derecho y Política*, núm. 31, 2020.

Gamero Casado, E. «Sistemas automatizados de toma de decisiones en el Derecho Administrativo Español». *Revista General de Derecho Administrativo*, núm. 63, 2023.

Guersenzvaig, A., Casacuberta, D. «Los peligros del algoritmo en tiempos del coronavirus», en https://elpais.com/tecnologia/2020-03-30/los-peligros-del-algoritmo-en-tiempos-del-coronavirus.html [Fecha consulta: 14 de julio de 2025].

MÁRQUEZ CARRASCO, C.; ORTEGA RAMÍREZ, J. A. «La COVID-19 y los desafíos de la vigilancia digital para los derechos humanos: a propósito de la app DataCOVID prevista en la Orden Ministerial SND/29/2020, de 27 de marzo». *Revista Bioética y Derecho*, núm. 50, 2020.

MARTÍNEZ GARCÍA, E., «Retos de la función jurisdiccional para un mundo interdependiente y ecodependiente». *Teoría & Derecho*, núm. 37 (Derecho e inteligencia artificial), 2024.

PAVÓN DE PAZ, I., et. al. La e-consulta como herramienta para la relación entre Atención Primaria y Endocrinología. Impacto de la epidemia por COVID-19 en su uso. *Journal of Healthcare Quality Research*, Vol. 37, núm. 3 (mayo-junio), 2022. DOI: 10.1016/j.jhqr.2021.10.006.

PONCE SOLÉ, J. «Inteligencia artificial, Derecho administrativo y reserva de humanidad: algoritmos y procedimiento administrativo debido tecnológico». *Revista General de Derecho Administrativo*, núm. 50, 2019.

PORTA FRUTOS, C. «Inteligencia artificial y proceso civil. Uso de los sistemas de IA en la adopción de medidas cautelares: riesgos, beneficios y perspectivas de futuro». *Revista Aranzadi de Derecho y nuevas tecnologías*, núm. 67 (enero-abril), 2025.

TEJADURA TEJADA, J., «El Estado de Derecho frente al COVID reserva de ley y derechos fundamentales». *Revista Vasca de Administración Pública (RVAP). Administrazio Publikoaren Euskal Aldizkaria*, núm. 120 (Mayo-agosto), 2021 (Ejemplar dedicado a: En recuerdo de Pablo Pérez Tremps).

TOACHE, E.A., ROSALES, M.A. «Preocupaciones éticas en el uso de inteligencia artificial, transparencia y derecho de acceso a la información. El caso de los chatbots en el gobierno de México, en el contexto de la Covid-19». *Estudios en Derecho a la Información*, 2023, pp. 85-111 DOI: 10.22201/iij.25940082e.2023.15.17472

VESTRI, G. «El acceso a la información algorítmica a partir del caso bono social vs. Fundación ciudadana Civio». *Revista General de Derecho Administrativo*, núm. 61, octubre-2022.

VESTRI, G., «Inteligencia artificial y chatbots en el proceso de transición digital del sector público» en M.D. Cervilla Garzón, M.A. Blandino Garrido (dirs.), y A. Nieto Cruz (coord.), *Declaración de voluntad en un entorno virtual*, Thomson Reuters-Aranzadi, Pamplona, España, 2021.

VESTRI, G., «La inteligencia artificial ante al desafío de la transparencia algorítmica. Una aproximación desde la perspectiva jurídico-administrativa». *Revista Aragonesa de Administración Pública*, núm. 56, 2021.

VESTRI, G. «Catarsis o plasticidad jurídica: el Derecho público contra las cuerdas», en A. Del Campo (Compilador), *Pensar la pandemia. Más allá de la sanidad y la economía*, Dykinson, Madrid, 2021.

WYNANTS, L. et. al. «Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal», *BMJ (Clinical research ed.),* 369, m1328. https://doi.org/10.1136/bmj.m1328.

DECISIONES AUTOMATIZADAS Y DERECHO A EXPLICACIÓN: SUPERVISIÓN Y TRANSPARENCIA EN LA ADMINISTRACIÓN PÚBLICA[1]

Jorge Castellanos Claramunt
Profesor Titular de Derecho constitucional
Universitat de València

SUMARIO: I. INTRODUCCIÓN. II. ATERRIZAJE NORMATIVO. 1. Reglamento General de Protección de Datos y decisiones automatizadas. 2. Reglamento (UE) 2024/1689 – Ley de Inteligencia Artificial (AI ACT). III. SUPERVISIÓN DE LAS DECISIONES AUTOMATIZADAS EN LA ADMINISTRACIÓN PÚBLICA. IV. EXPLICABILIDAD Y TRANSPARENCIA COMO PAUTA RELATIVA AL DERECHO A COMPRENDER LAS DECISIONES AUTOMATIZADAS. V. DIMENSIÓN PRÁCTICA DE LA SUPERVISIÓN Y EXPLICABILIDAD. VI. CONCLUSIONES. BIBLIOGRAFÍA.

I. INTRODUCCIÓN

La acelerada transformación digital que atraviesan las sociedades contemporáneas afecta de manera directa a la materia de protección de datos personales y a la regulación de las tecnologías emergentes. Obviamente afecta a muchísimos escenarios más, pero en lo concerniente a la temática a desarrollar, nos centraremos en esos aspectos que no son baladíes.

En el contexto europeo, la aprobación de instrumentos normativos como el Reglamento General de Protección de Datos (en adelante, RGPD) y el Reglamen-

1. Este trabajo se ha realizado en el marco del proyecto *Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas* [2023-2026] (PID2022-136439OB-I00), financiado por MCIN/AEI/10.13039/501100011033/ y «FEDER Una manera de hacer Europa». Asimismo, esta investigación se inserta en la investigación adscrita el Grupo de investigación PROMETEO Generalitat Valenciana «Algorithmic Law» (Prometeo/2024/016, 2025-2029).

to Europeo de Inteligencia Artificial (en adelante, AI Act) ha configurado un marco regulatorio complejo y dinámico en el que confluyen intereses públicos, empresariales y de los propios titulares de los datos. Lo cierto es que se trata de una apuesta extremadamente ambiciosa teniendo en cuenta la escasa relevancia internacional de la Unión Europea en estos temas, por lo que se ha apostado por una manera diferente de enfocar la IA y, en general, el desarrollo tecnológico, siendo Europa una especie de aldea gala que se resiste a la invasión tecnológica defendiendo los derechos fundamentales de los ciudadanos. Por supuesto hay que estudiar en profundidad lo dispuesto en estas normativas, como no puede ser de otra manera, pero también desde una perspectiva realista y ajustando las posibilidades de incidencia real en el escenario global, que son pocas.

Por su parte, España, como Estado miembro de la Unión Europea, se encuentra plenamente integrada en este ecosistema, lo que implica no solo la aplicación directa de la normativa comunitaria, sino también la adaptación de su legislación interna, como la Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales (en adelante, LOPDGDD). Asimismo, organismos como la Agencia Española de Protección de Datos (en adelante, AEPD) y el Supervisor Europeo de Protección de Datos (en adelante, SEPD) desempeñan un papel clave en la supervisión y aplicación de estas normas, reforzando la protección de los derechos fundamentales frente a los riesgos derivados de la digitalización[2].

A ello cabe agregar que el uso cada vez más generalizado de sistemas automatizados de decisión en el sector público implica que Administraciones públicas de distintos niveles recurran a algoritmos y herramientas de inteligencia artificial para tomar decisiones que afectan a los ciudadanos[3]. Todo ello se barniza desde una perspectiva de mejora en la eficiencia en la toma de decisiones, pero lo cierto es que también genera riesgos para los derechos fundamentales de las personas (como el derecho a la privacidad, la no discriminación o la tutela judicial efectiva). Por ello, el ordenamiento jurídico de la Unión Europea ha desarrollado un marco normativo para garantizar la supervisión adecuada de estos sistemas y la explicabilidad de sus decisiones. He aquí el objeto de este trabajo, analizar estos sistemas desde una perspectiva normativa europea. En este sentido, hay que centrar el tiro y advertir de qué se trata cuando hablamos de supervisión y explicabilidad. Así, con la supervisión se alude tanto al control humano y administrativo sobre los algoritmos (incluyendo la intervención de autoridades independientes) como a la existencia de mecanismos de control *ex*

2. L. Cervera Navas. «Las instituciones y organismos europeos de protección de datos: el Supervisor Europeo y el Comité Europeo de Protección de Datos», *El Cronista del Estado Social y Democrático de Derecho*, núm. 88-89 (Mayo-Junio), 2020 (Ejemplar dedicado a: Protección de datos: antes, durante y después del coronavirus), págs. 104-115.

3. A.E. Castagnedi Ramírez. «Inteligencia artificial: Cuando los algoritmos se convierten en neuronas», *Ius et scientia: Revista electrónica de Derecho y Ciencia*, vol. 8, núm. 2, 2022 (Ejemplar dedicado a: Medicina, biotecnología y derecho), págs. 136-145.

ante y *ex post*[4]. Por su parte, la explicabilidad hace referencia a la obligación de ofrecer explicaciones claras y comprensibles sobre cómo una decisión automatizada ha sido adoptada, de modo que los ciudadanos afectados puedan comprender las razones y fundamentos de esa decisión[5]. De ahí que la explicabilidad sea un elemento para tener en cuenta a la hora de poder impugnar o cuestionar una decisión automatizada de forma efectiva ya que, sin una explicación suficiente de la lógica de un algoritmo, los derechos de la persona a presentar alegaciones o recurrir carecerían de efectividad real[6]. Por tanto, supervisión y explicabilidad son pilares complementarios para garantizar que la inteligencia artificial en la Administración se alinee con los valores democráticos y el Estado de Derecho.

II. ATERRIZAJE NORMATIVO

1. Reglamento General de Protección de Datos y decisiones automatizadas

La arquitectura regulatoria en materia de protección de datos y tecnologías emergentes en Europa se apoya en un sistema multinivel en el que coexisten normas de aplicación directa —como los reglamentos de la Unión Europea— con transposiciones y desarrollos legislativos nacionales. Este entramado persigue un doble objetivo, por un lado, armonizar estándares de protección en el mercado interior y, por otro lado, preservar la soberanía normativa de los Estados miembros en aquellas materias que no están plenamente integradas.

El RGPD (Reglamento (UE) 2016/679) establece la primera capa normativa relevante para las decisiones automatizadas en cualquier ámbito, incluido el público. Y en lo que a esta investigación respecta, pese a que se constituya como un marco general de protección de datos personales, el RGPD contiene disposiciones específicas dirigidas a regular la toma de decisiones individuales automatizadas basadas en datos personales. Además, su carácter de norma directamente aplicable garantiza la uniformidad en los principios de tratamiento —licitud, lealtad, transparencia, minimización, exactitud, limitación de la conservación,

4. L. Cotino Hueso. «Transparencia y explicabilidad de la inteligencia artificial y «compañía» (comunicación, interpretabilidad, inteligibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad…). Para qué, para quién y cuánta», en L. Cotino Hueso y J. Castellanos Claramunt, *Transparencia y explicabilidad de la inteligencia artificial*, Tirant lo Blanch, Valencia, 2022, págs. 29-70.

5. L. Cotino Hueso y J. Castellanos Claramunt, *Transparencia y explicabilidad de la inteligencia artificial*, Tirant lo Blanch, Valencia, 2022.

6. G. Vestri. «De la inteligencia artificial y otros factores: requisitos de transparencia y explicabilidad en la contratación pública», C*ontratación administrativa práctica: revista de la contratación administrativa y de los contratistas*, núm. Extra 1, 2025 (Ejemplar dedicado a: El impacto de la IA en la contratación pública).

integridad y confidencialidad[7]— y en los derechos de las personas interesadas. Por su parte, la figura de las autoridades de control independientes, como la AEPD en España, asegura la vigilancia y cumplimiento del marco normativo, incluyendo la imposición de sanciones por infracciones graves y muy graves.

El RGPD tiene una aspiración de alcance extraterritorial, aplicándose a tratamientos realizados fuera de la UE cuando afectan a interesados en su territorio[8]. Además, los casos analizados por la AEPD, como la imposición de multas por el uso obligatorio de datos biométricos sin base legal suficiente[9] o la solicitud indebida de copias de documentos de identidad en el sector turístico[10], lo que tratan de poner sobre la mesa es su pretendido alcance transversal, de manera que sus efectos irradien a muchos y diversos escenarios de índole jurídica.

En particular, el artículo 22 RGPD consagra el derecho de todo interesado a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, cuando dicha decisión produzca efectos jurídicos o afecte significativamente al interesado[11]. Este precepto supone, de entrada, una prohibición general de las decisiones completamente automatizadas con efectos graves sobre la persona, salvo que concurra alguna de las excepciones previstas en el propio artículo 22(2)[12]. Además, aun en esos casos excepcionales, el artículo 22(3) exige al menos que el ciudadano tenga derecho a la intervención humana, a expresar su punto de vista y a impugnar la decisión[13].

La normativa de protección de datos, por tanto, condiciona fuertemente el uso de algoritmos decisorios en el sector público puesto que una administración no puede tomar decisiones con efectos relevantes apoyándose únicamente en un sistema automatizado, a menos que una norma lo autorice y se respeten garantías adicionales. A este respecto el TJUE ha aclarado el alcance del concepto de «decisión basada únicamente en tratamiento automatizado» con la sentencia C 634/21 (*OQ contra Land Hessen*, 7 de diciembre de 2023), en la que el Tribunal examinó el caso de un sistema privado de *credit scoring* (puntaje crediticio) cuyos resulta-

7. J. Muñoz Rodríguez. «Principios de protección de datos: licitud, lealtad, transparencia, minimización, exactitud, integridad y confidencialidad», *Economist & Jurist*, vol. 26, núm. 217, 2018, págs. 18-23.

8. E. Díaz Díaz, M.S.M. Del Busto Calosi. «De España a Iberoamérica: La influencia global del RGPD en las Leyes de Protección de Datos en Iberoamérica. Enfoque en Perú», *Revista Aranzadi Doctrinal*, núm. 7, 2025.

9. https://www.aepd.es/informes-y-resoluciones/criterios-juridicos-aepd/aepd-sanciona-tratamiento-datos-biometricos-ia

10. https://www.aepd.es/prensa-y-comunicacion/notas-de-prensa/aepd-informa-de-que-no-esta-permitido-solicitar-copia-dni-o-pasaporte-en-hospedajes

11. D. Sancho Villa. «Las decisiones individuales automatizadas, incluida la elaboración de perfiles (Comentario al artículo 22 RGPD)», en A. Troncoso Reigada (dir.), *Comentario al Reglamento General de Protección de Datos y a la Ley Orgánica de Protección de Datos personales y Garantía de los Derechos Digitales*, vol. 1, Thomson Reuters Aranzadi, Cizur Menor (Navarra), 2021, págs. 1725-1745.

12. Por ejemplo, que la decisión esté autorizada por una ley aplicable, que sea necesaria para un contrato, o que el ciudadano haya dado su consentimiento explícito.

13. V. Ivone. «Artículo 22 del RGPD y tratamiento automatizado de datos», en M.A. López-Suárez (dir.), *Big data y protección de datos*, Tirant lo Blanch, Valencia, 2025, págs. 129-178.

dos eran utilizados por bancos para conceder o denegar créditos. El TJUE declaró que la generación automatizada de un valor de solvencia (*score*) por una agencia de crédito constituye una «decisión individual automatizada» en el sentido del art. 22(1) RGPD cuando dicho *score* determina de manera decisiva que un tercero celebre o rescinda un contrato con la persona. Es decir, aunque formalmente la decisión final la tome un humano, si en la práctica el algoritmo tiene un peso determinante en esa decisión, la situación queda bajo el ámbito del artículo 22 RGPD. Con ello, el TJUE impidió eludir la protección del artículo 22 mediante la simple intervención nominal de un humano ya que, si la decisión humana se funda esencialmente en la recomendación automatizada, debe considerarse una decisión automatizada a efectos jurídicos. La consecuencia es que tal práctica estaría prohibida por principio, salvo que se cumpla alguna excepción del artículo 22(2) y se apliquen las salvaguardias del 22(3).

Junto al artículo 22, el RGPD impone obligaciones de transparencia específicas respecto a los sistemas de decisión automatizada[14]. Los artículos 13 y 14 RGPD[15] disponen que, si se prevé realizar decisiones automatizadas con esos datos, se debe informar de tal hecho y proporcionar «información significativa sobre la lógica aplicada, así como la importancia y las consecuencias previstas» de ese procesamiento automatizado. Asimismo, el artículo 15(1)(h) RGPD establece el derecho de acceso del interesado a obtener información similar, que es la relativa a la confirmación de si sus datos se usan en decisiones automatizadas y, en tal caso, «información significativa sobre la lógica aplicada» y las consecuencias previstas. Sobre ello cabe comentar que estas disposiciones fueron tradicionalmente fuente de debate sobre si consagraban o no un auténtico «derecho a explicación». En cualquier caso, la cuestión ha quedado sustancialmente esclarecida por la jurisprudencia reciente, de modo que encontramos la C 203/22[16], en la que el TJUE afirmó que la finalidad principal del derecho del artículo 15(1)(h) RGPD es permitir al afectado ejercer efectivamente sus derechos del art. 22(3) (a expresar su punto de vista e impugnar la decisión)[17]. Por tanto,

14. F.J. Blázquez Ruiz. «La paradoja de la transparencia en la IA: Opacidad y explicabilidad. Atribución de responsabilidad», *Revista Internacional de Pensamiento Político*, núm. 17 (1), 2022, págs. 261–272.

15. Relativos a la información que debe darse al interesado cuando se recopilan datos personales.

16. *CK contra D&B*, sentencia de 26 de enero de 2024.

17. El caso surge a raíz de la negativa de la empresa *Dun & Bradstreet Austria GmbH*, especializada en evaluaciones crediticias, a proporcionar a un ciudadano información detallada sobre la lógica aplicada en la elaboración de su perfil de solvencia. La controversia se centra en determinar hasta qué punto una persona afectada por una decisión automatizada tiene derecho a conocer los criterios y procedimientos empleados en su evaluación y cómo este derecho interactúa con otras consideraciones jurídicas, como la protección de secretos comerciales regulada en la Directiva 2016/943.

El Tribunal examina si el derecho de acceso reconocido en el RGPD exige que la empresa proporcione únicamente una explicación general del proceso de evaluación crediticia o si, por el contrario, debe detallar los criterios específicos utilizados para el cálculo del perfil del ciudadano afectado. Se plantea además la relación entre la transparencia en el tratamiento de datos personales y la protección de secretos comerciales, ya que la empresa argumentaba que revelar información

el interesado debe recibir una explicación de la decisión automatizada antes de tener que recurrirla, pues de otro modo sus derechos de defensa serían ilusorios. En línea con el considerando 71 RGPD, el Tribunal reconoció expresamente que el RGPD ofrece al interesado un genuino derecho a una explicación sobre el funcionamiento del mecanismo automatizado que produjo la decisión y sobre el resultado alcanzado. De este modo, la «información significativa sobre la lógica aplicada» debe interpretarse como un derecho a obtener una explicación comprensible del procedimiento y los principios en que se basó la decisión automatizada que le concierne.

La sentencia C 203/22 detalla qué exige esa explicación para ser adecuada. No basta con revelar fórmulas matemáticas complejas ni con describir en extremo detalle cada fase del algoritmo, pues tales enfoques no serían ni concisos ni inteligibles para el ciudadano medio. En lugar de ello, la explicación debe presentarse de forma concisa, transparente e inteligible, indicando qué datos del interesado se utilizaron y cómo se emplearon en el proceso automatizado para llegar a ese resultado. En consecuencia, el responsable del tratamiento tiene el deber de hallar formas sencillas de informar al interesado sobre la lógica o los criterios esenciales utilizados por el sistema. Por ejemplo, podría explicarse qué variables personales influyeron y en qué medida, o cómo un cambio en ciertos datos del interesado alteraría el resultado. Lo importante es que el afectado pueda entender las razones por las que, mediante el tratamiento automatizado de sus datos, se le asignó cierta conclusión, como ocurre con un perfil de riesgo crediticio.

Debe resaltarse que ni siquiera la invocación de secretos empresariales o derechos de propiedad intelectual por parte del responsable exime del deber de transparencia[18]. El TJUE, consciente de la tensión entre explicabilidad y secreto industrial, estableció un equilibrio en la C 203/22 de manera que, si el responsable

---

más detallada podría exponer su metodología interna y comprometer su ventaja competitiva. En este contexto, el Tribunal también evalúa la aplicación de la normativa austriaca, que limita el derecho de acceso cuando se afectan secretos comerciales, y su compatibilidad con el RGPD y el principio de proporcionalidad en la protección de datos.

En su fallo, el Tribunal concluye que el derecho de acceso previsto en el artículo 15.1.h del RGPD debe interpretarse de manera que garantice a la persona afectada la posibilidad de comprender cómo se ha tomado la decisión automatizada. No basta con informar de la existencia de un sistema de perfilado, sino que el interesado debe recibir una explicación clara, inteligible y transparente sobre los criterios específicos utilizados y cómo estos han afectado a su caso particular. Sin embargo, el Tribunal reconoce que este acceso no es absoluto y que, en situaciones donde exista un conflicto con otros derechos, como la protección de secretos comerciales o datos de terceros, la autoridad de control o el tribunal competente deben evaluar si la restricción del acceso está justificada y en qué medida se puede limitar la información proporcionada.

18. En el ámbito español, la LOPDGDD adapta el RGPD al ordenamiento nacional, incorporando particularidades como el derecho al olvido digital *post mortem*, la regulación de sistemas de información crediticia o la limitación de determinadas prácticas de videovigilancia laboral. Asimismo, contempla garantías específicas para el uso de datos en el ámbito electoral y en las relaciones laborales, como quedó patente en resoluciones de la AEPD sobre la inclusión no consentida de empleados en grupos de mensajería corporativa (Expediente Nº: EXP202404627, disponible: https://www.aepd.es/documento/ps-00393-2024.pdf).

alega que dar información sobre la lógica algorítmica revelaría secretos comerciales o datos de terceros, debe facilitar esa información confidencial a la autoridad de control (agencia de protección de datos) o al juez, para que dicha autoridad realice una ponderación independiente de los intereses en juego. Es decir, no cabe negar al interesado toda información amparándose en el secreto comercial; la información clave deberá al menos ser examinada por una autoridad imparcial, que decidirá qué se puede revelar al ciudadano sin comprometer secretos legítimos.

2. Reglamento (UE) 2024/1689 – Ley de Inteligencia Artificial (AI ACT)

Como complemento sectorial y tecnológico al RGPD, la Unión Europea promulgó el Reglamento (UE) 2024/1689 (conocido como *AI Act* o Ley de IA), aprobado el 13 de junio de 2024, que establece un marco jurídico uniforme para la comercialización, despliegue y uso de sistemas de inteligencia artificial en la UE[19]. A diferencia del RGPD, que se centra en la protección de datos personales, el *AI Act* adopta un enfoque basado en el riesgo del sistema de IA en sí mismo, abarcando incluso supuestos en que no medie tratamiento de datos personales. Su objetivo es garantizar que los sistemas de IA sean seguros, fiables y respetuosos con los derechos fundamentales, imponiendo requisitos a los proveedores (desarrolladores) y a los responsables del despliegue (usuarios finales de los sistemas, entre ellos las administraciones públicas) [20].

El *AI Act* clasifica las aplicaciones de IA por niveles de riesgo. De especial importancia son los sistemas de IA de alto riesgo, definidos en el artículo 6 y listados en el Anexo III del Reglamento. Muchas de las categorías de alto riesgo se refieren directamente a usos en el sector público o sectores de interés público. Por ejemplo, se consideran de alto riesgo los sistemas de IA destinados a ser utilizados por autoridades públicas para gestionar el acceso de personas a servicios y prestaciones esenciales (educación, formación profesional, servicios financieros básicos, asistencia sanitaria, ayudas sociales, etc.); los sistemas usados para la evaluación crediticia o solvencia de personas físicas; las herramientas de IA empleadas en procesos de contratación de personal; los sistemas utilizados por fuerzas del orden para evaluar riesgos de seguridad o elaborar perfiles criminales; los utilizados en la administración de justicia o procedimientos administrativos con efectos jurídicos; así como los sistemas aplicados en el ámbito migratorio y de control de fronteras (por ejemplo, para evaluar solicitudes de asilo o detectar riesgo de migración irregular). El propio Reglamento reconoce que en estos contextos públicos es «sumamente importante» que los

19. A. de Marcos Fernández. «Una doble historia de la inteligencia artificial: avance tecnológico y proceso de regulación en Europa», *Revista de privacidad y derecho digital*, vol. 9, Nnúm. 34, 2024, págs. 26-89.

20. J. Castellanos Claramunt. *DemocracIA: un análisis en clave constitucional*, Dykinson, Madrid, 2025.

sistemas de IA sean precisos, no discriminatorios y transparentes, pues de ellos dependen derechos fundamentales de las personas.

Para los sistemas de IA de alto riesgo, el *AI Act* impone una serie de requisitos obligatorios antes y durante su puesta en servicio. En términos de explicabilidad y supervisión, destacan los siguientes:

— *Documentación técnica, datos y registros*. El proveedor de un sistema de IA de alto riesgo debe elaborar documentación exhaustiva que describa sus características, finalidad, diseño y funcionamiento, incluyendo las técnicas algorítmicas utilizadas. Además, el sistema debe estar diseñado para registrar automáticamente eventos (*logs*) durante su funcionamiento, permitiendo conservar historiales de uso que faciliten auditorías y análisis a posteriori. En otras palabras, la IA debe generar trazas que posibiliten reconstruir qué ocurrió en cada decisión y los proveedores y usuarios deben conservar estos registros durante un período adecuado (al menos seis meses u otro plazo que establezca la normativa sectorial).
— *Transparencia y explicaciones a usuarios y personas afectadas*. Los sistemas de alto riesgo deben ir acompañados de instrucciones de uso claras para sus usuarios (por ejemplo, funcionarios que los apliquen) y de información que permita interpretar correctamente sus resultados[21]. En este sentido, el *AI Act* consagra expresamente un derecho de las personas afectadas a obtener una explicación de las decisiones individuales adoptadas con apoyo de IA[22]. El artículo 86 del Reglamento, titulado «Derecho a explicación de decisiones tomadas individualmente», establece que toda persona afectada por una decisión de un responsable público basada principalmente en la salida de un sistema de IA de alto riesgo —que produzca efectos jurídicos o le afecte considerablemente— tiene derecho a obtener del responsable explicaciones claras y significativas sobre el papel que tuvo la IA en el proceso de decisión y sobre los principales elementos de la decisión final. Esta disposición cubre precisamente el supuesto típico de las decisiones automatizadas en el sector público, imponiendo la obligación de explicar cómo influyó el algoritmo en el resultado. Se aplica a sistemas de IA de alto riesgo enumerados en el Anexo III (salvo los de la categoría 2, que corresponden a sistemas de identificación biométrica remota en espacios públicos, los cuales de hecho están prohibidos salvo excepciones estrictas)[23]. El derecho a explicación

21. S. Alayón Miranda. «El problema de la interpretabilidad de la Inteligencia Artificial y su impacto en la Administración Pública», *Revista Canaria de Administración Pública*, núm. 3, 2024, págs. 175-202.

22. P. Vargas Martínez. «La evolución de los conceptos de transparencia y explicabilidad: de los años 90 al Reglamento de Inteligencia Artificial», en I. Sánchez Frías e Y. Villegas Almagro (dirs.), *Derecho y entornos digitales*, Atelier, Barcelona, 2025, págs. 49-70.

23. M.R Torres Carlos y L. Míguez Macho. «Sistemas de IA prohibidos y sistemas de IA de alto riesgo», en M. Barrio Andrés (dir.), *El Reglamento Europeo de Inteligencia Artificial*, Tirant lo Blanch, Valencia, 2024, págs. 48-86.

del artículo 86 es subsidiario respecto a otras normas – es decir, solo se aplica en la medida en que el Derecho de la Unión no contemple ya un derecho similar. Esto busca armonizarlo con el RGPD; en la práctica, supone un refuerzo del derecho a explicación en contextos de IA, complementando el RGPD (que, como vimos *supra*, ya había sido interpretado en ese sentido por el TJUE). El *AI Act* elimina así cualquier duda de manera que, si un ciudadano se ve perjudicado por la decisión automatizada de una Administración, tiene el derecho explícito a solicitar y recibir una explicación comprensible de la contribución de la IA a dicha decisión, explicaciones que deben servirle de base para ejercer sus derechos y eventuales recursos.

— *Supervisión humana obligatoria*. El Reglamento de IA enfatiza la necesidad de control humano efectivo sobre los sistemas de alto riesgo. Los proveedores deben diseñar la IA de forma que pueda ser vigilada y controlada por personas durante su uso. Y los responsables del despliegue (usuarios finales, como las autoridades públicas) tienen la obligación de establecer medidas de supervisión humana apropiadas antes de utilizar el sistema, lo que incluye garantizar que el personal encargado de supervisar la IA esté debidamente formado y facultado para intervenir, corregir o desactivar el sistema cuando sea necesario. El objetivo de ello es que los humanos no actúen como meros «espectadores» de la decisión automatizada, sino que comprendan sus recomendaciones y tengan capacidad real de tomar decisiones informadas sobre si aceptarlas o modificarlas. Además, en ciertos usos particularmente sensibles, el *AI Act* impone supervisión humana reforzada, como por ejemplo en el caso de sistemas de identificación biométrica remota en tiempo real[24], donde se exige que al menos dos personas físicas confirmen la correspondencia antes de que la autoridad pueda actuar o tomar decisiones basadas en la identificación generada por la IA.

— *Evaluaciones de impacto y control ex ante*. De especial relevancia para el sector público, el *AI Act* introduce la figura de la evaluación de impacto relativa a los derechos fundamentales (FRIA). Conforme al artículo 27, antes de desplegar un sistema de IA de alto riesgo, los responsables del despliegue que sean organismos de Derecho público (así como entidades privadas que ofrezcan servicios públicos esenciales) deberán realizar una evaluación de impacto sobre cómo el uso de ese sistema puede afectar a los derechos fundamentales. Esta obligación cubre, entre otros, los sistemas de IA utilizados en la gestión de servicios sanitarios, seguridad social, justicia, policía, migración y otros ámbitos públicos especialmente relevantes. La evaluación de impacto debe identificar riesgos específicos de lesión de derechos o discriminación, determinar las categorías de personas afectadas, la duración y contexto del uso del sistema, las medidas de

24. Que sería en el caso de uso de cámaras con IA para reconocimiento facial en espacios públicos.

supervisión humana planeadas, y las estrategias para mitigar los riesgos o atender reclamaciones en caso de problemas[25]. Una vez completada la evaluación, sus resultados han de ser notificados a la autoridad nacional competente de vigilancia del mercado (la autoridad supervisora del *AI Act*), de manera que exista un control administrativo previo del despliegue de esos algoritmos en lo público. Esta evaluación de impacto en derechos fundamentales se concibe como complementaria a la evaluación de impacto en protección de datos exigida por el RGPD (DPIA del art. 35 RGPD), cuando la aplicación involucra datos personales. Es decir, en contextos donde ambos Reglamentos aplican, la administración deberá hacer un análisis integrado que cubra tanto los riesgos para la privacidad como otros impactos (no discriminación, debido proceso, etc.), reforzando así la coherencia normativa.

— *Registro y base de datos europea*. Para aportar transparencia pública, el *AI Act* establece la creación de una base de datos de la UE en la cual se registrarán los sistemas de IA de alto riesgo. Los proveedores deberán inscribir sus sistemas una vez obtengan el marcado CE o la autorización pertinente, proporcionando información básica sobre su finalidad, cumplimiento y resultados de evaluación de conformidad. Y los órganos públicos que usen sistemas de IA de alto riesgo estarán también sujetos a esta obligación de registro cuando actúen como responsables del despliegue. Hay que comentar que parte de esta base de datos será pública y accesible fácilmente, lo que permitirá que ciudadanos, investigadores o empresas sepan qué sistemas de IA de alto riesgo están activos, quién los utiliza y con qué fin. Sin embargo, por motivos de seguridad, ciertas aplicaciones, como las policiales o fronterizas, se registrarán en secciones no públicas a las que solo acceden la Comisión y las autoridades de supervisión. Con todo, se avanza hacia una transparencia institucional ya que un ciudadano podrá comprobar si su gobierno tiene desplegado un algoritmo para asignación de becas, o un periodista podrá escrutar qué IA se utilizan en juzgados o ayuntamientos.

Haciendo un análisis global de lo expuesto, podemos colegir que el *AI Act* añade una capa de regulación *ex ante* y de gobernanza institucional que complementa al RGPD[26]. De esta manera, mientras el RGPD ofrece remedios individuales (derechos de los interesados, reclamaciones ante Autoridades de Protección de Datos, etc.), la Ley de IA impone controles estructurales como la certificación de conformidad de sistemas, requisitos técnicos de diseño (seguri-

25. La Administración debe reflexionar y documentar antes de usar la IA sobre: ¿qué puede salir mal?, ¿a quién podría perjudicar?, ¿cómo se controlará?, ¿qué haremos si ocurre un resultado injusto?

26. C. Berenguer Albaladejo. «Transparencia y explicabilidad para prevenir la discriminación de los sistemas de inteligencia artificial: La interacción entre el RGPD y el RIA», en J.A. Moreno Martínez y P.J. Femenía López (coords.), *Inteligencia artificial y derecho de daños: Cuestiones actuales: Acorde al Reglamento (UE) 2024/1689*, Dykinson, Madrid, 2024, págs. 49-118.

dad, ausencia de sesgos, explicabilidad), evaluación de impactos y registro. En el sector público europeo, esto implica que las Administraciones tendrán que ajustar sus procedimientos. Así, si un Ministerio planea implementar un algoritmo para detectar fraude en prestaciones, deberá no solo cumplir la base legal del RGPD, sino también evaluar preventivamente los riesgos para derechos (evitar sesgos contra colectivos vulnerables), formar a sus funcionarios en la supervisión del sistema, registrar dicho algoritmo en la base europea, y estar listo para explicar a cada ciudadano afectado cómo influyó la IA en cualquier decisión adversa.

III. SUPERVISIÓN DE LAS DECISIONES AUTOMATIZADAS EN LA ADMINISTRACIÓN PÚBLICA

La supervisión de las decisiones automatizadas en el ámbito público adopta varias dimensiones complementarias: por un lado, la supervisión humana individual en el propio proceso de toma de decisión (evitando una automatización absoluta y sin control); por otro, la supervisión institucional o regulatoria por parte de organismos independientes que vigilan el cumplimiento normativo (agencias de protección de datos, autoridades de IA, tribunales); y finalmente la supervisión democrática en un sentido amplio (transparencia hacia la sociedad civil, rendición de cuentas política). Todas ellas persiguen impedir que los algoritmos operen como *cajas negras incontroladas*, garantizando que exista responsabilidad humana última sobre cada decisión pública[27].

Un principio cardinal derivado tanto del RGPD como del AI Act es que los sistemas automatizados deben estar al servicio de las personas y bajo su control, no al revés. En la práctica, esto significa que las administraciones que empleen IA han de diseñar sus procedimientos de manera que siempre intervenga un ser humano con capacidad real de decisión antes de adoptar actos individuales que afecten a ciudadanos. La figura del *humano en el bucle* (*human-in-the-loop*) resulta, por ende, clave para prevenir que una recomendación algorítmica sesgada o errónea se traduzca directamente en un acto administrativo perjudicial.

El RGPD fomenta esa intervención humana al prohibir, como regla general, las decisiones «únicamente automatizadas» con impacto significativo, excepto si el interesado dio consentimiento, es necesario para un contrato, o una ley lo autoriza con garantías. En el contexto público, no es frecuente recabar consentimiento válido para decisiones administrativas, de modo que la excepción habitual sería la existencia de una ley nacional o europea que habilite la decisión automatizada (art. 22(2)(b) RGPD). Ahora bien, esa ley debe incluir medidas para salvaguardar los derechos del interesado, como mínimo el derecho a ob-

27. J.I. Herce Maza. «Buena administración de la transparencia de los algoritmos de la Administración Pública: un instrumento para el control de las cajas negras decisionales», en G. Vestri (dir.), *La disrupción tecnológica en la Administración Pública: retos y desafíos de la inteligencia artificial*, Thomson Reuters Aranzadi, Cizur Menor (Navarra), 2022, 123-138.

tener intervención humana posterior, a expresar su opinión y a impugnar la decisión (art. 22(3)). En otras palabras, incluso cuando un Estado miembro decida permitir por ley cierta decisión automatizada en la administración (por ejemplo, la concesión automática de un permiso si se cumplen condiciones objetivas), deberá asegurar que la persona pueda solicitar que un funcionario revise el resultado o lo reconsidere, lo que implica formar adecuadamente a los funcionarios para comprender las sugerencias algorítmicas y alentar una cultura administrativa donde no se delegue ciegamente en «lo que diga el ordenador».

El *AI Act*, por su parte, traduce estos postulados en requisitos técnicos y organizativos. Como vimos, los sistemas de IA de alto riesgo deben incorporar funciones que permitan orientar e informar al operador humano durante su uso. Por ejemplo, un sistema de IA que ayuda a decidir ayudas sociales podría incluir alertas o explicaciones al empleado público sobre la fiabilidad de su recomendación o niveles de incertidumbre, de modo que el humano sepa cuándo debe ser especialmente cauteloso. Además, el Reglamento exige que las personas a cargo de la supervisión humana sean competentes en IA y tengan autoridad suficiente. No se trata solo de agregar un interventor humano, sino de asegurarse de que este tenga la alfabetización digital necesaria para entender cómo funciona el sistema, sus posibles sesgos, sus limitaciones y, en general, cuándo confiar o desconfiar de él. En los casos más sensibles (como el reconocimiento facial en vivo), se exige incluso una doble validación humana independiente antes de actuar, reconociendo que la sola presencia de un humano puede ser insuficiente si existe riesgo de error grave y presión de tiempo.

Desde la óptica del Derecho administrativo, esto enlaza con el principio general de que todo acto administrativo debe ser imputable a una autoridad y motivado. Un algoritmo carece de personalidad jurídica; quien «decide» formalmente es la autoridad pública, por lo que esta no puede escudarse en el software para eludir su responsabilidad. De hecho, países como Francia han legislado que cuando las administraciones usan algoritmos para tomar decisiones, deben informar al ciudadano y darle a conocer las reglas o parámetros de esa decisión automatizada[28]. España, por su parte, en la Ley 40/2015 de Régimen Jurídico del Sector Público, artículo 41, reconoció la validez de actos administrativos automatizados *siempre que* se establezca previamente el procedimiento preciso de producción del acto y se identifique al órgano que supervisará e interpretará el resultado. Esto muestra que, incluso en el plano interno de los Estados, se está imponiendo la idea de un control humano estructurado ya que, pese a reconocer que la automatización puede ahorrar trámites rutinarios, debe haber siempre un funcionario responsable último de que la decisión sea conforme a Derecho.

28. Artículo L311-3-1 del Código de Relaciones entre el Público y la Administración francés: «Sous réserve de l'application du 2° de l'article L. 311-5, une décision individuelle prise sur le fondement d'un traitement algorithmique comporte une mention explicite en informant l'intéressé. Les règles définissant ce traitement ainsi que les principales caractéristiques de sa mise en œuvre sont communiquées par l'administration à l'intéressé s'il en fait la demande.

Les conditions d'application du présent article sont fixées par décret en Conseil d'État».

No obstante, un desafío identificado es el fenómeno del sesgo de automatización derivado del hecho de que los humanos tendemos a confiar excesivamente en las sugerencias de los algoritmos, a veces otorgándoles más credibilidad que a nuestro propio juicio, lo que puede minar la efectividad de la supervisión humana, volviéndola acrítica[29]. Para contrarrestarlo, se sugiere que las administraciones implementen procedimientos de doble comprobación aleatoria de algunas decisiones automatizadas, o rotación de personal supervisor para evitar rutina, y que inculquen una cultura de pensamiento crítico respecto de las salidas de la IA[30]. Asimismo, el *AI Act* obliga a que los sistemas de IA de alto riesgo sean explicables para sus usuarios (no solo para los afectados externos), lo que conlleva que la interfaz y documentación deben ayudar al funcionario a interpretar correctamente los resultados. Un ejemplo sería una herramienta de apoyo a sentencias judiciales, que debe proporcionar al juez no solo una predicción, sino las razones principales que la llevaron a sugerir un fallo, permitiendo al juez evaluar si esas razones son pertinentes jurídica y fácticamente en su caso[31].

Además de la vigilancia interna, la UE ha dispuesto un robusto esquema de supervisión externa de las decisiones automatizadas públicas, a través de autoridades administrativas independientes y de los tribunales, capa fundamental para asegurar la *accountability* cuando fallan los controles iniciales o cuando un ciudadano considera lesionados sus derechos.

Por su parte, en materia de protección de datos, las Autoridades de Control (Agencias de Protección de Datos) de cada Estado miembro juegan un papel clave, puesto que son entes independientes encargados de vigilar y hacer cumplir el RGPD incluso frente a organismos públicos[32]. De esta manera los interesados pueden presentar reclamaciones ante ellas si creen que una decisión automatizada violó sus derechos de datos (por ejemplo, si se tomó sin base legal o sin ofrecer las explicaciones debidas), teniendo para ello potestades de investigación, incluyendo auditar sistemas algorítmicos y requerir información sobre su lógica.

Con la entrada en vigor del AI Act, surgen también nuevas autoridades de supervisión de IA, que cada Estado miembro designará como encargadas de la vigilancia del mercado de sistemas de IA (art. 64 y ss. del AI Act). Estas autoridades velarán por que los sistemas de IA comercializados y usados en su territorio cumplan los requisitos del Reglamento 2024/1689 y, en muchos casos, es previsible que los países asignen esa tarea a organismos ya existentes con experiencia en certificación o supervisión tecnológica.

---

29. J. Castellanos Claramunt. «Reflexiones sobre el sorteo de la Champions: la aplicación acrítica de la tecnología», *The Conversation*, 15 de diciembre de 2021.

30. J. Castellanos Claramunt. *DemocracIA: un análisis en clave constitucional*, op. cit.

31. J. Nieva Fenoll. *Inteligencia artificial y proceso judicial*, Marcial Pons, Madrid, 2018.

32. M. Bellón Yturriaga. «Transparencia y protección de datos en el uso de la inteligencia artificial por la administración pública», *Revista de privacidad y derecho digital*, vol. 10, núm. 36, 2025, págs. 25-72.

Una novedad del AI Act es la creación de la Oficina Europea de IA, que actuará como *hub* de coordinación entre autoridades nacionales y apoyo a la Comisión Europea. Esta Oficina podrá incluso ejercer competencias directas de supervisión en casos especiales, como los referentes a modelos de IA de uso general con impacto sistémico (por ejemplo, grandes modelos tipo GPT que afecten a toda Europa). En tales casos, la Comisión (a través de la Oficina) tendrá poderes para recabar documentación completa de esos desarrolladores y evaluar su conformidad[33].

La supervisión institucional también incluye la transparencia pública y el escrutinio por parte de la sociedad civil. Mecanismos como la base de datos pública de sistemas de IA de alto riesgo permitirán a ONGs, periodistas y ciudadanos monitorear qué IA utilizan las administraciones y con qué finalidades. Ya se ha visto un creciente activismo en este campo: organizaciones pro derechos digitales solicitan información (FOIA) sobre algoritmos públicos y han impulsado litigios estratégicos. Un ejemplo fuera del TJUE es el caso *SyRI* en Países Bajos, donde una coalición de ONG y particulares logró que un tribunal de La Haya en 2020 declarase ilegal un sistema gubernamental de perfilado de riesgo de fraude social por vulnerar la privacidad y el principio de proporcionalidad[34].

Por supuesto, el control judicial último recae en los tribunales nacionales y, en su caso, en el TJUE (para cuestiones de Derecho de la Unión) o el TEDH (si hay vulneración de derechos del Convenio Europeo de Derechos Humanos). Un ciudadano europeo que considere que una decisión automatizada pública ha violado sus derechos puede acudir a la jurisdicción contencioso-administrativa de su país. Allí, el juez podrá requerir a la administración la documentación del algoritmo y comprobar, por ejemplo, si la decisión fue arbitraria o discriminatoria. Aquí el derecho a explicación vuelve a ser crucial: gracias al RGPD y al AI Act, el individuo tiene derecho a obtener información sobre la lógica del algoritmo, lo que le permite presentar un recurso judicial con conocimiento de causa. En caso de negativa, el juez podría ampararse en la doctrina del TJUE para obligar a la administración a revelar los criterios o, si mediara secreto comercial de un proveedor privado, revisar él mismo *in camera* el funcionamiento del sistema.

IV. EXPLICABILIDAD Y TRANSPARENCIA COMO PAUTA RELATIVA AL DERECHO A COMPRENDER LAS DECISIONES AUTOMATIZADAS

El otro gran pilar temático es la explicabilidad de las decisiones automatizadas, estrechamente ligada al principio de transparencia. En una democracia, los ciudadanos tienen derecho a conocer las razones de las decisiones que les

33. J.F. Rodríguez Ayuso. «Artículo 64. Oficina de IA», en M. Barrio Andrés (dir.), *Comentarios al Reglamento Europeo de Inteligencia Artificial*, La Ley, Madrid, 2024, págs. 636-642.

34. G. Lazcoz Moratinos y J.A. Castillo Parrilla. «Valoración algorítmica ante los derechos humanos y el Reglamento General de Protección de Datos: el caso SyRI», *Revista chilena de derecho y tecnología*, vol. 9, núm. 1, 2020, págs. 207-225.

afectan, más aún si estas provienen de sistemas tecnológicos complejos[35]. La UE ha consagrado este derecho a entender las decisiones automatizadas tanto en normas jurídicas vinculantes como en principios éticos y directrices.

Como ya se analizó, el RGPD —vía art. 15(1)(h) y correlativos— proporciona a la persona un derecho de acceso a una explicación sobre la lógica de las decisiones automatizadas que le conciernen. Esta explicación equivale, en lo sustancial, a la motivación de un acto administrativo cuando dicho acto es en realidad fruto de un proceso automatizado. Tradicionalmente, las leyes de procedimiento administrativo exigen que los actos negativos o restrictivos de derechos estén motivados, es decir, que la autoridad exponga los fundamentos fácticos y jurídicos que le llevan a su decisión. En el contexto de la IA, la «motivación» puede provenir parcialmente de un algoritmo. Por ello, para cumplir con la obligación de motivar, la Administración debe ser capaz de traducir el razonamiento algorítmico a un lenguaje comprensible para el ciudadano.

La jurisprudencia C-203/22 clarifica que esta traducción no implica revelar *todo el código fuente* ni secretos técnicos detallados, sino explicar el criterio esencial que aplicó la máquina a los datos del individuo. En términos prácticos: si un sistema rechaza automáticamente una solicitud de beca porque el solicitante obtuvo un puntaje de 450 cuando el umbral es 500, la explicación al interesado debería indicar qué factores contribuyeron a ese puntaje y por qué no alcanzó el umbral. Podría señalar, por ejemplo: «Su solicitud ha sido valorada mediante un sistema automatizado considerando criterios objetivos (renta familiar, número de hermanos, expediente académico, etc.). El resultado obtenido (450 puntos) no alcanza el mínimo establecido (500 puntos) principalmente debido a que su renta familiar (X €) supera el promedio de solicitantes, lo cual reduce la puntuación según el baremo automatizado». Esta explicación, aunque simplificada, cumple con decirle al ciudadano qué datos suyos fueron relevantes y cómo. En cambio, comunicarle una mera cifra o una referencia interna al «algoritmo X vers. 3.4» no sería adecuado, por opaco. Además, el AI Act, con su artículo 86, consolida este derecho a obtener explicaciones claras y significativas cuando se usan sistemas de IA de alto riesgo en decisiones individuales.

Un aspecto relevante del art. 86 es que circunscribe el derecho a explicaciones *en la medida en que no esté previsto de otro modo por el Derecho de la Unión*. Esto se incluyó para evitar duplicidades con el RGPD, pero no significa que el ciudadano deba elegir entre RGPD y AI Act, sino que no haya acumulación de obligaciones idénticas. En la práctica, dado que el TJUE interpretó el RGPD en favor de un derecho a explicación, puede argumentarse que ya existe ese derecho «previsto por el Derecho de la Unión» (el RGPD). Sin embargo, la incorporación en el AI Act tiene valor añadido: (i) extiende la explicabilidad a sistemas de IA que quizá no impliquen datos personales (donde el RGPD no llegaría); (ii) cubre explícitamente los casos en que la decisión se basa «principalmente» en la IA y

35. M.C. Campos Acuña. «Sin transparencia no hay verdadera democracia y sin e-administración no hay transparencia», *Consultor de los ayuntamientos y de los juzgados: Revista técnica especializada en administración local y justicia municipal*, núm. 24, 2016, págs. 2761-2764.

afecta salud, seguridad o derechos fundamentales, aun si no hubiese un efecto jurídico formal, lo cual podría abarcar supuestos no claros en el RGPD; y (iii) lo enmarca como derecho fundamental ligado a la *accountability* de la IA.

Por otra parte, desde el punto de vista de la motivación jurídica, una explicación algorítmica podría formar parte del cuerpo de la resolución administrativa. Es decir, la propia resolución que se comunica al ciudadano puede incluir un apartado: «*Esta decisión ha sido adoptada mediante un procedimiento automatizado conforme a la normativa vigente. El sistema informático ha evaluado los siguientes datos… obteniendo el siguiente resultado… Por aplicación de la normativa X, y tras revisión por el órgano competente, se decide…*», lo que cierra el ciclo de la transparencia, integrando la lógica de la IA en la fundamentación formal del acto.

V. DIMENSIÓN PRÁCTICA DE LA SUPERVISIÓN Y EXPLICABILIDAD

La aplicación práctica de la normativa europea sobre decisiones automatizadas puede observarse en diversos sectores. En el ámbito educativo, el episodio del algoritmo de calificaciones en el Reino Unido de 2020 dio buena muestra de cómo un diseño opaco puede penalizar de manera desproporcionada a colectivos vulnerables. La rectificación política posterior confirmó la necesidad de transparencia y de mecanismos de recurso humano[36]. En el sector sanitario, estudios internacionales han demostrado que determinados algoritmos de priorización de pacientes reproducen sesgos estructurales (por ejemplo, asignando menos recursos a personas de raza negra al basarse en patrones de gasto sanitario), lo que refuerza el principio europeo de que la IA debe ser meramente asistencial: «el algoritmo aconseja, el profesional decide». En el terreno financiero, la sentencia del TJUE en el caso *Schufa* que hemos comentado consolidó la idea de que incluso la generación de un simple puntaje de solvencia puede constituir una decisión automatizada prohibida si determina de facto el acceso a un contrato. De ahí la exigencia de intervención humana significativa y del derecho a obtener explicaciones claras sobre los factores determinantes del resultado. En el ámbito laboral, los tribunales de Ámsterdam[37] y Bolonia[38] han

36. En 2020 el Reino Unido empleó un modelo matemático para calificar a los alumnos en sustitución de exámenes cancelados por la pandemia. El resultado fue polémico: el algoritmo redujo las calificaciones previstas de casi un 40% de los estudiantes, perjudicando especialmente a alumnos destacados de escuelas con menor rendimiento histórico (generalmente de entornos menos favorecidos). Las protestas forzaron la anulación de este sistema por considerarlo injusto y elitista.

37. En 2021, un tribunal de Ámsterdam ordenó a Uber readmitir a varios conductores que habían sido despedidos únicamente mediante un proceso automatizado (desactivación de sus cuentas por el algoritmo de la app, acusándolos de fraude sin pruebas). El juez concluyó que la rescisión fue ilegal por basarse exclusivamente en procesamiento automatizado sin evaluación humana, violando el artículo 22 RGPD.

38. Un tribunal de Bolonia falló en 2021 a favor de los repartidores de Deliveroo, declarando discriminatorio el algoritmo de asignación de pedidos de la empresa. El sistema penalizaba a los

declarado ilegales prácticas de plataformas digitales que despedían o penalizaban a trabajadores mediante procesos puramente algorítmicos. Estos fallos, junto con la Ley Rider española[39], reflejan una tendencia hacia la transparencia algorítmica como derecho colectivo de los trabajadores[40].

Finalmente, en el sector público, la anulación del sistema neerlandés SyRI marcó un hito en la protección de derechos fundamentales[41], al considerarse que su opacidad vulneraba la privacidad y podía dar lugar a discriminación indirecta. Casos como este muestran que la buena administración exige algoritmos explicables, revisables y sujetos a control judicial.

VI. CONCLUSIONES

El estudio realizado permite afirmar que la creciente incorporación de sistemas de decisión automatizada en la Administración pública plantea un desafío complejo que combina, al mismo tiempo, oportunidades y riesgos. La automatización ofrece la posibilidad de optimizar procesos, reducir cargas burocráticas y garantizar cierta uniformidad en la aplicación de las normas; sin embargo, también genera importantes tensiones en torno a la protección de los derechos fundamentales, la transparencia institucional y la legitimidad democrática de la acción administrativa.

Desde esta perspectiva, el ordenamiento jurídico europeo ha ido configurando un entramado normativo de gran densidad, en el que se entrelazan el Reglamento General de Protección de Datos (RGPD) y el Reglamento Europeo de Inteligencia Artificial (AI Act). Ambos textos cumplen funciones complementarias: mientras el RGPD establece límites sustantivos a la automatización de decisiones con efectos significativos sobre los individuos —prohibiéndolas en principio salvo supuestos excepcionales y garantizando la intervención humana, el derecho a impugnación y el acceso a explicaciones comprensibles—, el AI Act introduce mecanismos estructurales de control ex ante y ex post, que abarcan desde evaluaciones de impacto en derechos fundamentales hasta obligaciones de registro, trazabilidad, formación de supervisores y publicidad institucio-

---

riders que no se conectaban con frecuencia (por motivos como enfermedad o huelga), lo cual el juez consideró una discriminación indirecta y violación de derechos laborales. Ordenó a la empresa modificar el algoritmo y resarcir a los trabajadores

39. España aprobó en 2021 la pionera «Ley Rider» que, además de reconocer a los repartidores como empleados, estableció el derecho a la transparencia algorítmica: las empresas deben informar a la representación de los trabajadores sobre los parámetros, reglas e instrucciones de cualquier algoritmo que afecte a condiciones laborales (por ejemplo, reparto de pedidos, evaluaciones o despidos)

40. A. Todolí Signes. «Cambios normativos en la Digitalización del Trabajo: Comentario a la «Ley Rider» y los derechos de información sobre los algoritmos», *Iuslabor*, núm. 2, 2021.

41. L. Cotino Hueso. «Holanda: «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», *La Ley privacidad*, núm. 4 (Abril-junio 2020), 2020 (Ejemplar dedicado a: Lecciones aprendidas desde el confinamiento).

nal. De este modo, se consolida un modelo dual de garantías, que actúa tanto en el plano individual como en el sistémico.

La jurisprudencia del Tribunal de Justicia de la Unión Europea ha desempeñado un papel decisivo en la concreción de estas garantías, particularmente al reconocer que la información prevista en los artículos 13, 14 y 15 del RGPD debe traducirse en un auténtico derecho a explicación. Esta interpretación supone que los ciudadanos no solo tienen derecho a conocer que sus datos han sido tratados por un algoritmo, sino también a comprender de manera inteligible cuáles fueron los criterios relevantes que determinaron el resultado de la decisión. Con ello, se fortalece la capacidad de los individuos para ejercer una defensa efectiva de sus derechos y se evita que la opacidad tecnológica mine la efectividad del principio de motivación de los actos administrativos, que constituye una exigencia inherente al Estado de Derecho.

En este marco, la supervisión humana adquiere una centralidad indiscutible. El legislador europeo ha insistido en que la intervención de personas en el ciclo decisorio no puede reducirse a un control meramente formal, sino que debe revestir carácter sustantivo, dotando al operador público de la capacidad real de revisar, cuestionar o revertir la salida del sistema automatizado. Ello exige una adecuada formación técnica y jurídica de los funcionarios, así como una cultura organizativa que fomente el pensamiento crítico frente a los algoritmos y evite el riesgo del denominado «sesgo de automatización», es decir, la tendencia a confiar de manera acrítica en los resultados tecnológicos.

A su vez, la supervisión no se agota en la instancia individual. El modelo europeo contempla un entramado de supervisión institucional y democrática, que incluye el control ejercido por autoridades administrativas independientes (como las agencias de protección de datos y las futuras autoridades de supervisión de la IA), la revisión judicial y la transparencia frente a la sociedad civil mediante mecanismos como la base de datos europea de sistemas de alto riesgo. Esta pluralidad de controles persigue garantizar que la automatización se mantenga bajo el escrutinio público y no derive en un poder opaco ajeno a las exigencias del principio democrático.

La experiencia comparada en distintos Estados miembros y los casos judiciales analizados (como el asunto SyRI en Países Bajos o la jurisprudencia reciente sobre credit scoring) muestran que la falta de transparencia y de motivación en el uso de algoritmos puede desembocar en la anulación judicial de los sistemas automatizados y en un deterioro de la confianza ciudadana. Por ello, la exigencia de explicabilidad no debe considerarse un requisito accesorio, sino un componente esencial de la buena administración y de la preservación de la igualdad y la no discriminación en la acción pública.

En conclusión, la supervisión efectiva y la explicabilidad comprensible constituyen pilares indispensables para asegurar que el despliegue de la inteligencia artificial en el ámbito administrativo sea compatible con los valores constitucionales de la Unión Europea. Solo mediante la conjunción de estos elementos será posible integrar la innovación tecnológica en el funcionamiento de las instituciones públicas sin sacrificar los principios de legalidad, responsabilidad,

rendición de cuentas y respeto a los derechos fundamentales. La verdadera modernización administrativa, en consecuencia, no se mide únicamente por la adopción de nuevas tecnologías, sino por la capacidad de los poderes públicos para someterlas al Derecho y al control democrático, garantizando que la inteligencia artificial permanezca al servicio de las personas y nunca a la inversa.

BIBLIOGRAFÍA

Alayón Miranda, S. «El problema de la interpretabilidad de la Inteligencia Artificial y su impacto en la Administración Pública», *Revista Canaria de Administración Pública*, núm. 3, 2024, págs. 175-202.

Bellón Yturriaga, M. «Transparencia y protección de datos en el uso de la inteligencia artificial por la administración pública», *Revista de privacidad y derecho digital*, vol. 10, núm. 36, 2025, págs. 25-72.

Berenguer Albaladejo, C. «Transparencia y explicabilidad para prevenir la discriminación de los sistemas de inteligencia artificial: La interacción entre el RGPD y el RIA», en J.A. Moreno Martínez y P.J. Femenía López (coords.), *Inteligencia artificial y derecho de daños: Cuestiones actuales: Acorde al Reglamento (UE) 2024/1689*, Dykinson, Madrid, 2024, págs. 49-118.

Blázquez Ruiz, F.J. «La paradoja de la transparencia en la IA: Opacidad y explicabilidad. Atribución de responsabilidad», *Revista Internacional de Pensamiento Político*, núm. 17 (1), 2022, págs. 261–272. https://doi.org/10.46661/revintpensampolit.7526

Campos Acuña. M.C. «Sin transparencia no hay verdadera democracia y sin e-administración no hay transparencia», *Consultor de los ayuntamientos y de los juzgados: Revista técnica especializada en administración local y justicia municipal*, núm. 24, 2016, págs. 2761-2764.

Castagnedi Ramírez, A.E. «Inteligencia artificial: Cuando los algoritmos se convierten en neuronas», *Ius et scientia: Revista electrónica de Derecho y Ciencia*, vol. 8, núm. 2, 2022 (Ejemplar dedicado a: Medicina, biotecnología y derecho), págs. 136-145.

Castellanos Claramunt, J. «Reflexiones sobre el sorteo de la Champions: la aplicación acrítica de la tecnología», *The Conversation*, 15 de diciembre de 2021. Disponible en: https://theconversation.com/reflexiones-sobre-el-sorteo-de-la-champions-la-aplicacion-acritica-de-la-tecnologia-173786

Castellanos Claramunt, J. *DemocracIA: un análisis en clave constitucional*, Dykinson, Madrid, 2025.

Cervera Navas, L. «Las instituciones y organismos europeos de protección de datos: el Supervisor Europeo y el Comité Europeo de Protección de Datos», *El Cronista del Estado Social y Democrático de Derecho*, núm. 88-89 (Mayo-Junio), 2020 (Ejemplar dedicado a: Protección de datos: antes, durante y después del coronavirus), págs. 104-115.

Cotino Hueso, L. «Holanda: «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», *La Ley privacidad*, núm. 4 (Abril-junio 2020), 2020 (Ejemplar dedicado a: Lecciones aprendidas desde el confinamiento).

COTINO HUESO, L. «Transparencia y explicabilidad de la inteligencia artificial y «compañía» (comunicación, interpretabilidad, inteligibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad…). Para qué, para quién y cuánta», en L. Cotino Hueso y J. Castellanos Claramunt, *Transparencia y explicabilidad de la inteligencia artificial*, Tirant lo Blanch, Valencia, 2022, págs. 29-70.

COTINO HUESO, L., y J. Castellanos Claramunt, *Transparencia y explicabilidad de la inteligencia artificial*, Tirant lo Blanch, Valencia, 2022.

DE MARCOS FERNÁNDEZ, A. «Una doble historia de la inteligencia artificial: avance tecnológico y proceso de regulación en Europa», *Revista de privacidad y derecho digital*, vol. 9, núm. 34, 2024, págs. 26-89.

DÍAZ DÍAZ, E., y M.S.M. Del Busto Calosi. «De España a Iberoamérica: La influencia global del RGPD en las Leyes de Protección de Datos en Iberoamérica. Enfoque en Perú», *Revista Aranzadi Doctrinal*, núm. 7, 2025.

HERCE MAZA, J.I. «Buena administración de la transparencia de los algoritmos de la Administración Pública: un instrumento para el control de las cajas negras decisionales», en G. Vestri (dir.), *La disrupción tecnológica en la Administración Pública: retos y desafíos de la inteligencia artificial*, Thomson Reuters Aranzadi, Cizur Menor (Navarra), 2022, 123-138.

IVONE, V. «Artículo 22 del RGPD y tratamiento automatizado de datos», en M.A. López-Suárez (dir.), *Big data y protección de datos*, Tirant lo Blanch, Valencia, 2025, págs. 129-178.

LAZCOZ MORATINOS, G. y J.A. CASTILLO PARRILLA. «Valoración algorítmica ante los derechos humanos y el Reglamento General de Protección de Datos: el caso SyRI», *Revista chilena de derecho y tecnología*, vol. 9, núm. 1, 2020, págs. 207-225.

MUÑOZ RODRÍGUEZ, J. «Principios de protección de datos: licitud, lealtad, transparencia, minimización, exactitud, integridad y confidencialidad», *Economist & Jurist*, vol. 26, núm. 217, 2018, págs. 18-23.

NIEVA FENOLL, J. *Inteligencia artificial y proceso judicial*, Marcial Pons, Madrid, 2018.

RODRÍGUEZ AYUSO, J.F. «Artículo 64. Oficina de IA», en M. Barrio Andrés (dir.), *Comentarios al Reglamento Europeo de Inteligencia Artificial*, La Ley, Madrid, 2024, págs. 636-642.

SANCHO VILLA, D. «Las decisiones individuales automatizadas, incluida la elaboración de perfiles (Comentario al artículo 22 RGPD)», en A. Troncoso Reigada (dir.), *Comentario al Reglamento General de Protección de Datos y a la Ley Orgánica de Protección de Datos personales y Garantía de los Derechos Digitales*, vol. 1, Thomson Reuters Aranzadi, Cizur Menor (Navarra), 2021, págs. 1725-1745.

TODOLÍ SIGNES, A. «Cambios normativos en la Digitalización del Trabajo: Comentario a la «Ley Rider» y los derechos de información sobre los algoritmos», *Iuslabor*, núm. 2, 2021.

TORRES CARLOS, M.R., y L. MÍGUEZ MACHO. «Sistemas de IA prohibidos y sistemas de IA de alto riesgo», en M. Barrio Andrés (dir.), *El Reglamento Europeo de Inteligencia Artificial*, Tirant lo Blanch, Valencia, 2024, págs. 48-86.

VARGAS MARTÍNEZ. «La evolución de los conceptos de transparencia y explicabilidad: de los años 90 al Reglamento de Inteligencia Artificial», en I. Sánchez Frías e Y. Villegas Almagro (dirs.), *Derecho y entornos digitales*, Atelier, Barcelona, 2025, págs. 49-70.

VESTRI, G. «De la inteligencia artificial y otros factores: requisitos de transparencia y explicabilidad en la contratación pública», C*ontratación administrativa práctica: revista de la contratación administrativa y de los contratistas*, núm. Extra 1, 2025 (Ejemplar dedicado a: El impacto de la IA en la contratación pública).

La inteligencia artificial en la evaluación de solicitudes de asilo: un análisis crítico del caso alemán[1]

José Miguel Iturmendi Rubia
Profesor de Filosofía del Derecho
CUNEF Universidad

SUMARIO: I. INTRODUCCIÓN. II. CONTEXTUALIZACIÓN DEL PROBLEMA: EL CONTROL DE FRONTERAS EN EUROPA. III. EL CASO ALEMÁN. 1. Extracción de datos de dispositivos electrónicos en el sistema de asilo alemán: un análisis crítico desde la perspectiva jurídica y de derechos fundamentales. 2. Análisis lingüístico automatizado. 3. Críticas y debate jurídico 4. Jurisprudencia relevante. IV. CONCLUSIONES. BIBLIOGRAFÍA.

I. INTRODUCCIÓN

En los últimos años, el crecimiento y desarrollo de la tecnología ha producido una profunda transformación en las estructuras y lógicas de la administración pública. Entre los cambios más significativos, destaca la creciente automatización de procesos decisorios, respaldada por el desarrollo de sistemas algorítmicos y de inteligencia artificial (en adelante, IA)[2]. Esta reconfiguración tecnológica se inscribe en una narrativa de eficiencia, neutralidad y modernización del aparato estatal, que busca minimizar la discrecionalidad administrativa y acelerar la ges-

1. Esta investigación se ha realizado en el marco del proyecto de I+D+i *Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas* [2023-2026] (PID2022-136439OB-I00), financiado por MCIN/AEI/10.13039/501100011033/ y «FEDER Una manera de hacer Europa».

2. Para profundizar en esta cuestión, vid. A. Palma Ortigosa, *Decisiones automatizadas y protección de datos: especial atención a los sistemas de inteligencia artificial*, Dykinson, Madrid, 2022.

tión de procedimientos complejos. No obstante, el uso de estas herramientas plantea un número creciente de cuestiones críticas, especialmente cuando se aplican en ámbitos que afectan directamente a derechos fundamentales.

Uno de los terrenos donde esta tensión se manifiesta con mayor intensidad es el de la protección internacional y, más específicamente, la evaluación de la credibilidad en las solicitudes de asilo. Este procedimiento, ya de por sí delicado por el desequilibrio de poder entre solicitante y administración, se ha convertido en escenario de experimentación tecnológica bajo el pretexto de aumentar la objetividad y prevenir el fraude. Alemania, en particular, ha implementado dos mecanismos de carácter digital que revelan con claridad las implicaciones jurídicas y éticas de esta tendencia: por un lado, la extracción y análisis de datos contenidos en dispositivos electrónicos de los solicitantes de asilo; por otro, el uso de software de reconocimiento lingüístico para inferir el país de origen de los demandantes.

Estos instrumentos, si bien concebidos como auxiliares del proceso de toma de decisiones, han sido objeto de crecientes controversias. Diversos estudios han señalado su eficacia limitada, su falta de transparencia algorítmica y su potencial para generar decisiones erróneas o discriminatorias. La ausencia de garantías suficientes para salvaguardar los derechos de los solicitantes y la opacidad respecto al funcionamiento de las herramientas tecnológicas agravan la desconfianza hacia un modelo que amenaza con desplazar el juicio humano por una automatización opaca y posiblemente arbitraria. En ese sentido, la tecnologización de la administración no solo transforma los procedimientos, sino también los principios que los sustentan, como la proporcionalidad, la motivación individualizada de las resoluciones y el acceso efectivo a recursos.

Desde una perspectiva jurídica, estos desarrollos obligan a repensar el alcance del principio de legalidad y el papel del control judicial en un entorno cada vez más tecnificado. La implementación de tecnologías en la esfera administrativa, en particular cuando se afecta el ejercicio de derechos como el asilo, no puede prescindir de un marco normativo robusto que garantice la transparencia, la explicabilidad de los algoritmos utilizados, y mecanismos efectivos de revisión y reparación. En ausencia de estas condiciones, se corre el riesgo de consolidar una forma de «automatismo burocrático» incompatible con los valores democráticos del Estado de Derecho.

Como señala el profesor Castellanos[3], la IA representa un desafío constitucional significativo al influir cada vez más en la toma de decisiones que afectan derechos fundamentales como la privacidad, la igualdad, la libertad de expresión y el debido proceso. La creciente autonomía de los algoritmos y la opacidad de sus procesos pueden reproducir sesgos, generar discriminación y debilitar las garantías propias de un Estado de derecho. Frente a esto, se destaca la necesidad de establecer marcos normativos sólidos y éticamente fundamentados que

3. J. Castellanos Claramunt, «Sobre los desafíos constitucionales ante el avance de la Inteligencia Artificial. Una perspectiva nacional y comparada», *Revista de Derecho Político*, n.º 118, 2023, pp. 261-287.

aseguren la transparencia, la rendición de cuentas y la supervisión humana. Europa ha apostado por una regulación centrada en los derechos, en contraste con modelos como el estadounidense —más orientado al mercado— y el chino —con fuerte control estatal—. Además, es clave el papel de los tribunales y las instituciones democráticas para evitar que la IA erosione los principios democráticos y consolide formas de desigualdad estructural o control político. En definitiva, se requiere un enfoque multidisciplinar y preventivo que sitúe al ser humano y sus derechos en el centro del desarrollo tecnológico.

El presente artículo se propone examinar críticamente el uso de estas tecnologías en el procedimiento de asilo en Alemania, atendiendo tanto a su base legal como a su aplicación práctica y su recepción por parte de la jurisprudencia y la doctrina. A través de un enfoque multidisciplinar que articula elementos del derecho constitucional, el derecho administrativo y la teoría de los derechos fundamentales, se analizarán los desafíos que supone la integración de sistemas algoritmos en contextos de alta sensibilidad humana y jurídica. Asimismo, se considerará el impacto de estas prácticas sobre el principio de igualdad de trato, el derecho a un procedimiento justo y la obligación estatal de proteger a las personas solicitantes de protección internacional.

Lejos de rechazar per se el uso de la tecnología en la administración, este estudio pretende contribuir a una evaluación crítica y fundamentada de sus límites y condiciones de legitimidad. Para ello, se examinarán los fundamentos normativos que han permitido la adopción de estas prácticas, los mecanismos de control disponibles y las posibles vías para fortalecer el marco garantista frente a una tendencia creciente hacia la despersonalización de la justicia administrativa. Al hacerlo, se espera aportar elementos útiles para un debate informado sobre el papel de la tecnología en la administración de justicia, especialmente en escenarios donde están en juego la vida, la libertad y la dignidad de las personas.

II. CONTEXTUALIZACIÓN DEL PROBLEMA: EL CONTROL DE FRONTERAS EN EUROPA

En las últimas décadas, el control de las fronteras exteriores de la Unión Europea se ha convertido en un eje prioritario de la agenda política y legislativa comunitaria, como lo evidencia la creciente asignación de recursos a agencias como Frontex[4] (Agencia Europea de la Guardia de Fronteras y Costas), cuyo presupuesto ha pasado de 6 millones de euros en 2005 a más de 845 millones en 2023, y la entrada en vigor de sistemas como ETIAS[5] (El Sistema Europeo

4. Comisión Europea, «Agencia Europea de la Guardia de Fronteras y Costas (Frontex)», disponible en: https://european-union.europa.eu/institutions-law-budget/institutions-and-bodies/search-all-eu-institutions-and-bodies/frontex_es

5. Comisión Europea, «Sistema Europeo de Información y Autorización de Viajes (ETIAS)», disponible en: https://travel-europe.europa.eu/etias_en

de Información y Autorización de Viajes) o el Sistema de Entradas y Salidas (EES)[6], diseñados para registrar electrónicamente los movimientos de nacionales de terceros países. Estas iniciativas reflejan una voluntad clara de reforzar los mecanismos de control y vigilancia, articulando un marco político y tecnológico que busca anticipar, gestionar y disuadir la movilidad irregular, en un contexto marcado por crisis migratorias sucesivas y preocupaciones en torno a la seguridad interior de la Unión Europea, convirtiéndose en un eje prioritario de la agenda política y legislativa comunitaria. Este proceso ha estado profundamente influido por la creciente digitalización y el desarrollo de tecnologías avanzadas, particularmente la IA, que han sido incorporadas progresivamente a los sistemas de vigilancia y gestión migratoria. En este contexto, diversas investigaciones académicas han puesto de relieve los desafíos que esta transformación plantea desde el punto de vista de los derechos fundamentales, la legalidad y la ética democrática.

Uno de los aportes más relevantes en este ámbito es el de María Avello Martínez, quien analiza críticamente la utilización de la inteligencia artificial y el big data en las políticas migratorias de la Unión Europea. La autora sostiene, en su trabajo «EU Borders and Potential Conflicts Between New Technologies and Human Rights»[7], que, si bien estas herramientas tecnológicas han sido justificadas por la necesidad de reforzar la seguridad y prever los flujos migratorios, en la práctica están produciendo graves tensiones con el respeto a los derechos humanos. Señala que los sistemas automatizados implementados por agencias como EU-LISA[8] (Agencia de la Unión Europea para la Gestión Operativa de Sistemas Informáticos de Gran Magnitud en el Espacio de Libertad, Seguridad y Justicia) carecen de suficientes garantías jurídicas y mecanismos de control, lo que puede traducirse en vulneraciones del derecho a la privacidad, tratamientos discriminatorios y decisiones automatizadas sin posibilidad de impugnación. Para Avello Martínez, la regulación ética propuesta por la Comisión Europea[9] constituye un paso positivo, pero insuficiente, por lo que aboga por un marco jurídico más robusto, basado en el principio de precaución, que priorice el respeto a la dignidad humana[10].

6. Comisión Europea, «Sistema de Entradas y Salidas (EES)», disponible en: https://travel-europe.europa.eu/ees/what-ees_es

7. M. Avello Martínez, «EU Borders and Potential Conflicts Between New Technologies and Human Rights, Peace & Security – Paix et Sécurité Internationales», n.º 11, Universidad de Cádiz, Cádiz, 2023.

8. Comisión Europea, «Agencia de la Unión Europea para la gestión operativa de sistemas informáticos a gran escala en el espacio de libertad, seguridad y justicia (eu-LISA)», disponible en: https://european-union.europa.eu/institutions-law-budget/institutions-and-bodies/european-union-agency-operational-management-large-scale-it-systems-area-freedom-security-and_es

9. European Commission: «Directorate-General for Communications Networks, Content and Technology and Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji, Directrices éticas para una IA fiable», Publications Office, 2019, disponible en https://data.europa.eu/doi/10.2759/14078

10. M. Avello Martínez, «EU Borders and Potential Conflicts Between New Technologies and Human Rights», cit., p. 10.

De forma complementaria, Andrea Romano[11] centra su análisis en los sistemas de inteligencia artificial emocional aplicados a la gestión fronteriza, como el proyecto iBorderCtrl[12] (Intelligent Portable Border Control System). Este sistema, basado en el análisis de microexpresiones faciales para detectar presuntas mentiras en entrevistas automatizadas, plantea, según el autor, serias dudas sobre su fiabilidad científica y su compatibilidad con los derechos fundamentales. Romano destaca que la automatización de decisiones, enmarcada en el contexto del proyecto piloto iBorderCtrl financiado por la Unión Europea entre 2016 y 2019 y sujeto a controversias por su falta de validación científica y su posible vulneración del Reglamento General de Protección de Datos (RGPD), en contextos tan delicados como el migratorio puede derivar en discriminación, sesgos algorítmicos y una deshumanización del trato hacia las personas solicitantes de asilo o migrantes. En su opinión, el uso de estas tecnologías vulnera la dignidad humana al convertir a los individuos en objetos de análisis algorítmico, por lo que propone su prohibición expresa en el ámbito fronterizo, al menos hasta contar con garantías suficientes de transparencia, supervisión y rendición de cuentas.

Por su parte, Sandra Alonso Tomé[13] aporta una perspectiva más centrada en el análisis normativo, explorando el desarrollo legislativo europeo en materia de inteligencia artificial y su aplicación específica en el control de fronteras. La autora subraya que el Reglamento (UE) 2024/1689 representa un avance significativo al establecer directrices armonizadas para los sistemas de IA, pero advierte que este marco aún no es suficiente para garantizar una protección efectiva de los derechos fundamentales en un ámbito tan sensible. Alonso Tomé destaca que, si bien la UE ha apostado por un enfoque ético de la IA, los sistemas implementados, como ETIAS o iBorderCtrl, siguen careciendo de mecanismos eficaces de supervisión y de recursos adecuados para impugnar decisiones automatizadas. Considera que, para evitar la consolidación de una «sociedad de vigilancia», es imprescindible diseñar un marco jurídico que no solo sea flexible y tecnológicamente adaptativo, sino que sitúe la dignidad y los derechos de las personas en el centro de cualquier desarrollo o aplicación tecnológica.

En conjunto, los trabajos de Avello Martínez, Romano y Alonso Tomé ponen de manifiesto que la gestión de las fronteras exteriores de la UE mediante inteligencia artificial no puede abordarse exclusivamente desde lógicas de eficiencia o seguridad. Por el contrario, los autores coinciden en que el despliegue de estas tecnologías debe ir acompañado de un compromiso firme con los principios del Estado de derecho, la rendición de cuentas y la protección de los derechos humanos. Resulta indispensable, por tanto, que las instituciones europeas

11. A. Romano, «Derechos fundamentales e inteligencia artificial emocional en iBorderCtrl: retos de la automatización en el ámbito migratorio», *Revista Catalana de Dret Públic*, n.º 66, Generalitat de Catalunya, Barcelona, 2023.

12. https://cordis.europa.eu/project/id/700626

13. S. Alonso Tomé, «La aplicación de la inteligencia artificial en los controles de las fronteras exteriores de la Unión Europea: Regulación y desafíos», *Revista de Estudios Europeos*, nº 85, Universidad de Valladolid, Valladolid, 2025.

articulen mecanismos de regulación más exigentes, capaces de anticipar los efectos indeseados de los sistemas automatizados y de garantizar que la innovación tecnológica no se convierta en una vía de legitimación de prácticas discriminatorias o contrarias a la dignidad humana.

Este debate se inserta en un contexto más amplio de transformación del espacio europeo de libertad, seguridad y justicia, en el que la tecnología ocupa un lugar central en la configuración de nuevas formas de gobernanza. La tendencia a externalizar el control migratorio mediante acuerdos con terceros países, como el pacto entre la UE y Turquía en 2016[14] o el más reciente acuerdo con Túnez en 2023[15], ha tenido consecuencias directas sobre los derechos fundamentales de las personas migrantes. Estas políticas han sido criticadas por organizaciones de derechos humanos[16], que denuncian la falta de garantías jurídicas en los países receptores, la exposición de las personas migrantes a condiciones inhumanas y la imposibilidad de acceder a mecanismos efectivos de protección internacional. La transformación del espacio europeo de libertad, seguridad y justicia ha situado a la tecnología en el centro de nuevas formas de gobernanza. El fortalecimiento de agencias como Frontex, junto con la externalización del control migratorio mediante acuerdos con terceros países y la proliferación de sistemas interconectados de bases de datos biométricos y algoritmos predictivos, configuran un escenario en el que la frontera deja de ser únicamente una línea geográfica para convertirse en un espacio tecnológico de vigilancia, selección y control. Ante esta realidad, los marcos normativos europeos deben evolucionar no solo para acompañar la innovación, sino también para limitar sus excesos y garantizar la preservación del carácter garantista del orden jurídico. El control de las fronteras en la actualidad no puede desligarse de la profunda transformación que están experimentando las tecnologías aplicadas a la movilidad humana. La IA ofrece oportunidades indudables para mejorar la gestión y seguridad, pero también plantea riesgos estructurales que deben ser abordados con urgencia desde un enfoque integral. Ello exige no solo el fortalecimiento del marco legal, incluyendo propuestas concretas como la introducción de evaluaciones de impacto en derechos humanos[17] obligatorias para cualquier sistema de IA aplicado al control migratorio, la creación de mecanismos independientes de supervisión con capacidad sancionadora, y el establecimiento de vías efectivas de recurso para las personas afectadas por deci-

14. https://www.europarl.europa.eu/topics/es/article/20170426STO72401/relaciones-de-la-ue-con-turquia-entre-la-cooperacion-y-las-tensiones#:~:text=Acuerdo%20de%20la%20UE%20y%20Turqu%C3%ADa%20sobre%20migraci%C3%B3n&text=Lea%20m%C3%A1s%20sobre%20la%20respuesta,la%20UE%20para%20los%20refugiados.

15. https://www.europarl.europa.eu/news/es/agenda/briefing/2023-09-11/10/ue-tunez-debate-con-la-comision-y-el-consejo-sobre-el-acuerdo-migratorio

16. Amnistía Internacional:https://www.es.amnesty.org/en-que-estamos/noticias/noticia/articulo/ue-tunez-acuerdo-migracion/

17. L. Cotino Hueso, *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, Thomson Reuters Aranzadi, Navarra, España, 2022; y Cotino Hueso, L., «Nuevo paradigma en las garantías de los derechos fundamentales…», en la misma obra, 2022, pp. 69–105.

siones automatizadas, tal como han propuesto organizaciones como el European Center for Not-for-Profit Law[18] y académicos expertos en derechos digitales y migración, sino también una vigilancia continua desde la sociedad civil, la academia y los organismos garantes de derechos, para asegurar que el avance tecnológico no se traduzca en una regresión de las libertades y principios democráticos que definen a la Unión Europea.

III. EL CASO ALEMÁN

1. Extracción de datos de dispositivos electrónicos en el sistema de asilo alemán: un análisis crítico desde la perspectiva jurídica y de derechos fundamentales

Uno de los desarrollos más controvertidos y discutidos en el ámbito del derecho de asilo en Alemania es la capacidad de las autoridades migratorias para acceder y analizar la información contenida en los dispositivos electrónicos de los solicitantes de asilo. Esta práctica fue formalmente introducida mediante una reforma de la Ley de Asilo (Asylgesetz) en el año 2017, permitiendo que la Oficina Federal de Migración y Refugiados (Bundesamt für Migration und Flüchtlinge, BAMF) tenga acceso a datos personales contenidos en teléfonos móviles, tabletas u otros dispositivos, siempre que el solicitante no disponga de documentos oficiales que acrediten su identidad y nacionalidad. La medida se enmarca dentro de un intento de mejorar la eficiencia del sistema de asilo y evitar fraudes en la identificación, pero ha generado una considerable preocupación desde el punto de vista de los derechos fundamentales. Según establece la legislación, el acceso a dicha información sólo puede tener lugar cuando no exista un medio menos intrusivo para la determinación de la identidad. Además, se exige que el tratamiento de estos datos se lleve a cabo por personal especialmente capacitado y con autorizaciones comparables a las de una autoridad judicial.

Desde la perspectiva del Sistema Europeo Común de Asilo (SECA), este tipo de prácticas debe encuadrarse dentro de un marco garantista que armonice los procedimientos sin sacrificar los derechos fundamentales. La Directiva 2013/32/UE sobre procedimientos comunes para la concesión o la retirada de la protección internacional establece el principio de una decisión individual, objetiva e imparcial, basada en una evaluación completa de los elementos pertinentes[19].

18. ECNL: https://careernext.ceu.edu/resource/european-center-not-profit-law-ecnl

19. J. Cruz Ángeles y G. Vestri, «La nueva 'ley' de inteligencia artificial: una aliada necesaria para gestionar controles fronterizos, de asilo y migratorios en la Unión Europea», en *La disrupción tecnológica en la Administración Pública: retos y desafíos de la inteligencia artificial*, Aranzadi, Cizur Menor, 2022, pp. 97–121.

El trabajo de Andrea Romano[20] destaca particularmente la introducción de la tecnología en la evaluación de la credibilidad de los solicitantes de asilo, evidenciando cómo el análisis de los datos contenidos en los dispositivos electrónicos y los sistemas de análisis lingüístico automatizados se han consolidado como herramientas claves en Alemania. Esta doble vía de intervención tecnológica, aunque orientada a incrementar la eficiencia del proceso, pone en tensión los principios de objetividad y garantía procesal, generando serias dudas sobre su impacto en los derechos fundamentales. El uso de estas herramientas se enmarca en una estrategia de digitalización promovida por la BAMF tras la crisis migratoria de 2015-2016, momento en que se evidenciaron grandes retrasos y disfunciones en los procedimientos de asilo. La llamada Agenda para la Digitalización incluyó como eje central la incorporación de sistemas automatizados para reforzar la verificación de identidad. El problema es que estas prácticas no han ido acompañadas de un desarrollo suficiente de garantías jurídicas, lo que ha sido señalado por diversos autores y organizaciones como una carencia grave en el diseño institucional del sistema.

Por su parte, el cuadernillo elaborado por CEAR y CICrA Justicia Ambiental advierte que la digitalización, si bien ofrece oportunidades para la gestión migratoria, también puede tener efectos profundamente negativos sobre los derechos humanos, especialmente en un contexto de crisis climática y desplazamientos forzados. La vigilancia algorítmica, el perfilado digital y el monitoreo de movimientos mediante tecnologías de inteligencia artificial han transformado las fronteras en espacios de control digital intensivo, afectando de forma desproporcionada a personas refugiadas y solicitantes de asilo. Este enfoque tecnocrático de la gestión migratoria se alinea con un modelo de gobernanza fronteriza que tiende a la externalización de responsabilidades y la automatización del control. Caterina Rodelli[21], por ejemplo, señala que los sistemas de IA aplicados a la migración no son neutrales, y que contribuyen a categorizar a las personas según criterios opacos, muchas veces descontextualizados y sin mecanismos de rendición de cuentas efectivos.

La evaluación de la identidad mediante tecnología se justifica en la necesidad de acelerar los procesos y contrarrestar el uso fraudulento del sistema. Sin embargo, esta práctica puede redundar en una carga desproporcionada sobre los solicitantes que, por motivos ligados a la persecución, carecen de documentación válida. En estos casos, la lectura de dispositivos electrónicos se convierte en una fuente principal de prueba, pese a que la fiabilidad del contenido extraído y su interpretación puede estar sujeta a errores o malentendidos culturales, lingüísticos o contextuales.

---

20. A. Romano, «Derechos fundamentales e inteligencia artificial emocional en iBorderCtrl: retos de la automatización en el ámbito migratorio», cit., pp. 237–252.

21. https://www.elperiodico.com/es/internacional/20230926/oeneges-denuncian-union-europea-migrantes-inteligencia-artificial-reconocimiento-facial-derechos-92257762

https://algorights.org/la-ley-de-ia-de-la-ue-fracasa-en-garantizar-la-proteccion-de-los-derechos-humanos/

Desde el marco del SECA, la jurisprudencia del Tribunal de Justicia de la Unión Europea (TJUE) ha recordado que cualquier procedimiento de asilo debe respetar el derecho a la buena administración, incluyendo el derecho a ser escuchado y a recibir una decisión motivada. Así, la extracción de datos sin posibilidad de contradicción podría vulnerar este principio esencial.

Informes como el de CEAR (2024) alertan sobre los efectos de la digitalización acelerada de los procedimientos de asilo, subrayando la necesidad de garantizar que las innovaciones tecnológicas no comprometan los derechos fundamentales. Se pone de relieve, además, el impacto desigual que estas medidas pueden tener sobre ciertos colectivos vulnerables, como mujeres, niños o personas LGBTIQ+, cuyas narrativas pueden ser objeto de duda o malinterpretación automatizada.

Una perspectiva crítica ambiental complementa este análisis. La digitalización no es un proceso neutro en términos ecológicos ni políticos. Como expone Beatriz Felipe Pérez[22], el desarrollo digital global necesita una infraestructura que implica consumo energético intensivo, minería de minerales críticos y explotación de territorios, lo cual genera nuevos desplazamientos forzados, especialmente en el Sur global.

Desde una perspectiva constitucional, la aplicación de estas medidas debe considerarse bajo el prisma del principio de legalidad y del control judicial efectivo. Si bien el interés del Estado en verificar la identidad de los solicitantes es legítimo, no puede hacerse a expensas de los derechos fundamentales. En este contexto, la transparencia algorítmica emerge como una garantía indispensable para asegurar que el procesamiento automatizado de datos personales sea verificable, explicable y sujeto a control por parte de los propios interesados y de las autoridades judiciales independientes. En definitiva, se impone la necesidad de repensar el equilibrio entre seguridad y libertad en el contexto migratorio, reconociendo que los solicitantes de asilo no pierden sus derechos fundamentales al cruzar una frontera. La tecnología, si bien puede ser una aliada en la mejora de los procedimientos administrativos, debe estar subordinada a principios éticos, democráticos y jurídicos que aseguren su uso justo, proporcionado y respetuoso de la condición humana. Ello exige políticas de transparencia, participación de la sociedad civil en los procesos de diseño tecnológico y mecanismos efectivos de tutela judicial para las personas afectadas.

2. Análisis lingüístico automatizado

Otro de los instrumentos tecnológicos introducidos por las autoridades alemanas para evaluar las solicitudes de asilo ha sido el análisis lingüístico automatizado. Desde 2017, la Oficina Federal de Migración y Refugiados (BAMF) ha desarrollado un sistema basado en inteligencia artificial que graba y analiza la

22. B. Felipe Pérez, «Migraciones climáticas. Una aproximación al panorama actual», Fundación Ecología y Desarrollo (ECODES), 2018, p. 31.

forma de hablar de los solicitantes, con el objetivo de identificar su probable país de origen a partir del dialecto o variante lingüística que utilizan[23]. Esta práctica, enmarcada en la llamada Agenda para la Digitalización del BAMF, busca acelerar los tiempos de resolución de los procedimientos y reducir la incidencia de fraudes. El procedimiento consiste en una grabación de voz de aproximadamente dos minutos, durante la cual el solicitante describe una imagen. El sistema procesa la grabación, extrae rasgos fonológicos y léxicos, y los compara con una base de datos de muestras lingüísticas previamente clasificadas[24]. Inicialmente aplicado a variedades del árabe (levantino, magrebí, del Golfo, egipcio, iraquí), el sistema ha ido incorporando progresivamente otras lenguas como el dari, el pastún o el farsi[25]. El informe resultante asigna una probabilidad a diversos países o regiones de origen, siendo utilizado por los funcionarios como elemento auxiliar en la evaluación de la credibilidad[26].

Aunque el objetivo declarado es incrementar la objetividad y eficacia del sistema, el uso de este software ha suscitado críticas en múltiples planos. La literatura especializada ha cuestionado su fiabilidad, apuntando que el entrenamiento del algoritmo se basa en conjuntos de datos muy reducidos, sin transparencia sobre los criterios lingüísticos empleados ni posibilidad de auditoría externa[27]. Las tasas de acierto reconocidas oscilan, siendo superiores al 90 % para algunos dialectos, pero descendiendo a menos del 65 % en ciertos contextos, como el árabe sudanés[28]. Casos documentados evidencian errores graves: un solicitante kurdo fue identificado erróneamente como hablante de turco y hebreo, lo que condujo a una resolución negativa de su solicitud. En otro expediente, se atribuyó un origen egipcio a un demandante palestino sirio con base en un supuesto acento[29].

Estas prácticas han sido señaladas por organizaciones independientes como potencialmente discriminatorias. No sólo por el sesgo técnico, sino por la lógica subyacente que tiende a desconfiar de la declaración del solicitante, invirtiendo de facto la carga de la prueba[30]. Desde el punto de vista lingüístico, además, se advierte que el habla de una persona puede verse modificada por numerosos factores: el entorno de migración, la escolarización, la influencia del intérprete, o incluso el estrés del procedimiento. La posibilidad de que una persona adapte o mimetice ciertas formas de hablar complica aún más la supuesta objetividad del sistema[31].

---

23. A. Romano, «Credibilidad y derecho: el rol de la tecnología en la evaluación de las solicitudes de protección internacional. El caso de la ley de asilo alemana», *IDP. Revista de Internet, Derecho y Política*, n.º 39, pp. 1-12, doi:10.7238/idp.v0i39.417280. 2023, p. 7.
24. BAMF, Agenda Digitalisierung, 2022.
25. A. Romano, «Credibilidad y derecho…», cit., p. 12.
26. EUAA, Practical guide on evidence assessment, 2023, p. 176.
27. Gesellschaft für Freiheitsrechte (GFF), Analyseverfahren des BAMF, 2020. https://edri.org/our-work/germany-invading-refugees-phones-security-or-population-control/
28. A. Romano, «Credibilidad y derecho…», cit., p. 13.
29. Deutsche Welle, «Flüchtling aus Syrien falsch erkannt», 2023.
30. European Digital Rights (EDRi), Automated decision-making in migration, 2022.
31. A. Romano, «Credibilidad y derecho…», cit., p. 16.

Jurídicamente, el análisis lingüístico automatizado plantea serias dudas. Como recuerda la doctrina, la determinación de la nacionalidad o identidad del solicitante de asilo es un aspecto clave en la evaluación de su credibilidad, pero no puede fundarse exclusivamente en pruebas técnicas cuya validez no ha sido consensuada ni verificada por peritos[32]. En varias resoluciones, tribunales administrativos alemanes han subrayado que este tipo de informes no puede reemplazar al juicio del funcionario responsable ni servir como único fundamento para denegar protección[33]. El derecho a un procedimiento justo exige, además, que el solicitante tenga acceso al informe, pueda impugnarlo con asistencia técnica, y que su voluntad de no colaborar con el sistema no se traduzca automáticamente en una denegación.

El análisis comparado muestra que, aunque algunos Estados europeos recurren a peritajes lingüísticos en procesos de asilo, Alemania es el único país que ha implementado una herramienta automatizada de este tipo a escala estatal. Las agencias europeas estudian replicar el modelo, con la creación de plataformas comunes de apoyo técnico, pero persisten fuertes reservas éticas y jurídicas[34]. En todo caso, como señalan expertos y organismos internacionales, cualquier uso de tecnologías para evaluar credibilidad debe guiarse por los principios de proporcionalidad, transparencia, y respeto a la dignidad de los solicitantes[35].

3. Críticas y debate jurídico

La implementación de tecnologías como la extracción de datos de dispositivos electrónicos y el análisis lingüístico automatizado en los procedimientos de asilo ha suscitado un debate jurídico de considerable envergadura[36]. Desde una perspectiva doctrinal, estas herramientas han sido objeto de escrutinio por su potencial para vulnerar principios fundamentales del derecho administrativo y constitucional[37], especialmente en lo que respecta a la proporcionalidad, la intimidad, la presunción de veracidad y el derecho a un procedimiento justo[38].

32. W. Marx, «Zur Verfassungsmäßigkeit automatisierter Sprachanalyse», Asylmagazin, 2021.

33. Verwaltungsgericht Berlin, Urteil vom 10.04.2021 – VG 12 K 34.21.doi.org/10.5771/9783748913979

34. EUAA, Digital strategy, 2022.

35. ACNUR, Procedural Standards for Refugee Status Determination, 2019.

36. M. Avello Martínez, EU borders y potential conflicts between new technologies y human rights, Peace & Security – Paix et Sécurité Internationales, nº 11, 2023. Disponible en: https://dialnet.unirioja.es/descarga/articulo/9115695.pdf.

37. J. Castellanos Claramunt, «Sobre los desafíos constitucionales ante el avance de la Inteligencia Artificial. Una perspectiva nacional y comparada», *Revista de Derecho Político*, nº 118, 2023, pp. 261–287.

38. J. Castellanos Claramunt, «Una reflexión acerca de la influencia de la inteligencia artificial en los derechos fundamentales», en F. Ramón Fernández (dir.), *Ciencia de datos y perspectivas de la inteligencia artificial*, Tirant lo Blanch, Valencia, 2024, pp. 271-300; L. Cotino Hueso, «Big data e inteligencia artificial. Una aproximación a su tratamiento jurídico desde los derechos fundamentales», *Dilemata*, nº 24, 2017, pp. 131–150.

En cuanto a la extracción de datos de dispositivos electrónicos, diversos autores han alertado sobre la posible vulneración del derecho a la autodeterminación informativa, reconocido en Alemania por el Tribunal Constitucional Federal desde la célebre sentencia del Censo (Volkszählungsurteil) de 1983, que estableció un derecho fundamental a controlar la propia información personal[39]. El acceso indiscriminado a los contenidos de móviles o tabletas, sin las debidas garantías judiciales, representa un riesgo cierto de extralimitación del poder administrativo[40]. Si bien la normativa vigente exige que esta medida se aplique sólo cuando no existan otros medios menos invasivos para verificar la identidad, informes como los elaborados por la Gesellschaft für Freiheitsrechte (GFF) han documentado su uso sistemático en numerosos expedientes sin una justificación adecuada de necesidad ni proporcionalidad[41].

También desde el punto de vista del derecho europeo, la Directiva 2013/32/UE sobre procedimientos de asilo exige que las decisiones se basen en una evaluación individual y objetiva[42]. Además, la jurisprudencia del Tribunal Europeo de Derechos Humanos ha establecido que el uso de tecnologías invasivas debe respetar el derecho a la vida privada (artículo 8 del CEDH), siendo necesaria una base legal clara, un objetivo legítimo y una proporcionalidad estricta[43] (caso Szabó y Vissy c. Hungría, 2016)[44].

Autores como Lenaerts han enfatizado que la digitalización administrativa no puede desarrollarse a costa de debilitar el control judicial, ya que la eficacia nunca puede prevalecer sobre la legalidad[45]. En este sentido, el principio de tutela judicial efectiva se convierte en un criterio central para evaluar la legitimidad de las tecnologías utilizadas en procedimientos que afectan derechos fundamentales.

Desde la doctrina española, autores como Asunción Ventura Franch han subrayado que la utilización de sistemas algorítmicos en la administración exige una redefinición del principio de transparencia y del derecho de acceso a la información pública, especialmente cuando estas decisiones afectan a personas en situación de vulnerabilidad[46]. Del mismo modo, Rafael Bustos Gisbert ha advertido sobre los riesgos de delegar la función administrativa en sistemas técnicos opacos, pues ello puede conducir a un vaciamiento del principio de

39. Tribunal Constitucional Federal Alemán, BVerfGE 65, 1 – Volkszählungsurteil, 1983.

40. F. Pulido Catalán; F. Castro-Rial Garrone; A. Jarillo Aldeanueva, «El intercambio de información entre estados de la UE y las nuevas amenazas a la seguridad: estudio particular de la Guardia Civil», 2023. Disponible en: https://dialnet.unirioja.es/servlet/exttes?codigo=315246

41. Gesellschaft für Freiheitsrechte (GFF), Digitalisierte Asylverfahren auf dem Prüfstand, Informe, Berlín, 2021.

42. Directiva 2013/32/UE del Parlamento Europeo y del Consejo, de 26 de junio de 2013, sobre procedimientos comunes para la concesión o la retirada de la protección internacional.

43. J. Castellanos Claramunt, «Una reflexión acerca de la influencia de la inteligencia artificial en los derechos fundamentales», cit., pp. 271–300.

44. TEDH, Szabó y Vissy c. Hungría, núm. 37138/14, sentencia de 12 de enero de 2016.

45. K. Lenaerts, «La protección judicial efectiva en el Derecho de la Unión Europea», *Revista de Derecho Comunitario Europeo,* núm. 52, 2015, pp. 45-79.

46. A. Ventura Franch, «Inteligencia artificial, administración y garantías», *Revista Aragonesa de Administración Pública*, núm. 57, 2021, pp. 11-38.

responsabilidad y a una erosión del garantismo propio del Estado de Derecho[47]. ¿Puede realmente hablarse de garantías efectivas cuando la persona afectada no comprende ni puede rebatir los mecanismos que determinaron su destino?

La jurisprudencia del Tribunal Constitucional español también ofrece una base de análisis crítico. En sentencias como la STC 292/2000 o la STC 114/2006, se ha destacado la necesidad de que toda actuación administrativa que afecte derechos fundamentales esté debidamente motivada y sujeta a control judicial[48]. En este contexto, el artículo 103 de la Constitución Española, que exige que la administración actúe con pleno sometimiento a la ley y al derecho, adquiere especial relevancia[49].

Por otra parte, el análisis lingüístico automatizado ha sido objeto de críticas particularmente intensas por parte de la comunidad académica y expertos en derechos digitales. Más allá de los problemas técnicos relacionados con la fiabilidad del sistema y la opacidad algorítmica, se ha subrayado que este tipo de tecnologías puede introducir sesgos discriminatorios[50]. Como argumenta Koskenniemi, cualquier sistema de evaluación automatizada que transforme declaraciones humanas en datos requiere una contextualización cultural y lingüística que los algoritmos actuales no están en condiciones de ofrecer[51].

Autores como Shoshana Zuboff han advertido que el uso de IA en la administración, sin regulación democrática, constituye una forma de «capitalismo de vigilancia» que erosiona la autonomía individual y despoja al sujeto de su capacidad de interlocución política[52]. Esta crítica cobra especial relevancia cuando se traslada al ámbito migratorio, donde los solicitantes de asilo ya parten de una posición de vulnerabilidad estructural. ¿Estamos preparados para asumir las consecuencias de una tecnología que, en lugar de humanizar la administración, contribuye a despersonalizarla aún más?

La literatura también ha destacado que el uso de estas tecnologías puede revertir la carga de la prueba, situando al solicitante en una posición de debilidad al obligarlo a refutar los resultados de un sistema cuya lógica le es des-

47. R. Bustos Gisbert, «Administración pública y algoritmos: riesgos y garantías», *Revista Española de Derecho Administrativo*, núm. 215, 2022, pp. 113-142.

48. Tribunal Constitucional, STC 292/2000, de 30 de noviembre; STC 114/2006, de 5 de abril.

49. Constitución Española, artículo 103.1.

50. J. Castellanos Claramunt, «La influencia de la Inteligencia Artificial en la concepción tradicional de los derechos fundamentales: un nuevo paradigma tecnológico y jurídico», en W, Arellano Toledo (dir.), *Derecho, Ética e Inteligencia Artificial*, Tirant Lo Blanch, Valencia, 2023, pp. 139–176.

51. M. Koskenniemi, «The Politics of International Law», *European Journal of International Law*, vol. 1, núm. 1, 1990, pp. 4-32.

M, Koskenniemi, «The Politics of International Law – 20 Years Later», *The European Journal of International Law* Vol. 20 no. 1, 2009, p.7-19. doi: 10.1093/ejil/chp006

52. S. Zuboff, *The Age of Surveillance Capitalism, Public Affairs*, Nueva York, 2019. https://www.researchgate.net/publication/346844216_Shoshana_Zuboff_The_age_of_surveillance_capitalism_the_fight_for_a_human_future_at_the_new_frontier_of_power_New_York_Public_Affairs_2019_704_pp_ISBN_978-1-61039-569-4_hardcover_978-1-61039-270-0_eboo

conocida[53]. En este sentido, el Comité de Derechos Humanos de Naciones Unidas ha insistido en la necesidad de asegurar que cualquier procedimiento que afecte derechos fundamentales sea transparente y permita la contradicción, garantizando así el principio del debido proceso[54].

Estudios empíricos realizados por el Max Planck Institute for Social Anthropology han revelado que los resultados de estos sistemas son frecuentemente utilizados como base principal para denegar solicitudes, incluso cuando presentan tasas de error superiores al 30% en algunos dialectos[55]. Este dato refuerza la preocupación por el uso desproporcionado de tecnologías que, en lugar de servir como herramientas auxiliares, tienden a sustituir el juicio humano[56].

Desde una perspectiva teórica, David Lyon propone que estas prácticas no sólo afectan la privacidad, sino que transforman la relación entre ciudadano y Estado, instaurando una forma de «gobernanza algorítmica» que socava los principios democráticos tradicionales[57]. Esta observación es clave para comprender que el problema no reside únicamente en el funcionamiento técnico de los sistemas, sino en el modelo de gobernanza que introducen.

En este contexto, se ha propuesto la necesidad de establecer evaluaciones de impacto en derechos fundamentales para toda tecnología aplicada al ámbito migratorio. Autores como Cathryn Costello y Petra Molnar han planteado la exigencia de una moratoria sobre el uso de IA en procedimientos de asilo hasta que se disponga de un marco normativo robusto, transparente y sujeto a supervisión judicial efectiva[58]. También se ha sugerido la creación de órganos independientes de auditoría algorítmica, con competencias sancionadoras, inspirados en modelos como el de la Agencia Nacional para la Supervisión de la IA en Canadá. Además, estudios recientes han evidenciado que las herramientas de reconocimiento lingüístico tienden a reproducir jerarquías coloniales en la clasificación de acentos y dialectos[59], lo que genera una exclusión epistémica de los sujetos migrantes, tal como ha sido denunciado por autores como E. Sanjurjo y M. Bhat[60]. La tecnología no es neutra: refleja estructuras de poder, sesgos históricos y lógicas de exclusión que deben ser

53. F.J. Garrido Carrillo y V. Faggiani, «Los límites de la inteligencia artificial en la evaluación y tratamiento de las solicitudes de asilo y refugio», en *La necesaria reconfiguración del derecho de asilo*, Editorial Aranzadi, 2023, pp. 273–311.

54. Comité de Derechos Humanos de Naciones Unidas, Observación General núm. 32, CCPR/C/GC/32, 2007.

55. Max Planck Institute for Social Anthropology, Automated Decision-Making and Asylum Procedures, Informe de investigación, Halle/Saale, 2022.

56. N. Enache y M.C. González Rabanal, *Relaciones laborales e impacto económico de los trabajadores inmigrantes rumanos en España*, 2022.

57. D. Lyon., «Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination», *Polity Press*, Cambridge, 2003.

58. C. Costello y P. Molnar, «The Future of Refugee Status Determination: AI and Human Rights», *Refugee Survey Quarterly*, vol. 40, núm. 2, 2021, pp. 129-150.

59. Y. Taki, «No signal. Desconexión e hiperconexión: la discriminación algorítmica en la era digital», *El Pájaro de Benín*, nº 8, 2022, pp. 135–145

60. E. Sanjurjo y M. Bhat, «Language, Power and Technology in Migration Governance», *Migration Studies*, vol. 9, núm. 4, 2021, pp. 831-849.

interrogadas críticamente. Por ello, es imprescindible situar este debate en el marco más amplio de la digitalización de la administración pública, que si bien ofrece oportunidades para la mejora de la eficiencia, no puede desarrollarse a costa de erosionar derechos fundamentales[61]. Como recuerda el Informe Anual de la Agencia de los Derechos Fundamentales de la Unión Europea (FRA), la tecnología no es neutra, y su implementación debe regirse por principios de legalidad, necesidad, proporcionalidad y supervisión independiente[62]. La construcción de una administración digital democrática requiere no solo límites técnicos, sino también una visión normativamente robusta del papel del Estado y los derechos de las personas[63].

4. Jurisprudencia relevante

El debate jurídico en torno al uso de tecnologías invasivas en los procedimientos de asilo no puede comprenderse sin atender a la evolución jurisprudencial, especialmente en el ámbito alemán, donde los tribunales han sentado criterios fundamentales sobre la proporcionalidad y la necesidad de tales medidas.

Un hito importante se encuentra en la sentencia del Tribunal Administrativo de Berlín (VG 9 K 135/20 A, de 1 de junio de 2021), que resolvió el caso de una solicitante de asilo afgana[64]. En dicha resolución, el tribunal declaró ilegítimo el acceso por parte de la Oficina Federal de Migración y Refugiados (BAMF) a los datos de su teléfono móvil, al considerar que las autoridades debieron agotar previamente medios menos intrusivos para comprobar su identidad. El tribunal enfatizó que la medida vulneraba los principios de necesidad y proporcionalidad, al haberse aplicado de forma automática sin valorar alternativas menos restrictivas. Este razonamiento constituye un precedente significativo, pues afirma que la protección de los derechos fundamentales no puede supeditarse a criterios de eficiencia administrativa, especialmente cuando se trata de personas en situación de vulnerabilidad, como los solicitantes de asilo[65].

En este fallo, además, se reconoce de manera implícita el derecho de los solicitantes a no ser sometidos a un escrutinio tecnológico indiscriminado. Tal como han señalado algunos comentaristas de la sentencia, se trata de un paso

61. J.F. López Aguilar, «Una nueva gobernanza para el siglo XXI: la arquitectura institucional europea», *Sistema,* nº 271, 2024, pp. 79–89; y F. Pulido Catalán; F. Castro-Rial Garrone; A. Jarillo Aldeanueva, «El intercambio de información entre estados de la UE y las nuevas amenazas a la seguridad: estudio particular de la Guardia Civil», 2023.

Disponible en: https://dialnet.unirioja.es/servlet/exttes?codigo=315246

62. Agencia de los Derechos Fundamentales de la Unión Europea (FRA), Fundamental Rights Report 2023, Luxemburgo, Oficina de Publicaciones de la UE.

63. L. Parra Membrilla y J. Martín Ostos, «La inteligencia artificial en los sistemas de asilo y migración: ¿avance o retroceso?», en *Inteligencia artificial y derecho*, Sevilla: Astigi, 2024, pp. 141–149.

64. Verwaltungsgericht Berlin (VG 9 K 135/20 A), sentencia de 1 de junio de 2021. Disponible en: Gesellschaft für Freiheitsrechte. https://freiheitsrechte.org/

65. L. Riemer, «Handydatenauswertung bei Asylsuchenden: Rechtsprechung und Grundrechte», *Zeitschrift für Ausländerrecht und Ausländerpolitik*, núm. 6, 2021, pp. 219-230. https://background.tagesspiegel.de/digitalisierung-und-ki/briefing/handydatenauswertung-im-asylverfahren-rechtwidrig

hacia la consolidación de un estándar más exigente de garantías en la gestión de procedimientos migratorios[66]. ¿No debería este mismo principio extenderse a todos los Estados miembros de la Unión Europea como exigencia mínima de respeto a los derechos fundamentales?

La línea iniciada por el tribunal berlinés fue posteriormente reforzada por la Sentencia del Tribunal Supremo de lo Contencioso-Administrativo alemán (BVerwG, 1 C 19.21, de 16 de febrero de 2023)[67]. En este caso, el tribunal reafirmó que las autoridades no pueden aplicar un acceso generalizado y posterior a los datos, sino que deben realizar una evaluación de proporcionalidad antes de proceder a la extracción de información de dispositivos electrónicos. Este matiz resulta crucial: si la proporcionalidad se analiza una vez realizada la medida, se incurre en una lógica de justificación ex post que vacía de contenido las garantías preventivas exigidas por el Estado de Derecho[68].

De este modo, el BVerwG confirma que la actuación administrativa debe ser estrictamente individualizada y que la carga de motivar la necesidad de la medida corresponde a la administración, no al solicitante. Esta exigencia se vincula con el principio de inversión de la carga de la prueba: no puede pedirse al solicitante que demuestre la innecesariedad de la intervención tecnológica, sino que son las autoridades quienes deben justificar con precisión su adopción.

En la doctrina española, Juan Antonio García Amado ha analizado críticamente el principio de proporcionalidad como criterio de control judicial, advirtiendo que a menudo se convierte en una fórmula retórica que legitima la decisión ya adoptada en lugar de un verdadero examen de adecuación, necesidad y ponderación de intereses. Según este autor, el riesgo radica en que la proporcionalidad se reduzca a un «comodín argumentativo» que carezca de capacidad efectiva para limitar al poder público[69]. Su reflexión resulta especialmente pertinente en este ámbito: si el control de proporcionalidad se ejerce de manera formal o superficial, se corre el peligro de vaciar de contenido las garantías que la jurisprudencia alemana busca precisamente reforzar.

Sin embargo, esta postura ha generado un interesante debate con otros autores, como el profesor Manuel Atienza, quien ha defendido la utilidad del principio de proporcionalidad en la práctica jurídica como un instrumento que, aunque susceptible de abusos, permite articular un razonamiento más transparente y racional en la ponderación de derechos[70]. Atienza sostiene que, lejos de ser un mero recurso retórico, el principio proporciona un marco metodológico

66. Tribunal Europeo de Derechos Humanos, Szabó y Vissy c. Hungría, núm. 37138/14, sentencia de 12 de enero de 2016. Disponible en: https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-223725%22]}

67. Bundesverwaltungsgericht (BVerwG 1 C 19.21), sentencia de 16 de febrero de 2023. Disponible en: Bundesverwaltungsgericht. https://www.bverwg.de/

68. D. Ehlers. Europäische Grundrechte und Grundfreiheiten, Berlin, New York: De Gruyter, 2009. https://doi.org/10.1515/9783899496543

69. J.A. García Amado, *Argumentación jurídica: Fundamentos teóricos y elementos prácticos*, Tirant lo Blanch, Valencia, 2023.

70. M. Atienza, *Filosofía del Derecho y transformación social,* Trotta. 2017. M. Atienza, «El derecho como argumentación», Ariel, Barcelona, 2006, pp. 215-238.

que obliga a los jueces a explicitar las razones de sus decisiones, aumentando así la legitimidad del razonamiento jurídico. La polémica entre ambos autores ilustra dos visiones contrapuestas: la desconfianza de García Amado frente a la potencial banalización del principio y la confianza de Atienza en su valor heurístico para guiar la deliberación judicial.

Este planteamiento contrasta con la doctrina del Tribunal de Justicia de la Unión Europea (TJUE), que ha consolidado el principio de proporcionalidad como uno de los pilares del Derecho de la Unión. En casos como Digital Rights Ireland (C-293/12 y C-594/12), el TJUE anuló la Directiva de conservación de datos por considerar que imponía una injerencia desproporcionada en los derechos fundamentales a la vida privada y a la protección de datos personales. El Tribunal ha insistido reiteradamente en que toda restricción a un derecho fundamental debe ser idónea, necesaria y proporcionada en sentido estricto, exigiendo un análisis riguroso de la medida y de sus alternativas menos restrictivas[71]. Este contraste revela un dilema interesante: mientras García Amado alerta de la banalización del principio en la práctica judicial, el TJUE lo utiliza como instrumento clave para limitar la expansión de medidas tecnológicas invasivas.

De igual modo, el Tribunal Constitucional español ha desarrollado una aplicación sistemática del principio de proporcionalidad en contextos de derechos fundamentales, especialmente en materia de intervenciones en el derecho a la intimidad y a la protección de datos. En sentencias como la STC 292/2000 o la STC 114/2006, se estableció que toda injerencia en derechos fundamentales exige un triple juicio: adecuación de la medida para alcanzar el fin legítimo, necesidad de que no existan alternativas menos lesivas, y proporcionalidad en sentido estricto mediante la ponderación entre el beneficio obtenido y el sacrificio del derecho afectado[72]. Esta metodología, más estructurada, contrasta con la crítica de García Amado, pues dota al principio de un carácter operativo que, al menos en teoría, busca evitar su uso meramente retórico.

Estas decisiones, consideradas en conjunto, suponen un claro recordatorio de que el recurso a nuevas tecnologías no puede quedar fuera del control jurisdiccional. Al contrario, los jueces están llamados a garantizar que la digitalización de los procedimientos administrativos se desarrolle en estricto respeto a los derechos fundamentales. Asimismo, la jurisprudencia alemana anticipa posibles conflictos de armonización con el marco europeo, ya que el Reglamento de Eurodac y el paquete de fronteras inteligentes avanzan en una dirección contraria, que tiende a normalizar la extracción sistemática de datos electrónicos[73]. En consecuencia, esta jurisprudencia no sólo tiene un valor interno en

71. TJUE, Digital Rights Ireland (C-293/12 y C-594/12), sentencia de 8 de abril de 2014. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex:62012CJ0293

72. Tribunal Constitucional, STC 292/2000, de 30 de noviembre. https://hj.tribunalconstitucional.es/es-ES/Resolucion/Show/4276; STC 114/2006, de 5 de abril https://www.boe.es/diario_boe/txt.php?id=BOE-T-2006-8144.

73. Tribunal Constitucional, STC 292/2000, de 30 de noviembre. https://hj.tribunalconstitucional.es/es-ES/Resolucion/Show/4276; STC 114/2006, de 5 de abril https://www.boe.es/diario_boe/txt.php?id=BOE-T-2006-8144.

Alemania, sino que plantea interrogantes de alcance más amplio: ¿será capaz la Unión Europea de integrar estos estándares en su legislación común sobre asilo y fronteras, o persistirá la tensión entre eficiencia administrativa y protección de derechos?[74] La respuesta a esta cuestión será decisiva para el futuro del sistema común europeo de asilo[75].

La jurisprudencia alemana, española y europea demuestra que el principio de proporcionalidad se ha convertido en la piedra angular del control de las tecnologías aplicadas a los procedimientos de asilo. No obstante, el modo en que se aplica varía significativamente: mientras Alemania exige una evaluación preventiva y concreta de cada caso, el TJUE utiliza la proporcionalidad como mecanismo de control estructural de la legislación europea, y el Tribunal Constitucional español desarrolla un método tripartito que, aunque sistemático, corre el riesgo de caer en la formalidad. Estas diferencias plantean una pregunta crucial: ¿podrá alcanzarse un estándar común europeo que combine eficacia administrativa y respeto a los derechos fundamentales, o persistirá la fragmentación entre los distintos sistemas jurisdiccionales? Este dilema será central para los debates futuros sobre la legitimidad del uso de tecnologías en el ámbito migratorio.

Cuadro comparativo: aplicación del principio de proporcionalidad

Tribunal	Caso relevante	Enfoque sobre proporcionalidad	Observación crítica
Alemania (VG y BVerwG)	VG 9 K 135/20 A (2021); BVerwG 1 C 19.21 (2023)	Exige que el análisis de proporcionalidad se realice antes de acceder a los datos; énfasis en necesidad y alternativas menos intrusivas.	Refuerza garantías preventivas, evita justificaciones ex post.
España (TC)	STC 292/2000; STC 114/2006	Triple juicio de proporcionalidad (adecuación, necesidad, proporcionalidad en sentido estricto).	Modelo estructurado, pero riesgo de formalismo señalado por García Amado.
Unión Europea (TJUE)	*Digital Rights Ireland* (C-293/12 y C-594/12)	Proporcionalidad como criterio central para anular medidas invasivas; análisis estricto de alternativas.	Instrumento eficaz de control frente a legislaciones expansivas.

Fuente: elaboración propia.

74. A. Gálvez del Valle, «Inmigración, derechos humanos y modelo europeo de fronteras», *Revista de Estudios Jurídicos y Criminológicos*, núm. 2, 2020, pp. 145-210. https://doi.org/10.25267/REJUCRIM.2020.i2.07

75. S. Peers, *EU Justice and Home Affairs Law*, Oxford University Press, Oxford, 2021, p. 412.

IV. CONCLUSIONES

El examen realizado permite constatar que la irrupción de tecnologías de inteligencia artificial y big data en los procedimientos de asilo plantea tensiones profundas entre la eficiencia administrativa y la protección de los derechos fundamentales. Los casos analizados en Alemania, junto con la doctrina y la jurisprudencia española y europea, evidencian que el principio de proporcionalidad se ha convertido en el eje de control, aunque no está exento de polémicas en la doctrina, como muestra la controversia entre Robert Alexy y Jürgen Habermas en Alemania o el debate, en nuestro país, entre García Amado y Manuel Atienza.

Un paralelismo actual resulta inevitable: lo que sucede hoy en el ámbito del asilo refleja un fenómeno más amplio de «gobernanza algorítmica» que está permeando otros sectores clave de la vida social y jurídica. En la justicia penal, por ejemplo, los sistemas de *predictive policing* han suscitado objeciones semejantes en torno a sesgos, opacidad y proporcionalidad. En el campo laboral, los algoritmos de selección de personal reproducen discriminaciones estructurales que recuerdan a las alertadas en el análisis lingüístico automatizado de solicitantes de asilo. Y en la gestión administrativa general, la automatización de decisiones reproduce la misma lógica: una tensión entre eficacia técnica y garantías jurídicas, que cuestiona los fundamentos del Estado de Derecho.

Este paralelismo no solo confirma la transversalidad de los problemas, sino que permite advertir un riesgo: lo que hoy se experimenta con solicitantes de asilo —personas en situación de especial vulnerabilidad— puede convertirse en laboratorio de prácticas que luego se extienden al conjunto de la ciudadanía. Como subraya Cotino Hueso, el reto consiste en que la digitalización no erosione el núcleo de los derechos fundamentales, sino que se desarrolle bajo parámetros claros de legalidad, necesidad y proporcionalidad.

A este respecto, conviene recordar las políticas de endurecimiento migratorio impulsadas por la administración Trump en Estados Unidos, que incluyeron prácticas de separación familiar en la frontera, detenciones prolongadas y un incremento del uso de tecnologías de vigilancia. Dichas políticas evidenciaron cómo el control migratorio puede convertirse en un espacio de experimentación de medidas excepcionales que luego impactan en el marco general de derechos. El paralelismo con la situación europea resulta inquietante: ¿podría la utilización de tecnologías algorítmicas en fronteras derivar en una forma de persecución institucionalizada que erosione los principios del Estado de Derecho, tal como ocurrió con el endurecimiento de las políticas migratorias en EE. UU.?

El papel de los tribunales, tanto nacionales como europeos, es decisivo para trazar los límites. Alemania, España y el TJUE han puesto de manifiesto que la proporcionalidad, bien entendida, es un mecanismo capaz de frenar los excesos. Sin embargo, el riesgo señalado por García Amado de que se convierta en un recurso retórico sigue latente. De ahí que resulte imprescindible reforzar la transparencia, la motivación judicial y la existencia de contrapesos efectivos.

El paralelismo entre el asilo y otros ámbitos de aplicación de la IA obliga a preguntarse: ¿estamos ante un modelo de administración digital que se configura como herramienta de emancipación o de control? La respuesta dependerá de la capacidad de los Estados y de la Unión Europea para garantizar que la innovación tecnológica no se imponga sobre los derechos, sino que los consolide. En última instancia, la cuestión remite a un dilema político y jurídico central: decidir si queremos una inteligencia artificial al servicio de la dignidad humana o una dignidad humana subordinada a la inteligencia artificial.

BIBLIOGRAFÍA

Alamillo Domingo, I., «La regulación de la tecnología: la superación del modelo papel como elemento de transformación digital innovadora», en I. Martín Delgado (dir.), *La reforma de la Administración electrónica: una oportunidad para la innovación desde el derecho*, Instituto Nacional de Administración Pública, Madrid, 2017.

Alonso Tomé, S., «La aplicación de la inteligencia artificial en los controles de las fronteras exteriores de la Unión Europea: Regulación y desafíos», *Revista de Estudios Europeos*, nº 85, Universidad de Valladolid, 2025.

Amnistía Internacional, «UE-Túnez: acuerdo migratorio», disponible en: https://www.es.amnesty.org/en-que-estamos/noticias/noticia/articulo/ue-tunez-acuerdo-migracion/

Atienza, M., *El derecho como argumentación*, Ariel, Barcelona, 2006.

Atienza, M., *Filosofía del Derecho y transformación social*, Trotta, Madrid, 2017.

Avello Martínez, M., «EU Borders and Potential Conflicts Between New Technologies and Human Rights», *Peace & Security – Paix et Sécurité Internationales*, nº 11, Universidad de Cádiz, Cádiz, 2023.

BAMF, *Agenda Digitalisierung*, Berlín, 2022.

Bundesverwaltungsgericht (BVerwG 1 C 19.21), sentencia de 16 de febrero de 2023, disponible en: https://www.bverwg.de/

Castellanos Claramunt, J., «Sobre los desafíos constitucionales ante el avance de la Inteligencia Artificial. Una perspectiva nacional y comparada», *Revista de Derecho Político*, nº 118, 2023, pp. 261-287.

Castellanos Claramunt, J., «Una reflexión acerca de la influencia de la inteligencia artificial en los derechos fundamentales», en F. Ramón Fernández (dir.), Ciencia de datos y perspectivas de la inteligencia artificial, Tirant lo Blanch, Valencia, 2024, pp. 271-300

Castellanos Claramunt, J., «La influencia de la Inteligencia Artificial en la concepción tradicional de los derechos fundamentales: un nuevo paradigma tecnológico y jurídico», en W. Arellano Toledo (dir.), *Derecho, Ética e Inteligencia Artificial*, Tirant lo Blanch, Valencia, 2023, pp. 139-176.

Comisión Europea, «Agencia Europea de la Guardia de Fronteras y Costas (Frontex)», disponible en: https://european-union.europa.eu/

Comisión Europea, «Agencia de la Unión Europea para la gestión operativa de sistemas informáticos a gran escala en el espacio de libertad, seguridad y justicia (eu-LISA)», disponible en: https://european-union.europa.eu/

Comisión Europea, «Sistema Europeo de Información y Autorización de Viajes (ETIAS)», disponible en: https://travel-europe.europa.eu/etias_en

Comisión Europea, «Sistema de Entradas y Salidas (EES)», disponible en: https://travel-europe.europa.eu/ees/

CORDERO, J., «Proyecto europeo iBorderCtrl (CORDIS)», disponible en: https://cordis.europa.eu/project/id/700626

Corte Europea de Derechos Humanos, *Szabó y Vissy c. Hungría*, nº 37138/14, sentencia de 12 de enero de 2016.

COSTELLO, C., y Molnar, P., «The Future of Refugee Status Determination: AI and Human Rights», *Refugee Survey Quarterly*, vol. 40, nº 2, 2021, pp. 129-150.

COTINO HUESO, L., «Big data e inteligencia artificial. Una aproximación a su tratamiento jurídico desde los derechos fundamentales», *Dilemata*, nº 24, 2017, pp. 131-150.

COTINO HUESO, L., *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, Thomson Reuters Aranzadi, Cizur Menor, 2022.

COTINO HUESO, L., «Nuevo paradigma en las garantías de los derechos fundamentales y una nueva protección de datos frente al impacto social y colectivo de la inteligencia artificial», en *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, Thomson Reuters Aranzadi, Cizur Menor, 2022, pp. 69-105.

CRUZ ÁNGELES, J., y VESTRI, G., «La nueva 'ley' de inteligencia artificial: una aliada necesaria para gestionar controles fronterizos, de asilo y migratorios en la Unión Europea», en *La disrupción tecnológica en la Administración Pública: retos y desafíos de la inteligencia artificial*, Aranzadi, Cizur Menor, 2022, pp. 97-121.

DEUTSCHE WELLE, «Flüchtling aus Syrien falsch erkannt», 2023.

Directiva 2013/32/UE del Parlamento Europeo y del Consejo, de 26 de junio de 2013, sobre procedimientos comunes para la concesión o la retirada de la protección internacional.

EHLERS, D., *Europäische Grundrechte und Grundfreiheiten*, De Gruyter, Berlín/Nueva York, 2009.

ENACHE, N., y GONZÁLEZ RABANAL, M.C., *Relaciones laborales e impacto económico de los trabajadores inmigrantes rumanos en España*, 2022.

European Commission, *Ethics Guidelines for Trustworthy AI*, Publications Office, Luxemburgo, 2019.

European Digital Rights (EDRi), *Automated decision-making in migration*, Bruselas, 2022.

European Union Agency for Asylum (EUAA), *Digital strategy*, 2022.

European Union Agency for Asylum (EUAA), *Practical guide on evidence assessment*, 2023.

FRA (Agencia de los Derechos Fundamentales de la Unión Europea), *Fundamental Rights Report 2023*, Luxemburgo, Oficina de Publicaciones de la UE.

Gálvez del Valle, A., «Inmigración, derechos humanos y modelo europeo de fronteras: Propuestas conceptuales», *Revista de Estudios Jurídicos y Criminológicos*, nº 2, 2020, pp. 145-210.

García Amado, J.A., *Argumentación jurídica: fundamentos teóricos y elementos prácticos*, Tirant lo Blanch, Valencia, 2023.

Garrido Carrillo, F.J., y Faggiani, V., «Los límites de la inteligencia artificial en la evaluación y tratamiento de las solicitudes de asilo y refugio», en *La necesaria reconfiguración del derecho de asilo*, Aranzadi, Cizur Menor, 2023, pp. 273-311.

Gesellschaft für Freiheitsrechte (GFF), *Analyseverfahren des BAMF*, Informe, Berlín, 2020.

Gesellschaft für Freiheitsrechte (GFF), *Digitalisierte Asylverfahren auf dem Prüfstand*, Informe, Berlín, 2021.

Koskenniemi, M., «The Politics of International Law», *European Journal of International Law*, vol. 1, nº 1, 1990, pp. 4-32.

Koskenniemi, M., «The Politics of International Law – 20 Years Later», *European Journal of International Law*, vol. 20, nº 1, 2009, pp. 7-19.

Lenaerts, K., «La protección judicial efectiva en el Derecho de la Unión Europea», *Revista de Derecho Comunitario Europeo*, nº 52, 2015, pp. 45-79.

López Aguilar, J.F., «Una nueva gobernanza para el siglo XXI: la arquitectura institucional europea», *Sistema: revista de ciencias sociales*, nº 271, 2024, pp. 79-89.

Lyon, D., *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination*, Polity Press, Cambridge, 2003.

Marx, W., «Zur Verfassungsmäßigkeit automatisierter Sprachanalyse», *Asylmagazin*, 2021.

Max Planck Institute for Social Anthropology, *Automated Decision-Making and Asylum Procedures*, Informe de investigación, Halle/Saale, 2022.

Palma Ortigosa, A., *Decisiones automatizadas y protección de datos: especial atención a los sistemas de inteligencia artificial*, Dykinson, Madrid, 2022.

Parra Membrilla, L., y Martín Ostos, J., «La inteligencia artificial en los sistemas de asilo y migración: ¿avance o retroceso?», en *Inteligencia artificial y derecho*, Astigi, Sevilla, 2024, pp. 141-149.

Peers, S., *EU Justice and Home Affairs Law*, Oxford University Press, Oxford, 2021.

Pulido Catalán, F., Castro-Rial Garrone, F., y Jarillo Aldeanueva, A., «El intercambio de información entre estados de la UE y las nuevas amenazas a la seguridad: estudio particular de la Guardia Civil», 2023, disponible en: https://dialnet.unirioja.es/servlet/exttes?codigo=315246

Romano, A., «Credibilidad y derecho: el rol de la tecnología en la evaluación de las solicitudes de protección internacional. El caso de la ley de asilo

alemana», *IDP: Revista de Internet, Derecho y Política*, nº 39, 2023, pp. 1-12.

ROMANO, A., «Derechos fundamentales e inteligencia artificial emocional en iBorderCtrl: retos de la automatización en el ámbito migratorio», *Revista Catalana de Dret Públic*, nº 66, 2023, pp. 237-252.

SANJURJO, E., y Bhat, M., «Language, Power and Technology in Migration Governance», *Migration Studies*, vol. 9, nº 4, 2021, pp. 831-849.

TAKI, Y., «No signal. Desconexión e hiperconexión: la discriminación algorítmica en la era digital», *El Pájaro de Benín*, nº 8, 2022, pp. 135-145.

TEDH, *Szabó y Vissy c. Hungría*, núm. 37138/14, sentencia de 12 de enero de 2016.

TJUE, *Digital Rights Ireland* (C-293/12 y C-594/12), sentencia de 8 de abril de 2014.

Tribunal Constitucional, STC 292/2000, de 30 de noviembre; STC 114/2006, de 5 de abril.

Tribunal Constitucional Federal Alemán, *BVerfGE 65, 1 – Volkszählungsurteil*, 1983.

VENTURA FRANCH, A., «Inteligencia artificial, administración y garantías», *Revista Aragonesa de Administración Pública*, nº 57, 2021, pp. 11-38.

Verwaltungsgericht Berlin (VG 9 K 135/20 A), sentencia de 1 de junio de 2021, disponible en: Gesellschaft für Freiheitsrechte, https://freiheitsrechte.org/

ZUBOFF, S., *The Age of Surveillance Capitalism*, Public Affairs, Nueva York, 2019.

HACIA UNA TRANSFORMACIÓN POST-BUROCRÁTICA EN EL USO RESPONSABLE DE LA INTELIGENCIA ARTIFICIAL EN LA ADMINISTRACIÓN PÚBLICA EUROPEA: DESAFÍOS, RIESGOS Y NUEVOS HORIZONTES PARA UNA GOBERNANZA ÉTICA DE LA IA[1]

María Teresa García-Berrio Hernández
Profesora Titular de Filosofía del Derecho
Universidad Complutense de Madrid

SUMARIO: I. HACIA UN CAMBIO DE PARADIGMA EN LA ADMINISTRACIÓN PÚBLICA. II. DE LA BUROCRACIA CLÁSICA A LA NUEVA GESTIÓN PÚBLICA. III. EL IDEAL POST-BUROCRÁTICO: VALOR PÚBLICO Y SATISFACCIÓN CIUDADANA. IV. IA Y TRANSFORMACIÓN DEL SECTOR PÚBLICO EUROPEO: OPORTUNIDADES, RETOS Y EXPERIENCIAS. V. POR UNA ADMINISTRACIÓN PÚBLICA INTEROPERABLE: LA IA COMO MOTOR Y DESAFÍO DE UNA GOBERNANZA RESPONSABLE PARA UNA CIUDADANÍA DIGITAL. VI. HACIA UNA HERMENÉUTICA CRÍTICA PARA LA VIABILIDAD DE UNA GOBERNANZA HUMANA EN EL USO DE UNA IA CONFIABLE. VII.CONCLUSIONES. *BIBLIOGRAFÍA*.

1. Esta investigación se ha realizado en el marco del proyecto de I+D+i *Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas* [2023-2026] (PID2022-136439OB-I00), financiado por MCIN/AEI/10.13039/501100011033/ y «FEDER Una manera de hacer Europa».

I. HACIA UN CAMBIO DE PARADIGMA EN LA ADMINISTRACIÓN PÚBLICA

La revolución asociada a las Nuevas Tecnologías disruptivas como la Inteligencia Artificial (en lo sucesivo IA) está redefiniendo los procesos administrativos y la relación entre los ciudadanos y las instituciones gubernamentales. La IA nos ofrece la quimera de unos servicios públicos digitales más ágiles y eficientes y, lo que es más importante, el ideal de una Administración pública más justa. Hoy en día se aborda constantemente en foros de expertos la pertinencia del uso de la IA en la modernización y optimización de los procesos administrativos públicos y se suscitan constantes debates sobre cómo el Sector Público puede aprovechar las ventajas de estas tecnologías digitales para la mejora en la atención ciudadana.

En todo el mundo, pero sobre todo en el entorno de la Unión Europea, los Estados están apostando por un proceso de transformación digital de los servicios públicos a través de un «modelo post-burocrático automatizado» en el que la IA aparece cada vez mas presente en la gestión de procesos administrativos estandarizados, con el objeto de reducir así los inevitables retrasos asociados a los procedimientos administrativos tradicionales.

Este interés creciente por el fenómeno de la digitalización de las administraciones públicas obedece, en parte, a la capacidad de aprendizaje automatizado que presentan las últimas versiones de sistemas de IA Generativa, pero, en gran medida, a la posibilidad de recurrir al empleo de algoritmos para *predecir a futuro* el comportamiento de los ciudadanos[2]. Asimismo, la IA se nos ofrece como una vía de consolidación de un modelo de «Gobernanza inteligente» (*Smart Governance*) capaz de transformar aquellas estructuras administrativas que resultan problemáticas, en cuanto a su gestión y accesibilidad, y se perfila hoy como una aliada en la ardua tarea de erradicación de los marcos clientelistas de la Administración burocrática tradicional[3].

No podemos negar el papel dinamizador que puede aportar la IA en la mejora de la prestación de servicio públicos y en el incremento de la calidad de la toma de decisiones. Sin embargo, una transformación digital adecuada no sólo implica un despliegue tecnológico y cada vez son más las voces críticas que apuestan por emprender un profundo debate ético, normativo y político sobre cómo la IA puede contribuir a mejorar la gobernanza pública sin devaluar lo humano, centrándose en las necesidades de los ciudadanos y la rendición de cuentas, y garantizando siempre la alineación de la tecnología con los principios de gestión del valor público y la democracia representativa. En este contexto, cobra actualidad el vaticinio de autores como Osborne y Gaebler, quienes en la década de los 90 del siglo pasado se aventuraron a anticipar la caducidad de los modelos administrativos tradicionales y la necesidad de «reinventar el

2. J.I. Criado, «Inteligencia Artificial y Administración Pública», *Eunomía: Revista en Cultura de la Legalidad*, 2021.

3. Ibid.

Gobierno»[4] para liberarlo de los modelos procedimentales paquidérmicos[5] que hacen énfasis en la eficiencia de los medios más que en su *valor público.*

II. DE LA BUROCRACIA CLÁSICA A LA NUEVA GESTIÓN PÚBLICA

Los cimientos de la administración burocrática tradicional presentan cada vez más limitaciones —algunas de ellas irresolubles— en la coyuntura de las sociedades contemporáneas, sobre todo en su adaptación a la era digital, y todo ello conduce inexorablemente a la ineficiencia de la gestión pública. Por este motivo, con el objeto de superar el llamado *burocratismo* —entendido como la hegemonía del modelo burocrático tradicional—, la doctrina ha apostado en las últimas décadas por un nuevo tipo de *racionalidad administrativa*[6] y por una nueva cultura organizacional con nuevos enfoques de gestión y nuevas tecnologías administrativas[7].

La historia de la Administración pública occidental ha estado marcada por la hegemonía de modelo weberiano del llamado «paradigma burocrático», entendido como un ideal de organización administrativa sustanciado en la racionalidad, la legalidad y la eficiencia y orientado a estudiar organizaciones grandes y complejas —especialmente el Estado moderno—con el propósito de asegurar su funcionamiento eficiente y predecible a través de reglas y procesos formales. En efecto, Max Weber en su obra *Economía y Sociedad*[8] concibe la burocracia como la forma más racional de organización, imprescindible pues para institucionalizar el poder y la dominación legítima a través del Estado moderno.

El modelo burocrático weberiano se sustenta en cuatro pilares esenciales[9]: *(i) la jerarquización de la autoridad,* mediante el diseño de niveles bien definidos de autoridad en los que cada miembro responde ante un superior; *(ii) la impersonalidad de los procedimientos*, con el objeto de garantizar imparcialidad y transparencia en la gestión; *(iii) la estratificación* de las tareas del funcionariado y división del trabajo en base a la capacitación técnica y mérito, con el objeto de maximizar la eficiencia a través de la profesionalización y especiali-

4. D. Osborne, y T. Gaebler, *La reinvención del Gobierno*, Addison-Wesley, New York, 1992. El término «reinvención del Gobierno» consiste para estos autores en la disolución del paradigma weberiano de la administración federal en los Estados Unidos de América. Esta propuesta buscaba ser implementada como el plan de reforma del Gobierno Norteamericano por la comisión que presidió el Vicepresidente Al Gore en la administración Clinton.

5. S. Chica Vélez, «Una mirada a los nuevos enfoques de la gestión pública», *Administración & Desarrollo*, 39 (53), 2011, pp. 57-74. Cita: Chica Vélez, S., op. cit. p. 71.

6. L.C. Bresser Pereira, *Lo público no estatal en la reforma del Estado*, Ed. Paidós, Buenos Aires, 1998.

7. M. Barzelay, *El nuevo paradigma de la gestión pública*, Fondo de Cultura Económica, México, 2003, p. 17.

8. M. Weber, *Economía y sociedad*, Fondo de Cultura Económica, México, 1922.

9. M. Weber, *The Theory of Social and Economic Organization*. Free Press, Nueva York, 1947.

zación en la asignación de tareas y, por último, (iv) *el respeto debido a la formalidad de la acción administrativa*, mediante procedimientos estandarizados que buscan, a través de la rutina y la reiteración de los procedimientos, garantizar la trazabilidad de las comunicaciones administrativas y asegurar la eficiencia y previsibilidad en el funcionamiento de la organización estatal.

El paradigma weberiano del modelo burocrático constituyó durante décadas la base de legitimidad y eficiencia del un modelo ideal del Estado moderno que buscaba blindar la Administración pública frente a la arbitrariedad política y establecer un margen de previsibilidad en la toma de decisiones administrativas, contribuyendo así al desarrollo de instituciones perdurables y confiables para la ciudadanía. No obstante, como el propio Weber nos advierte a través de la paradigmática expresión de «jaula de hierro» (*Stahlhartes Gehäuse*): todo exceso de rigidez burocrática conduce inexorablemente a la subordinación de la libertad individual en aras al cumplimiento estricto de la validez formal de las normas y de las reglas procedimentales[10].

Si bien el modelo burocrático weberiano consolidó una administración formalmente imparcial y estable, también fue poco a poco decayendo en sus objetivos fundacionales debido a la lentitud, centralización y desvinculación de su sistema de gestión respecto de los resultados y expectativas ciudadanas. En efecto, la priorización de la validez formal de los procedimientos sobre la eficacia y la obsesión por el cumplimiento normativo llegaron a obstaculizar la creatividad organizativa, el servicio y la proximidad hacia la ciudadanía en la gestión de lo público.

Por todo ello, desde finales de los años setenta del siglo XX, el modelo de Administración Pública inspirada en el ideal burocrático weberiano parecía no encajar con el desarrollo de los modelos de Estado de bienestar de los principales países desarrollados; asimismo, la consecuente expansión de una Administración cada vez más descentralizada para poder garantizar la prestación de servicios públicos generalistas y el acceso a los recursos públicos (educativos, sanitarios, de seguridad ciudadana, etc…) de la generación de los *Baby*-boomers durante las décadas de los ochenta y noventa, acabaría por tensionar definitivamente dicho modelo tradicional.

En este contexto de transformación surgirían *modelos alternativos post-burocrático*s de gestión pública que incorporarían progresivamente instrumentos y lógicas de acción importados del sector privado, los cuales transitarían desde el *interés por lo públic*o a un concepto de resultados medidos desde la valoración que la propia ciudadanía —a modo clientelar— asigna a su propia experiencia con la Administración pública.

Desde este nuevo paradigma, el llamado *modelo post-burocrático*, la Administración se entiende ya en términos de producción, busca la calidad y la generación de valor y, como apunta J. Ignacio Criado, se avanza hacia un nuevo modelo de gestión de lo público que debe «(…) superar el enfoque de

10. S.M. Martínez Castilla, «La burocracia: elemento de dominación en la obra de Max Weber», *Revista Misión Jurídica*, núm. 10, 2016, pp. 141-154.

la eficiencia desligado de la eficacia»[11]. En la misma línea se pronuncian otros autores influyentes en el campo de estudio de la gestión pública como Mark Barzelay, quien introduce como aspecto más relevante de los modelos alternativos post-burocráticos la *rendición de cuentas*. En efecto, este autor es tajante al erradicar en su estudio titulado *El nuevo paradigma de la gestión pública*[12] incluso la palabra 'eficiencia' del léxico de la Administración Pública, argumentando para ello que el enfoque burocrático tradicional sustanciado en la eficiencia no nos permite valorar adecuadamente las actividades gubernamentales. En su lugar, Barzeley sostiene que al deliberar acerca de la naturaleza y del valor de las actividades gubernamentales, los servidores públicos deben recurrir de forma preferente a otros conceptos, como los de producto o servicio, calidad y valor.

Precisamente, es en este contexto de crisis funcional del modelo burocrático clásico donde emerge en la década de los noventa del siglo XX el modelo de la llamada *Nueva Gestión Pública* (en lo sucesivo, NGP) [también conocida por su nomenclatura técnica inglesa como New Public Management (NPM)][13]. La NGP se plantea como un modelo de sustitución de la administración burocrática que promueve la introducción de un espíritu empresarial en la cultura administrativa, entendiendo al Estado como una especie de 'gestor-mediador' entre la sociedad y el mercado. Este enfoque buscaba modernizar la forma en que se concibe y ejecuta la gestión pública, orientándola a resultados 'medibles' o cuantificables, con una clara influencia por tanto de las dinámicas propias del pensamiento económico neoclásico, el cual centra su acción en la satisfacción del cliente y en el que se valora especialmente la rentabilidad y la competitividad, apostando para ello por un nuevo tipo de *racionalidad administrativa* de corte empresarial[14].

En esta misma línea de pensamiento se pronunciarían otros autores favorables a modelos alterativos *post-burocráticos*, como Peter Aucoin[15], para quien

11. Para un recorrido avanzado por las principales corrientes que han inspirado la gestión pública contemporánea, tratando de evaluar sus dimensiones clave y las aportaciones más relevantes a las administraciones públicas contemporáneas: J.I. Criado, «Las administraciones públicas en la era del gobierno abierto. Gobernanza inteligente para un cambio de paradigma en la gestión pública», *Revista de Estudios Políticos* (173), 2016, pp. 245-275. DOI: http://dx.doi.org/10.18042/cepc/rep.173.07.

12. M. Barzelay, *El nuevo paradigma de la gestión pública*, op. cit.

13. Michael Barzelay en un segundo estudio titulado *La nueva gerencia pública. Un acercamiento a la investigación y al debate de las políticas públicas* plantearía que la expresión *Nueva Gestión Pública* hace referencia a aquel cúmulo de decisiones sobre políticas públicas de las últimas décadas que han supuesto un giro sustancial en la gestión del «sector estatal» en casos concretos como son el Reino Unido, Nueva Zelanda, Australia, Escandinavia y América del Norte. Véase al respecto: M. Barzelay, *La nueva gerencia pública. Un acercamiento a la investigación y al debate de las políticas públicas*, Fondo de Cultura Económica, México, 2005. Cita: Barzelay, M., op. cit., 16-30.

14. J. Feffer, *Nuevos rumbos en la teoría de las organizaciones*, Fondo de Cultura Económica, México, 2000.

15. P. Aucoin, *Reforma administrativa en la gestión pública: Paradigmas, principios, paradojas y péndulos*, en Brugue, Q. y Subirats, J. (comps.). *Lecturas de gestión pública*, Instituto Nacional de Administración Pública, Madrid, 1996, pp. 293- 515.

en la propia expresión NGP reside el cambio de paradigma de las políticas de la Administración pública orientadas a la reducción de costos mediante la medición cuantitativa de la eficiencia en la prestación y entrega de los servicios públicos. A juicio de este autor, entre las innovaciones centrales que aporta el modelo de la NGP destacan *(i)* la sustitución de estructuras burocráticas piramidales por modelos organizativos más horizontales a través de equipos de trabajo multidisciplinares que favorezcan la toma de decisiones y la resolución de problemas de manera más ágil; *(ii)* la adopción de mecanismos de competencia interna y externa en los que sea el tercer sector, a través de agencias públicas y mediante la privatización parcial de servicios, el que presta buena parte de servicios públicos bajo sistemas de rendición de cuentas; *(iii)* la descentralización de la propia Administración pública, lo que permite mayor cercanía con la ciudadanía, fomentando la autonomía local; *(iv)* la promoción de una gestión orientada al ciudadano, concebido como 'cliente', cuya satisfacción se convierte en el indicador principal del éxito público y, por último, *(v)* la automatización de los procesos administrativos, abandonando el ideal de la rigidez de la fórmula kelseniana de hegemonía de la validez formal procedimental por nuevos mecanismos de acceso a la Administración que priorizan la eficacia y el impacto social de las políticas públicas sobre la ciudadanía.

Si bien el propósito inicial del modelo *post-burocrático* de la NGP era el de modernizar la Administración pública —para poder superar, en definitiva, la rigidez formalista del modelo weberiano—, su alcance e impacto han propiciado una profunda descentralización organizacional y una redefinición del papel de los servidores públicos; los cuales dejan de ser meros ejecutores de normas para convertirse en gestores responsables de productos y servicios de calidad.

No obstante, no podemos eludir en nuestra exposición que esta nueva orientación ha generado críticas sustanciales entre los especialistas en el estudio de la gestión de la Administración pública. En este sentido se pronuncia Lynn, para quien la traslación acrítica y mecánica de doctrinas empresariales puede suponer una pérdida de los valores fundamentales que distinguen a la Administración pública —especialmente la igualdad y la seguridad jurídica—, dando como resultado políticas públicas sustanciadas en la discrecionalidad excesiva o incluso en la ineficiencia cuando el modelo de la NGP no se integra en una cultura de servicio público. Por su parte, Pollitt y Bouckaert advierten que la gestión flexible que promueve el modelo de la NGP puede debilitar los sistemas de control y aumentar la discrecionalidad en la Administración pública, abriendo la puerta al clientelismo y a la corrupción; alejándose pues de la profesionalización y meritocracia que deben regir el servicio público[16]. Asimismo, Du Gay enfatiza el riesgo de que la excesiva flexibilidad que aporta el modelo de la NGP, cuando no viene acompañada de una auténtica cultura de responsabilidad pública, pueda derivar en una ineficiencia congénita del Sector Público;

16. C. Pollitt & G. Bouckaert, *Public Management Reform. A Comparative Analysis: Into the Age of Austerity*, Oxford University Press, Cambridge, 2017.

en su opinión, las reformas que sugiere la NGP pueden crear estructuras demasiado maleables que carezcan de controles efectivos y transparencia[17].

En definitiva, todas estas voces disidentes y reaccionaras a asumir la conveniencia del tránsito hacia el modelo *post-burocrático* de la NGP en la Administración pública, convergen en la centralidad que otorgan a la creación de valor público y a la satisfacción ciudadana. Se abandona pues progresivamente la lógica de la eficiencia meramente administrativa —propia del modelo weberiano tradicional— y se aboga por una gestión orientada a resultados, donde el éxito administrativo no se mide sólo en términos de cumplimiento formal, sino en función del impacto social, la mejora del bienestar colectivo, la confianza ciudadana y la producción efectiva de *valor público*.

El ideal de este modelo *post-burocrático* de Administración pública reclama un nuevo tipo de relación entre ciudadanos y administraciones, superando el tradicional papel pasivo de la ciudadanía y promoviendo la corresponsabilidad y la colaboración en la definición de prioridades, la gestión de recursos y la evaluación de políticas[18]. La toma de decisiones se descentraliza y democratiza, implicando activamente a los empleados públicos de todos los niveles operativos y promoviendo una cultura institucional centrada en la misión y el servicio. En este debate destaca particularmente la figura de Mark Moore, para quien los ciudadanos deben dejar de ser considerados como usuarios-adoptantes, receptores, clientes o meros consumidores de servicios gubernamentales para pasar a convertirse en «socios activos» de modelo de colaboración entre la Administración y la ciudadanía que prioriza el *valor público*[19] y en el que se asegura la inclusividad y democratización[20].

III. EL IDEAL POST-BUROCRÁTICO: VALOR PÚBLICO Y SATISFACCIÓN CIUDADANA

El concepto de *valor público*, planteado por Mark Moore en 1995 en su tratado *Gestión estratégica y creación de valor en el sector público* trata de establecer una estructura de razonamiento práctico que suponga una guía para los gestores públicos; en este contexto, el gestor público es un actor que debe ex-

17. P. Du Gay, *In Praise of Bureaucracy: Weber, Organization, Ethics*, Sage, London, 2000.

18. L.C. Bresser Pereira, *Lo público no estatal en la reforma del Estado*, Ed. Paidós, Caracas, 1998.

19. M.H. Moore, *Creating Public Value: Strategic Management in Government*, Cambridge, MA: Harvard University Press, 1995. En su traducción al castellano: M.H. Moore, *Gestión estratégica y creación de valor en el sector público*, Ed. Paidós, Barcelona, 1998.

20. La literatura sobre administración electrónica identifica cuatro modelos paradigmáticos para conceptualizar al ciudadano como usuario de servicios públicos digitales: *(i)* un *ideal de profesionalidad*, según el cual, el ciudadano es cliente; *(ii)* un *ideal de eficiencia*, según el cual, los ciudadanos son receptores anónimos y la prioridad es la rentabilidad y la productividad administrativa; *(iii)* un *ideal de servicio*, según el cual, el ciudadano es usuario-adoptante y, por último, *(iv)* un *ideal de participación*, según el cual, los ciudadanos son co-productores y colaboradores con los organismos públicos.

plotar el potencial del contexto político y organizativo en el que está inmerso con el fin de crear *valor público*[21].

Moore propone en su estudio que la función del sector público debe ir más allá de regular y controlar para, en su lugar, generar beneficios cuantificables para la ciudadanía[22]. Esto último exige por parte de la Administración pública una gestión estratégica orientada a los resultados ya que, a juicio de este autor, el llamado *valor público* debe ser más que un mero enfoque de impactos o valoraciones monetarias y debe incluir también beneficios sociales desde la percepción de los propios ciudadanos. Con este objetivo, Moore entiende la gestión pública como una acción estratégica que se orienta a aquellos resultados especialmente demandados por la ciudadanía.

No obstante, si bien en el sector privado el éxito empresarial se mide en beneficio económico, en el sector público la tasa de éxito debe evaluarse en términos de bienestar social, satisfacción y confianza ciudadana[23]. Para Moore, el éxito de la gestión pública reside precisamente en la inserción de aquellos principios especialmente valorados por los clientes de la organización y, tal y como nos desvela con especial ironía, evaluar la satisfacción ciudadana no resulta tarea fácil, ya que «(…) son muchos los ciudadanos que escriben para quejarse sobre la prestación de un servicio, pero son muy pocos los que escriben para felicitar al consistorio, porque cada mañana al abrir su ducha, efectivamente sale un chorro de agua potable»[24].

El enfoque del *valor público* nos obliga, por tanto, a pensar en una mejora continua de los servicios públicos que se ofrecen a la ciudadanía, sustentándose para ello en tres premisas que garantizan el equilibrio entre la eficacia en los servicios, de una parte, y los principios de equidad y legitimidad, de otra; en concreto, *(i)* la promoción de la soberanía ciudadana, la cual reconoce que los valores prioritarios deben provenir de procesos democráticos; *(ii)* el blindaje de la autoridad estatal, responsable de administrar recursos colectivos en aras al beneficio común y, por último, *(iii)* un marco normativo mixto que garantice de forma correlativa tanto la eficiencia como la justicia en la gestión pública.

Los sistemas burocráticos con los que opera la Administración pública suelen ser rígidos y poco flexibles y, por este motivo, la adopción de un enfoque de *valor público* implica forzosamente llevar a cabo una transformación de las prácticas convencionales y de los procesos representativos de las democracias liberales, ya que requiere el compromiso por parte de la ciudadanía en su conjunto. Como aprecia Moore, «(…) no basta con decir que los gestores públicos crean resultados valiosos, sino que deben de ser capaces de demostrar que los resultados obtenidos se pueden comparar tanto con el consumo privado como con la libertad del gestor a la hora de producir los resultados deseados»[25].

---

21. M.H. Moore, *Creating Public Value: Strategic Management in Government*, Cambridge, MA: Harvard University Press, 1995. Cita: Moore, op. cit, pp. 18-19.
22. Moore, op. cit. supra., p.19
23. Ibid., pp. 18-19.
24. Ibid., p. 64.
25. Ibid., p. 61.

Históricamente, la evaluación de las políticas públicas ha sido todo un rompecabezas. En este sentido, la tecnología, en general, se revela esencial para una gestión estratégica y eficaz en el Sector Público al posibilitar la automatización de procesos y la mejora continua en la prestación de servicios. Efectivamente, gracias a herramientas digitales —tales como portales de transparencia y sistemas de evaluación electrónica— se ha facilitado la implementación progresiva de mecanismos de rendición de cuentas y transparencia que permiten a los ciudadanos evaluar el impacto de las políticas públicas. Por ejemplo, en los últimos años son numerosas las entidades públicas que han implantado sistemas digitales de monitoreo para supervisar el cumplimiento de sus objetivos de *valor público*. Estas plataformas permiten recibir retroalimentación ciudadana de manera constante, lo que contribuye a la mejora de la calidad de los servicios públicos y a la adaptación de las políticas públicas a las necesidades reales de la ciudadanía. Los portales de transparencia de las Administraciones públicas y las plataformas de participación ciudadana han mostrado ser especialmente efectivos en la promoción del valor público, al facilitar también el fomento de una cultura de confianza y colaboración entre la ciudadanía y sus instituciones. La digitalización, asimismo, permite que los ciudadanos accedan a información en tiempo real, aspecto crucial para asegurar la transparencia en la gestión pública.

IV. IA Y TRANSFORMACIÓN DEL SECTOR PÚBLICO EUROPEO: OPORTUNIDADES, RETOS Y EXPERIENCIAS

La IA ha irrumpido en el Sector Público como una herramienta susceptible de transformar la gestión administrativa hacia un modelo *post-burocrático* de *valor público* que haga frente a la creciente complejidad de las sociedades contemporáneas y a la pluralidad de intereses en comunidades cada vez más globalizadas. Ante la imposibilidad de gestionar de forma racional la ingente demanda ciudadana de eficiencia en la Administración pública, la implementación de sistemas de automatización de los procesos administrativos mediante algoritmos y sistemas de decisión automatizada (en lo sucesivo, ADS) hace que crezcan las expectativas en la ciudadanía de una mayor objetividad, transparencia y racionalidad en la gestión pública[26]. En este contexto, los sistemas de IA permiten procesar grandes volúmenes de datos, identificar patrones y anticipar problemas, abriendo nuevas posibilidades para la intervención pública y la regulación eficiente[27]. Es más, lejos de ser una mera herramienta, la IA se revela —como a continuación analizaremos—como un auténtico *motor de gobernanza*

26. T.Z. Zarsky, «The Trouble with Algorithmic Decisions: An Analytic Road Map to the Legal Debate», *Science, Technology, & Human Values*, 41(1), 2016, pp. 118-132.

27. T. Gillespie, «The Relevance of Algorithms», in Gillespie, T. Boczkowski, P.J. & Foot, K.A. (Eds.), *Media Technologies: Essays on Communication, Materiality and Society*, MIT Press, Massachussets, 2014, pp. 167-194.

ética para la optimización de la gestión de servicios ante los complejos desafíos derivados de la era digital.

Es evidente que la implementación de la IA en el Sector Público ofrece múltiples beneficios al permitir optimizar tareas repetitivas (tales como el procesamiento de solicitudes, la gestión documental, el análisis de datos, etc..), al liberar recursos humanos para funciones de mayor valor añadido y al mejorar la prestación de servicios en base a las necesidades ciudadanas, mediante la personalización y canalización de información en función del perfil de cada usuario. Por último, al fomentar la transparencia y la participación ciudadana mediante plataformas digitales que habiliten a la ciudadanía canales innovadores para evaluar el desempeño de la Administración pública en tiempo real, los sistemas de IA se convierten en auténticos aliados para la transparencia y el acceso a la información pública de manera eficiente; lo que contribuye significativamente a la rendición de cuentas y la participación ciudadana en la gestión pública. Adicionalmente, la IA también nos ofrece la posibilidad de minimizar la influencia de factores subjetivos o discriminatorios en los procesos de decisión administrativos; en efecto, el empleo de algoritmos basados en criterios predefinidos puede favorecer la toma de decisiones más imparciales y transparentes por parte de la Administración Pública y, de esta forma, se minimiza el impacto de un posible error humano o del recurso a sesgos discriminatorios en la toma de decisiones, lo que promueve una mayor equidad y objetividad en la consecución por parte de la Administración pública del interés general.

En este sentido, Europa se ha posicionado en los últimos años como líder en la regulación y adopción responsable de la IA en el Sector Público. Entre las buenas prácticas adoptadas por las Administraciones públicas europeas en los últimos años destacan especialmente la implementación de sistemas de IA para la gestión de trámites administrativos, la predicción de demanda de servicios sociales y la optimización de recursos en salud y educación[28].

En efecto, la IA ha irrumpido en el Sector Público europeo como un instrumento esencial para gestionar información, agilizar procesos, personalizar servicios y, como analizaremos con más detalle en el siguiente epígrafe, para mejorar la llamada *interoperabilidad* entre instituciones de diferentes países europeos. Entre sus aplicaciones más relevantes destacan: *(i)* la automatización de procesos administrativos, reduciendo cargas de trabajo repetitivas y aumentando la eficiencia operativa; *(ii)* el análisis predictivo y la anticipación de necesidades sociales en la ciudadanía, permitiendo intervenciones preventivas basadas en el análisis de grandes volúmenes de datos; *(iii)* la gestión inteligente de servicios urbanos, como la movilidad, la seguridad y el medioambiente; y *(iv)* la mejora de la interoperabilidad institucional, facilitando el intercambio ágil y seguro de información entre organismos públicos de una misma Administración o de varias.

28. H. Margetts & C. Dorobantu, «Rethinking Public Policy in the Era of AI: Lessons from the United Kingdom», *Nature Communications*, 11, 2020, pp. 1-3.

En los últimos años se han sucedido proyectos pioneros en España, Dinamarca, Estonia y los Países Bajos. En el caso de Estonia, el 99% de los servicios públicos están digitalizados y disponibles en línea, permitiendo a los ciudadanos realizar casi cualquier trámite sin acudir a oficinas físicas; asimismo, se utilizan sistemas de IA para resolver juicios menores y para la transcripción automática de juicios y sesiones parlamentarias[29]. Por su parte, en el marco de los países escandinavos, Dinamarca ha desarrollado plataformas digitales que integran la IA para facilitar la interacción ciudadana y mejorar la eficiencia gubernamental, así como Suecia ha integrado sistemas de algoritmos predictivos para la monitorización de bosques, cosechas, nieve, hielo e inundaciones. En el caso de Países Bajos, relevantes esfuerzos se han llevado a cabo en la implementación de algoritmos predictivos para la detección del fraude fiscal, la asignación de subsidios o ayudas públicas y la gestión inteligente del tráfico urbano. Por su parte, en el caso de España, el uso de sistemas de chatbots y plataformas de participación ciudadana se ha generalizado en las administraciones locales con el objeto de analizar en tiempo real opiniones ciudadanas recogidas desde plataformas digitales, proporcionando respuestas rápidas y precisas a consultas de carácter cotidiano, lo que permite mejorar de esta forma la experiencia del ciudadano y reducir los tiempos de espera en la atención personal de los vecinos. A modo de ejemplo, podemos reseñar una reciente iniciativa del Ayuntamiento de Valencia adoptada tras la reciente tragedia sufrida en la región a raíz de las inundaciones catastróficas de 2024, de un prototipo de «gemelo digital urbano» que combina herramientas de IA y tecnologías de simulación para anticiparse a las emergencias y mejorar la gestión y la planificación en el término municipal del desarrollo de infraestructuras críticas, lo que permitirá que la ciudadanía pueda evaluar políticas públicas antes de su implementación efectiva. Otro ejemplo destacado del uso responsable de la IA en el Sector Público español es la herramienta conocida como *Administració Oberta de Catalunya* (AOC) que automatiza la verificación de datos personales de ciudadanos para identificar vulnerabilidad energética, asegurando de este modo tanto la interoperabilidad entre administraciones públicas como la debida protección social ante casos de vulnerabilidad y pobreza energética.

Todos estos ejemplos ponen de manifiesto cómo la IA está transformando la Administración pública —de forma más inmediata en el ámbito local—a través de la absorción progresiva de tareas que van desde la automatización de trámites y la mejora de la eficiencia, hasta la detección de fraude, la participación ciudadana y la innovación en la gestión de servicios públicos. Estas experiencias europeas nos ofrecen valiosas lecciones sobre la necesidad de abrirnos a una Administración pública algorítmica. No obstante, todas ellas ponen también de manifiesto la necesidad de abordar de forma efectiva los riesgos asociados a la privacidad, la discriminación y la opacidad que plantea el uso de sistemas algorítmicos y de IA susceptibles de reproducir sesgos estructurales.

29. P. Kettunen & J. Kallio, «Digital Government in the Nordic Countries: From e-Government to Smart Government», *Government Information Quarterly*, 38(1), 2021, pp. 101.

En efecto, el despliegue, fomento y generalización progresiva del uso de herramientas de IA en la Administración pública europea se encuentra marcado por importantes limitaciones técnicas, organizativas, éticas y legales; entre las que destacan: *(i)* la necesidad de que los sistemas algorítmicos empleados en el Sector Público sean comprensibles y auditables tanto por los propios ciudadanos como por los supervisores de servicios y gestores; *(ii)* el cumplimiento estricto del Reglamento General de Protección de Datos (RGPD) para garantizar la anonimización de datos sensibles y la garantía de que la automatización no vulnere la privacidad de los ciudadanos; *(iii)* el establecimiento de mecanismos claros de rendición de cuentas que permitan determinar la responsabilidad en caso de error, sesgo o daño producido por sistemas de IA empleados en la gestión pública; y *(iv)* la mitigación de sesgos discriminatorios que pudieran coartar o impedir el acceso igualitario de toda la ciudadanía a servicios públicos automatizados, evitando así que las herramientas de IA empleadas por la Administración Pública permitan reproducir desigualdades sociales.

Efectivamente, la complejidad de los modelos actuales de IA dificulta la comprensión de sus resultados, por lo que es fundamental garantizar la transparencia en su funcionamiento, incluyendo instrucciones claras que permitan a los responsables el uso informado y apropiado de los sistemas. Además, se deben desarrollar protocolos para la responsabilidad ética, especialmente ante decisiones erróneas o resultados sesgados, junto a una supervisión humana continua de los sistemas de IA que presente algún tipo de riesgo para los usuarios del Sector Público. En efecto, la protección de la privacidad y los datos personales constituye otro desafío de primer orden debido al gran volumen de datos que gestiona la Administración pública a diario, y esto último exige el establecimiento de mecanismos robustos que salvaguarden los derechos fundamentales de la ciudadanía. Por otra parte, la capacitación técnica del funcionariado y personal de las Administraciones públicas europeas —el proceso conocido como *alfabetización digital*—cobra especial relevancia en este contexto a la hora de afrontar con éxito la entrada en vigor de la normativa comunitaria pionera en materia de IA y servicios digitales[30].

Como se ha expuesto, la integración de la Inteligencia Artificial en el Sector Público europeo suscita significativos interrogantes respecto a la compatibilidad de los sistemas de IA con los principios democráticos y con un modelo ético de

30. La Comisión Europea ha propuesto a los Estados miembros la adaptación de marcos de acción para desarrollar competencias técnicas, éticas y organizativas en IA entre los empleados públicos. Algunas de estas estrategias de desarrollo de competencias técnicas a destacar serían: *(i)* la formación continuada en ética y competencias digitales, dirigida a todos los niveles de la Administración pública*; (ii)* la elaboración de códigos de conducta y decálogos de principios para el uso ético de la IA; *(iii)* la supervisión permanente y evaluación de impacto del empleo de sistemas de IA, incluyendo auditorías internas y externas; *(iv)* el fomento de la participación ciudadana en el diseño y revisión de sistemas algorítmicos y, por último*, (v)* la colaboración entre administraciones europeas.

gobernanza pública[31]. El panorama regulatorio de la IA en la UE avanza inexorablemente, bajo el impulso del Reglamento europeo en materia de IA, hacia un enfoque ético y fiable. En este nuevo escenario programático y normativo, la IA es reconocida como una herramienta clave para *(i)* la automatización de los procesos administrativos, incrementando la eficiencia y reduciendo los errores; *(ii)* la estandarización y transformación de procedimientos administrativos heterogéneos e ineficaces; *(iii)* la personalización y adaptación de los servicios públicos a las necesidades reales de los usuarios y, lo que es más relevante, *(iv)* la anticipación de los retos vinculados al cumplimiento del requisito de *interoperabilidad de la Administración pública europea* a través del empleo de sistemas predictivos de IA.

Por tanto, la implementación de la IA en el ámbito de la Administración pública europea evidencia dos aspectos fundamentales a considerar: por un lado, el potencial innovador indiscutible de esta nueva tecnología, pero, por otro, la necesidad acuciante de que establezcamos un *marco ético de Gobernanza que regule su uso responsable*. Dado que la implementación de la IA tiene, en efecto, el potencial de transformar la gestión pública; debemos abordar el desarrollo de una estrategia integral en el marco europeo favorable a la *interoperabilidad de la Administración pública,* con el objeto de contemplar todos los aspectos necesarios para garantizar una implementación confiable de la IA. Por todo ello, el desarrollo de soluciones que garanticen la sostenibilidad, la escalabilidad y la interoperabilidad entre países, sectores y plataformas institucionales que operan en la UE, promoviendo así la cooperación europea, constituye un aspecto de suma relevancia que debe ser objeto de nuestro análisis en el siguiente apartado.

31. Si bien esta problemática será objeto de análisis en el epígrafe final del presente estudio, en el caso concreto del control ético de la IA en la Administración pública, debemos articular cuatro mecanismos esenciales: (i) El respecto debido a principios éticos europeos; en efecto, el desarrollo y la implantación de sistemas algorítmicos deben respetar valores como la dignidad humana, la equidad, la sostenibilidad, la supervisión humana y la explicabilidad tecnológica. (ii) La evaluación de impacto algorítmico, es decir, la UE establece la obligación de realizar evaluaciones de impacto para los sistemas de IA en funciones públicas de alto riesgo. Estas evaluaciones deben analizar efectos potenciales sobre derechos fundamentales, inclusión, no discriminación y privacidad, garantizando la participación de expertos y ciudadanía. (iii) Garantizar la transparencia y la creación de registros públicos de algoritmos empleados en la Administración que garanticen la trazabilidad, la transparencia y la rendición de cuentas sobre las decisiones automatizadas. (iv) Por último, la supervisión y el control continuo mediante la realización de auditorías internas y externas, supervisión institucional y garantías procedimentales para la monitorización continua de los sistemas de IA.

V. POR UNA ADMINISTRACIÓN PÚBLICA INTEROPERABLE: LA IA COMO MOTOR Y DESAFÍO DE UNA GOBERNANZA RESPONSABLE PARA UNA CIUDADANÍA DIGITAL

El Reglamento (UE) 2024/903 *sobre una Europa Interoperable*[32] representa un hito en la transformación digital del Sector Público europeo, especialmente por el papel que se reconoce a la Inteligencia Artificial en la *interoperabilidad* y en la evolución de los servicios públicos digitales. Para ello, el Reglamento 2024/903 establece medidas orientadas a fomentar el intercambio transfronterizo de datos, información y conocimientos mediante servicios digitales para, más adelante, lograr la eliminación de barreras legales, técnicas, semánticas y organizativas en el intercambio de datos entre las Administraciones públicas de la UE, de forma que se puedan garantizar *servicios públicos eficientes, transparentes y plenamente conectados en toda la Unión Europea*. Su finalidad, por tanto, no es otra que la coordinación de los servicios públicos digitales a nivel internacional y europeo, eliminando para ello aquellos obstáculos que dificulten la transferencia de información entre Administraciones públicas en el entorno de la UE.

Entre sus objetivos estratégicos destacan: *(i)* garantizar la *interoperabilidad transfronteriza de los servicios públicos digitales*, eliminando barreras y facilitando el acceso integrado de ciudadanos y empresas a trámites clave [como por ejemplo, el reconocimiento de títulos académicos, la gestión de datos sanitarios o determinados procedimientos fiscales]; *(ii)* establecer un marco común y seguro para el intercambio de datos, estandarizando criterios y armonizando procedimientos entre diferentes países miembros de la UE*; (iii)* impulsar la transformación digital del sector público, priorizando la disponibilidad en línea de los servicios esenciales hasta 2030; y, por último, *(iv)* fomentar la cooperación y gobernanza compartida mediante la creación de un *Comité europeo de interoperabilidad* y a través de la designación de autoridades nacionales competentes.

Desde enero de 2025, toda la Administración pública europea que ofrezca servicios digitales transfronterizos debe realizar *evaluaciones sistemáticas de interoperabilidad* previas a la adopción de nuevos servicios digitales transeuropeos, con el objeto de detectar de forma anticipada aquellos obstáculos legales y técnicos que imposibilitan o limitan la eficacia de los servicios públicos digitales en el marco de la UE. En efecto, los sistemas de IA se nutren de herramientas automatizadas y de IA —tales como el procesamiento del lenguaje natural, el aprendizaje automático y el modelado semántico— para extraer, transformar y conciliar datos, garantizando así su homogenización para un intercambio e integración eficientes de dichos datos. Los algoritmos que emplea la IA poseen la capacidad de analizar y comprender el contenido y el contexto de los datos, permitiendo así la identificación de patrones, lo que facilita la automatización del proceso de integración de datos entre sistemas

32. Reglamento (UE) 2024/903 del Parlamento Europeo y del Consejo, de 13 de marzo de 2024, *por el que se establecen medidas a fin de garantizar un alto nivel de interoperabilidad del sector público en toda la Unión (Reglamento sobre la Europa Interoperable*). Véase: http://data.europa.eu/eli/reg/2024/903/oj

dispares, el intercambio fluido de información entre fuentes heterogéneas y, en último término, su *interoperabilidad*.

Por tanto, las referidas *evaluaciones sistemáticas de interoperabilidad* exigidas por el Reglamento (UE) 2024/903 optimizan la integridad y la celeridad de los datos, minimizando al tiempo posibles imprecisiones y repeticiones. Asimismo, la ejecución de *evaluaciones sistemáticas de interoperabilidad* previas a la adopción de nuevos servicios digitales paneuropeos permite asimismo garantizar la calidad y la uniformidad de dichos servicios digitales, promoviendo para ello la movilidad y la cooperación ciudadanas y compartiendo soluciones técnicas, documentación y códigos según establece el llamado *Marco Europeo de Interoperabilidad*[33].

En definitiva, la obligación de este tipo de evaluaciones *ex ante* nos conduce a la adopción de un modelo de Gobernanza basado en la *interoperabilidad* como requisito ineludible a la hora de generar ecosistemas colaborativos de servicios públicos digitales de calidad, transparentes, eficientes y equitativos. Tan sólo mediante un modelo ético, responsable y colaborativo de la IA, podrá Europa aprovechar verdaderamente todas las ventajas que ofrece esta tecnología disruptiva para mejorar significativamente la vida de su ciudadanía[34].

La hoja de ruta europea hacia una Administración pública inteligente impulsada por la IA no se puede limitar a una mera implementación tecnológica. Más al contrario, la verdadera *Gobernanza inteligente* requiere la adopción de un enfoque holístico capaz de armonizar la innovación tecnológica asociada a la IA con un modelo ético que sustancia su uso responsable a través de la supervisión humana y, lo que es más importante, a través de un enfoque firme en el *valor público* capaz de fomentar el bienestar de la ciudadanía para garantizar que la IA sirva como herramienta para el progreso social.

En el siguiente epígrafe de nuestro estudio exploraremos las dimensiones hermenéuticas que implican la integración de la IA en los marcos de la Administración pública. En efecto, con el objeto de disponer de una base conceptual sólida, en la sección final de nuestro análisis, trataremos de evaluar la viabilidad del modelo de Gobernanza humana de la IA basado en el cumplimiento de principios bioéticos y el *valor público*, con el objeto de reacondicionar de este modo los paradigmas teóricos que estructuran la relación entre ciudadanía y Administración digital en Europa.

---

33. El *Marco Europeo de Interoperabilidad* (también conocido como MEI o EIF, por sus siglas en inglés) hace referencia a aquel conjunto de directrices, estándares y recomendaciones que orientan a las Administraciones públicas de la Unión Europea para asegurar que sus sistemas y servicios digitales puedan operar de forma conjunta, eficaz y segura, tanto a nivel nacional como transfronterizo. Por tanto, el MEI busca permitir la prestación de servicios públicos digitales *interoperables* a nivel europeo, facilitando el intercambio de datos, la cooperación administrativa y el acceso a servicios por parte de ciudadanos y empresas, especialmente cuando estos implican interacciones entre diferentes Estados miembros

34. A. Barsekh-Onji, Z. Torres Hernández y O.E. Cardoso Espinosa; «Advancing smart public administration: Challenges and benefits of Artificial intelligence», *Urban Governance* (5), 2025, pp. 279-292.

VI. HACIA UNA HERMENÉUTICA CRÍTICA PARA LA VIABILIDAD DE UNA GOBERNANZA HUMANA EN EL USO DE UNA IA CONFIABLE

Cada vez son más numerosas las voces críticas de quienes, como los filósofos españoles Jesús Conil o Adela Cortina[35], admiten una preocupación creciente por la viabilidad de una *Gobernanza humana de la IA*. Dado que se trata de tecnologías que han sido diseñadas para actuar de forma autónoma y adaptarse a nuevas circunstancias sin necesidad de supervisión o control humanos, la capacidad de la IA de operar sin supervisión humana necesariamente suscita importantes debates entorno a las limitaciones que los sistemas de IA disponen para comprender la percepción humana sobre los límites morales y la conducta ética, lo cual conlleva un desajuste crítico entre las capacidades humanas y tecnológicas, sobre todo cuando se abordan las consideraciones éticas que deben adoptarse en la toma de decisiones en los colectivos humanos.

Por tanto, la imparcialidad en el uso de recursos tecnológicos en la gestión del Sector Público que se ha mantenido hasta el siglo XX, en el contexto actual de la IA aplicada a la Administración pública, se considera incompleta y obsoleta para poder establecer un marco ético integral que permita alinear el uso de la IA con los valores humanos y, de este modo, garantizar una IA confIAble.

La implementación de un marco ético integral para la IA se presenta como una herramienta para contrarrestar los posibles desafíos asociados al determinismo tecnológico, fomentando para ello una hermenéutica crítica que promueve la interdependencia a través de la comprensión colectiva de los límites éticos que la IA no debe sobrepasar.

En este sentido, las instituciones de la UE han manifestado en los últimos años un respaldo político significativo hacia esta iniciativa favorable a una hermenéutica critica del uso de la IA, lo que se ha manifestado en el establecimiento de una serie de directrices basadas en cuatro principios bioéticos fundamentales aplicables al entorno digital: *autonomía, beneficencia, no maleficencia y justicia.*

(I) La *autonomía* es un principio fundamental en Bioética que se emplea para referirse a la capacidad de un individuo para tomar decisiones deliberadas respecto a sus objetivos personales y actuar de acuerdo con ellas. Sin embargo, este principio también implica autorregulación y autodeterminación. (I.1) Por un lado, la *autodeterminación* implica la libertad de elegir entre seguir las leyes establecidas o rechazarlas y, como apunta Cortina, seleccionar tanto las normas idiosincrásicas como las leyes uni-

35. J. Conill, *Ética hermenéutica*. Tecnos. Madrid, 2006. J. Conill, *Intimidad personal y persona humana. De Nietzsche a Ortega y Zubiri,* Madrid, Taurus. 2019. A. Cortina, «Ética de la Inteligencia artificial», *Anales de la RACMYP*, Madrid, 2019, 379-394. Cortina A., *Ética cosmopolita*, Paidós, Barcelona. 2021.

versales[36]. (I.2) Por lo que respecta a la *autoregulación*, la situación se vuelve más compleja ya que podría parecer que los usuarios de sistemas de IA ceden voluntariamente parte de su poder de decisión a las máquinas. De hecho, el término «sistemas autónomos» podría inducirnos a error, ya que podría sugerir una ausencia de verdadera autonomía entre los usuarios. Sin embargo, es crucial reconocer que la responsabilidad moral no puede atribuirse a la «tecnología autónoma» y que la acción humana sigue siendo primordial para determinar la responsabilidad moral; en definitiva, la responsabilidad de la toma de decisiones recae en los operadores humanos, que son los únicos capaces de justificar las decisiones que toman[37].

(II) El principio de *beneficencia* supone buscar el beneficio óptimo para el mayor número de personas e implica la obligación moral de evitar o mitigar el daño. En definitiva, este principio presupone comportamientos que conducen a la promoción del bien o a la mitigación del mal, así como a la neutralización del daño. Además, este principio implica la ausencia de cualquier acción que pueda causar daño o perjuicio[38].

(III) El principio de *no maleficencia* puede entenderse como el compromiso de actuar de manera que se eviten daños y se garantice el bienestar de todos los seres vivos. Este principio se alinea con la máxima clásica *primum non nocere* —traducida regularmente como «ante todo, no hacer daño»— y significa la obligación de abstenerse de infligir daño intencionalmente. El principio de *no maleficencia* plantea pues importantes desafíos en el contexto de los sistemas de IA, ya que nos advierte contra diversas consecuencias perjudiciales que pueden derivarse del uso indebido de las tecnologías de IA, incluidas las violaciones de la privacidad personal por el uso de algoritmos en el procesamiento de datos y la manipulación de grandes volúmenes de datos personales. La naturaleza intrincada de los daños que se derivan del empleo de algoritmos hace que el proceso de identificación de las partes responsables sea particularmente arduo y, por ello, la apelación al principio de no maleficencia nos permite que los sistemas de IA den prioridad a la seguridad y la dignidad de los posibles usuarios digitales, al tiempo que minimiza el riesgo y mejora la transparencia y la explicabilidad de los sistemas de IA.

(IV) El principio de *justicia*, en el sentido de equidad en la distribución de obligaciones y ventajas, funciona como el estándar fundamental para determinar la ética de una acción. La distribución desigual de los recursos tecnológicos y científicos puede poner en peligro la cohesión social; aplicándolo a los sistemas de IA, este principio de justicia exigiría una distribución equitativa de los beneficios derivados de las nuevas tecnologías,

36. A. Cortina, op.cit. supra., 2024, pp. 70-71.

37. Ibid., p. 73.

38. M. Kottow, «Bioética entre filosofía y medicina», *Revista De Filosofía*, 2016, pp. 49-58. https://doi.org/10.5354/0718-4360.1996.43330

> ya que el impacto de un acceso libre a la IA generalizado fomentaría la inclusión social, una distribución equitativa de los medios tecnológicos y permitiría contrarrestar la brecha digital entre países desarrollados y en vías de desarrollo.

No obstante, esta tentativa favorable al establecimiento de directrices para una IA confIAble basada en principios bioéticos ha sido objeto de un escrutinio crítico exhaustivo por parte de especialistas en ética de datos y filósofos especializados en tecnología, que han expresado en los últimos años fuertes reticencias respecto a la iniciativa de la UE de establecer un marco regulatorio de la IA en base a la integración y cumplimiento de directrices éticas. Esto es así porque las normas éticas relativas a la tecnología pueden desempeñar un papel precursor en el desarrollo de los marcos jurídicos, pero carecen de la capacidad intrínseca para reemplazar o suplantar la legislación normativa. Asimismo, la ética carece de la legitimidad democrática y el carácter vinculante necesarios para su aplicación efectiva por parte de autoridades gubernamentales y judiciales y eso, junto con la ausencia de mecanismos legales de rendición de cuentas, convierten a los principios éticos de la IA en inoperativos.

Dado el profundo impacto de los sistemas de IA en las sociedades contemporáneas, la cuestión entorno a la *gobernanza de la IA* requiere asimismo un análisis contextual centrado en los valores sociales y éticos, y, sobre todo, en un análisis que involucre a la comunidad y la ciudadanía en su conjunto en los procesos de toma de decisiones.

Como acertadamente expone el jurista italiano Alessandro Mantelero, se trata de *centrarse en los cambios que la IA traerá a la sociedad y no así en remodelar todas las áreas en las que se puede aplicar la IA*[39]. En relación a esta cuestión, Alessandro Mantelero sostiene que el debate sobre la ética de los datos ha estado siempre marcado por una superposición impropia entre Ética y Derecho, en general y por lo que respecta a los derechos humanos y derechos fundamentales, en particular. En este sentido, como apunta este autor, se ha sugerido que los desafíos éticos deben abordarse «(…) fomentando el desarrollo y las aplicaciones de la ciencia de datos y garantizando al mismo tiempo el respeto de los derechos humanos y de los valores que configuran sociedades de la información abiertas, pluralistas y tolerantes»[40]. Asimismo, para el profesor Mantelero, resulta pertinente diferenciar entre dos enfoques posibles. *(i)* De una parte, el enfoque de la «ética primero», que sostiene que la ética es la raíz de toda regulación y debe guiar el desarrollo tecnológico cuando la ley es insuficiente; este enfoque considera que la ética es el *humus prejurídico* clave cuando las normas legales son insuficientes para responder a nuevos desafíos tecnológicos. *(ii)* De otra parte, el enfoque de la «ética después» según el cual, la aplicación práctica de los derechos humanos o fundamentales implica equilibrar

39. A. Mantelero, *Regulating AI. In: Beyond Data. Information Technology and Law Series*, vol 36. T.M.C. Asser Press, The Hague, 2022. https://doi.org/10.1007/978-94-6265-531-7_4

40. Ibid., p. 95.

intereses a la luz de valores éticos ya que, en la práctica, los derechos humanos no bastan por sí solos. Por último, Alessandro Mantelero sostiene que ambos enfoques son insuficientes si se consideran por separado, ya que, a su juicio, los derechos humanos y derechos fundamentales se nutren de la ética, pero adquieren sentido y eficacia solo a través de los marcos legales y jurisprudenciales concretos. Como acertadamente señala este autor : « (…) El punto aquí no es cortar las raíces éticas, sino reconocer que los derechos y libertades florecen sobre la base de la forma que les otorgan las disposiciones legales y la jurisprudencia. No existe conflicto entre valores éticos y derechos humanos, sino que estos últimos representan una cristalización específica de dichos valores, circunscrita y contextualizada por disposiciones legales y decisiones judiciales»[41].

Sobre este mismo debate se pronunciará también el filósofo de la tecnología, contemporáneo de Mantelero, Luciano Floridi, quien insiste en la necesidad de establecer una divergencia entre las regulaciones éticas y legales relativas a la IA. En concreto, Floridi sostiene que la llamada Ética blanda (también conocido como *Soft Ethics*) sigue siendo un elemento indispensable de la «buena ciudadanía» en aquellos contextos en los que la legislación es inexistente, ambigua o necesita de interpretación[42]. Además, este autor subraya la idea de que la llamada *Ética blanda* podría contribuir al proceso de autorregulación de la IA, operando de manera complementaria a la legislación.

No obstante, con la evolución de la IA Generativa y ante los efectos devastadores que sobre los derechos fundamentales de los usuarios pudiera conllevar su uso generalizado en plataformas digitales, el empleo de algoritmos discriminatorios e incluso su impacto en el Estado de derecho y los valores democráticos, a juicio del profesor Floridi, la autorregulación de la IA debe ser reemplazada a la mayor brevedad posible por un *enfoque integral de gobernanza de la IA* sustanciada en un sistema de control de riesgos[43].

Efectivamente, cuando las regulaciones normativas relativas a la IA se integran con los principios clásicos de la Bioética en un *enfoque integral de gobernanza digital*, los marcos regulatorios se ven dotados de suficiente perspicacia interpretativa y jurídica para promover y salvaguardar los valores humanos y los derechos fundamentales. Además, la adopción de un enfoque integral tiene el potencial de ofrecer una salvaguardia más completa de los valores humanos fundamentales en comparación con las limitaciones de un marco puramente sustantivo y procedimental, construyendo así regulaciones más sensibles a los contextos sociales y a los principios éticos compartidos en una comunidad. Por

41. Ibid., p. 96.

42. L. Floridi, *Soft ethics, the governance of the digital and the General Data Protection Regulation,* Philosophical Transactions of the Royal Society Mathematical, Physical and Engineering Sciences, 376 (2133), 2018. L. Floridi, «Translating Principles into practices of digital ethics: Five risks of being unethical», *Philosophy & Technology*, 32(2), 2019, pp.185-193.

43. L. Floridi, L., «The End of an Era: from Self-Regulation to Hard Law for the Digital Industry», *Philos. Technol.* 34, 2021, pp. 619-622. https://doi.org/10.1007/s13347-021-00493. Floridi, L., «The European Legislation on AI: A brief analysis of its philosophical approach», *Philosophy & Technology*, 34(2), 2021, pp. 215-222.

último, este enfoque integrador tiene el potencial de garantizar la coherencia de las líneas de acción elegidas con un sistema más amplio de creencias morales y de imperativos ideológicos de orden superior dentro de una sociedad.

VII. CONCLUSIONES

La magnitud y la complejidad de la transformación que la Inteligencia Artificial está impulsando en el ámbito de la Administración pública europea es inconmensurable. Lejos de contribuir a la aceleración de la digitalización de los servicios públicos, la implementación del uso de la IA en el Sector Público ha reacondicionado los fundamentos mismos de la gestión administrativa, lo que resulta en un desplazamiento progresivo de los modelos burocráticos tradicionales hacia nuevos paradigmas *post-burocráticos* centrados en la creación de *valor público* y la satisfacción ciudadana.

La transición desde la burocracia clásica hacia modelos de Nueva Gestión Pública y, más recientemente, hacia un modelo de *Gobernanza inteligente basada en la IA*, ha supuesto una redefinición de los roles, las responsabilidades y las expectativas tanto de los gestores públicos como de la propia ciudadanía en su conjunto. En este contexto, la IA emerge como un elemento de eficiencia, transparencia y personalización de los servicios administrativos que ofrece el Sector Público, permitiendo optimizar procesos administrativos, liberar recursos para tareas de mayor valor añadido, mejorar la capacidad de anticipación y respuesta de las instituciones y fomentar la transparencia y la participación ciudadana a través de plataformas digitales. Las experiencias pioneras de digitalización casi total de los servicios públicos en Estonia, la gestión predictiva en los países nórdicos y la implementación de sistemas de participación ciudadana en España, son ejemplos que ilustran este potencial transformador de la IA en la gestión pública europea. No obstante, dicho potencial innovador se encuentra ineludiblemente vinculado a una serie de riesgos y limitaciones que van desde la opacidad de los algoritmos, la posibilidad de reproducción de sesgos discriminatorios, la amenaza a la privacidad y la protección de datos y, lo que resulta más complejo en su erradicación, la dificultad para establecer mecanismos efectivos de rendición de cuentas. Por tanto, esta complejidad técnica inherente a los sistemas de IA impone, por añadidura, la necesidad imperante de una capacitación continua del personal público de las Administraciones y el desarrollo de marcos normativos y éticos robustos que aseguren la protección de los derechos fundamentales y la confianza ciudadana.

En este sentido, el Reglamento (UE) 2024/903 *sobre una Europa Interoperable* constituye un avance significativo en la articulación de un espacio digital común en el que la IA desempeña un papel central para garantizar la *interoperabilidad, la eficiencia y la calidad de los servicios públicos transfronterizos y paneuropeos*. La implementación de evaluaciones sistemáticas de interoperabilidad y la creación de marcos comunes para el intercambio de datos constituyen

pasos decisivos hacia una Administración pública europea más conectada, más transparente y, sobre todo, más orientada al ciudadano.

Sin embargo, la mera implementación tecnológica de sistemas de IA no resulta suficiente. La Gobernanza de la IA requiere asimismo la integración de la innovación tecnológica con un enfoque ético y humano, en el que la supervisión, la transparencia y la rendición de cuentas sean principios irrenunciables. Como han preconizado destacadas figuras como Alessandro Mantelero y Luciano Floridi, ni *la ética puede reemplazar la regulación jurídica, ni el derecho puede ignorar los valores éticos que fundamentan un marco de gobernanza integral de la IA,* que permita dotar a la regulación de la IA de la flexibilidad, sensibilidad y eficacia necesarias para responder a los desafíos de una sociedad digital plural y dinámica. En este sentido, la adopción de *principios bioéticos* —tales como la autonomía, la beneficencia, la no maleficencia y la justicia—en conjunción con la exigencia de marcos normativos vinculantes y mecanismos efectivos de control y rendición de cuentas, constituye la base de una *IA confiable y legítima* en el ámbito de la Administración pública.

A pesar de los importantes avances normativos y tecnológicos que se han llevado a cabo, y que hemos avanzado en el presente estudio, el debate sobre la integración de la IA en la Administración pública europea está lejos de concluirse. Las nuevas adaptaciones de la IA Generativa nos enfrentan cada vez más, y más rápido, a nuevas controversias y desafíos que requieren de una reflexión hermenéutica crítica y de un diálogo interdisciplinario continuo. Entre los asuntos que se postulan como prioritarios y que deberán ser objeto de debate en ulteriores análisis sobre la materia, destacan los siguientes:

(i) La implementación de mecanismos para la identificación, mitigación y corrección de sesgos inherentes a los sistemas de IA se vislumbra como una necesidad ineludible. Asimismo, resulta imperativo garantizar que la automatización no reproduzca ni amplifique las desigualdades estructurales, promoviendo así la inclusión y la equidad en la sociedad.

(ii) Por lo que respecta a la protección de los derechos fundamentales en el entorno digital, es evidente que los marcos normativos actuales son insuficientes para salvaguardar la privacidad, la protección de datos y la no discriminación en la era de la IA; resultando pues inevitable emprender reformas legales y adoptar garantías adicionales.

(iii) En el ámbito de la democracia ante la IA, la participación ciudadana constituye un pilar fundamental a la hora de garantizar la legitimidad del sistema democrático ante el uso de la IA.

(iv) La implementación de modelos de participación y control social se erige como un desafío imperativo con el propósito de involucrar a la ciudadanía en el proceso de Gobernanza de la IA. En este contexto, se plantea la interrogante respecto a los mecanismos más adecuados para evitar la tecnocracia y, simultáneamente, promover la corresponsabilidad en la definición de prioridades y la evaluación de políticas públicas.

Por todo lo expuesto, debemos ser cautos a la hora de adoptar códigos de *gobernanza* de la IA que carezcan de obligaciones normativas definidas y debemos orientarnos mejor a aquellos mecanismos de control que favorezcan la rendición de cuentas de forma inequívoca. La cuestión fundamental no es sólo determinar si es necesario implementar medidas legales para abordar las preocupaciones éticas asociadas a la IA, sino también establecer cómo deben implementarse dichas medidas para garantizar el desarrollo y la aplicación de la IA en términos de responsabilidad ética y social. En ese sentido, como hemos visto, son numerosos los especialistas en ética y filósofos de la tecnología que vienen preconizando la necesidad de establecer una interacción complementaria entre el ámbito jurídico y la ética; así pues, en los últimos años venimos observando una tendencia recurrente a un enfoque holístico de gobernanza de la IA que invoca la Ética y el Derecho de forma conjunta en un esfuerzo colaborativo con el objeto de proteger y promover de manera efectiva los valores humanos.

En conclusión, la integración de la IA en la Administración pública europea se manifiesta como un proceso abierto, dinámico y profundamente transgresor, en el que está en juego no solo la eficiencia y la modernización del Sector Público, sino también la calidad democrática, la protección de los derechos fundamentales y la confianza ciudadana en las instituciones. La gobernanza de la IA, por tanto, requiere la adopción de un enfoque hermenéutico crítico, reflexivo y colaborativo, capaz de anticipar riesgos, corregir desviaciones y aprovechar las oportunidades que ofrece esta tecnología para construir un uso *antropocéntrico* de la IA en Administración pública más inclusivo y orientado al Bien común.

BIBLIOGRAFÍA

AUCOIN, P., *Reforma administrativa en la gestión pública: Paradigmas, principios, paradojas y péndulos*, en Brugue, Q. y Subirats, J. (comps.). *Lecturas de gestión pública*, Instituto Nacional de Administración Pública, Madrid, 1996, pp. 293- 515.

BARZELAY, M., *El nuevo paradigma de la gestión pública*, Fondo de Cultura Económica, México, 2003.

BARZELAY, M., *La nueva gerencia pública. Un acercamiento a la investigación y al debate de las políticas públicas*. Fondo de Cultura Económica, México, 2005.

BRESSER PEREIRA, L. C., *Lo público no estatal en la reforma del Estado*. Ed. Paidós, Buenos Aires, 1998.

BARSEKH-ONJI, A., TORRES HERNÁNDEZ, Z. y CARDOSO ESPINOSA; O.E., «Advancing smart public administration: Challenges and benefits of Artificial intelligence», *Urban Governance* (5), 2025, pp. 279-292.

CHICA VÉLEZ, S., «Una mirada a los nuevos enfoques de la gestión pública», *Administración & Desarrollo*, 39 (53), 2011, pp. 57-74.

CONILL, J., *Ética hermenéutica*. Tecnos. Madrid, 2006.

CONILL, J., *Intimidad personal y persona humana. De Nietzsche a Ortega y Zubiri,* Madrid, Taurus. 2019.

CORTINA, A., «Ética de la Inteligencia artificial», *Anales de la RACMYP*, Madrid, 2019, pp. 379-394.

CORTINA A., *Ética cosmopolita*, Paidós, Barcelona. 2021.

CRIADO, J. I., «Las administraciones públicas en la era del gobierno abierto. Gobernanza inteligente para un cambio de paradigma en la gestión pública», *Revista de Estudios Políticos* (173), 2016, pp. 245-275. DOI: http://dx.doi.org/10.18042/cepc/rep.173.07.

CRIADO, J. I., «Inteligencia Artificial y Administración Pública», *Eunomía: Revista en Cultura de la Legalidad*, 2021.

DU GAY, P., *Praise of Bureaucracy: Weber, Organization, Ethics*, Sage, London, 2000.

FEFFER, J., *Nuevos rumbos en la teoría de las organizaciones*, Fondo de Cultura Económica, México, 2000.

FLORIDI, L., *Soft ethics, the governance of the digital and the General Data Protection Regulation*. Philosophical Transactions of the Royal Society Mathematical, Physical and Engineering Sciences, 376 (2133), 2018.

FLORIDI, L., «Translating Principles into practices of digital ethics: Five risks of being unethical», *Philosophy & Technology*, 32(2), 2019, pp.185-193.

FLORIDI, L., «The End of an Era: from Self-Regulation to Hard Law for the Digital Industry», *Philos. Technol.* 34, 2021, pp. 619-622. https://doi.org/10.1007/s13347-021-00493.

FLORIDI, L., «The European Legislation on AI: A brief analysis of its philosophical approach», *Philosophy & Technology*, 34(2), 2021, pp. 215-222.

GILLESPIE, T., «The Relevance of Algorithms», en Gillespie, T. Boczkowski, P.J. & Foot, K.A. (Eds.), *Media Technologies: Essays on Communication, Materiality and Society*, MIT Press, Massachussets, 2014, pp. 167-194.

MARGETTS, H., & Dorobantu, C., *Rethinking Public Policy in the Era of AI: Lessons from the United Kingdom, Nature Communications*, Num. 11, 2020.

MANTELERO, A., *Regulating AI. In: Beyond Data. Information Technology and Law Series*, vol 36. T.M.C. Asser Press, The Hague, 2022. https://doi.org/10.1007/978-94-6265-531-7_4

MARTÍNEZ CASTILLA, S.M., «La burocracia: elemento de dominación en la obra de Max Weber», *Revista Misión Jurídica*, num.10, 2016, pp. 141-154.

MOORE, M. H., *Creating Public Value: Strategic Management in Government*. Cambridge, MA: Harvard University Press, 1995.

MOORE, M. H., *Gestión estratégica y creación de valor en el sector público*. Ed. Paidós, Barcelona, 1998.

OSBORNE, D. y GAEBLER, T., *La reinvención del Gobierno*, Addison-Wesley, New York, 1992.

POLLITT, C. & BOUCKAERT, G., *Public Management Reform. A Comparative Analysis: Into the Age of Austerity*, Oxford University Press, Cambridge, 2017.

KETTUNEN, P., & KALLIO, J., «Digital Government in the Nordic Countries: From e-Government to Smart Government», *Government Information Quarterly*, 38(1), 2021.

KOTTOW, M., «Bioética entre filosofía y medicina», *Revista De Filosofía*, 2016, pp. 49-58. https://doi.org/10.5354/0718-4360.1996.43330

WEBER, M., *Economía y sociedad*, Fondo de Cultura Económica, México, 1922.

WEBER, M., *The Theory of Social and Economic Organization*. Free Press, Nueva York, 1947.

ZARSKY, T. Z., «The Trouble with Algorithmic Decisions: An Analytic Road Map to the Legal Debate», *Science, Technology, & Human Values*, 41(1), 2016, pp. 118-132.

Las decisiones automatizadas relevantes en el reglamento general de protección de datos: un análisis a las últimas resoluciones judiciales del TJUE[1]

Adrián Palma Ortigosa
Profesor Ayudante Doctor de Derecho Administrativo
Universidad de Valencia

SUMARIO: I. INTRODUCCIÓN. II ORIGEN DE LAS DECISIONES AUTOMATIZADAS EN LA NORMATIVA DE PROTECCIÓN DE DATOS. ¿DE DÓNDE VENIMOS Y DÓNDE ESTAMOS? 1. La ley francesa de 1978. 2. La Directiva de protección de datos de 1995. 3. El Reglamento Europeo de Protección de Datos y el avance de la inteligencia artificial. III. LAS DECISIONES AUTOMATIZADAS RELEVANTES. 1. Las decisiones plenamente automatizadas. 2. Los efectos jurídicos o significativamente similares. Los efectos relevantes. IV. BASES DE LEGITIMACIÓN DE LAS DECISIONES AUTOMATIZADAS RELEVANTES. 1. El principio de prohibición general y su fundamento. 2. Las excepciones a las decisiones automatizadas relevantes. V. GARANTÍAS DEL INTERESADO FRENTE A LAS DECISIONES AUTOMATIZADAS RELEVANTES. 1. Las garantías expresas del artículo 22. 2.Otras garantías aplicables a las decisiones automatizadas relevantes. VI. EL ARTÍCULO 22 Y LOS MODELOS DE INTELIGENCIA ARTIFICIAL GENERATIVA. VII. CONCLUSIONES. BIBLIOGRAFÍA.

1. Esta investigación se ha realizado en el marco del proyecto de I+D+i *Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas* [2023-2026] (PID2022-136439OB-I00), financiado por MCIN/AEI/10.13039/501100011033/ y «FEDER Una manera de hacer Europa».

I. INTRODUCCIÓN

El presente trabajo analiza el régimen jurídico de las decisiones automatizadas reguladas en el artículo 22 del Reglamento General de Protección de Datos (RGPD). Este precepto constituye una de las disposiciones más complejas y relevantes del marco normativo europeo de protección de datos, especialmente en el contexto del desarrollo de la inteligencia artificial y los sistemas algorítmicos. El trabajo examina el origen histórico de esta regulación, delimita el concepto de decisión automatizada relevante, analiza las bases de legitimación que permiten su implementación y estudia las garantías específicas establecidas en favor de los interesados tomando como referencia las últimas resoluciones judiciales del TJUE. Además, se plantean interrogantes sobre la aplicación de este precepto a los nuevos avances ligados a la inteligencia artificial generativa.

II. ORIGEN DE LAS DECISIONES AUTOMATIZADAS EN LA NORMATIVA DE PROTECCIÓN DE DATOS. ¿DE DÓNDE VENIMOS Y DÓNDE ESTAMOS?

1. La ley francesa de 1978

La primera aproximación a la regulación de las decisiones automatizadas tuvo su origen a través la *Ley Francesa relativa a la tecnología de la información, los ficheros y las libertades* de 1978. Esta norma establecía en su artículo 2 una prohibición revolucionaria que sentaría las bases del régimen legal objeto de análisis de este trabajo. Este precepto establecía que «*ninguna decisión judicial que implique una evaluación del comportamiento humano puede basarse en el procesamiento automatizado de información que dé una definición del perfil o la personalidad de la persona en cuestión. Ninguna decisión administrativa o privada que implique una evaluación del comportamiento humano puede tener como única base un tratamiento automatizado de la información que dé una definición del perfil o personalidad del interesado*». Esta formulación francesa ya contenía algunos de los elementos esenciales que posteriormente desarrollaría el derecho europeo ligados a la prohibición de decisiones basadas únicamente en tratamiento automatizado[2]. Para comprender plenamente el alcance innovador de la ley francesa de 1978, es preciso situarla en su contexto tecnológico y social. Los años 70 presenciaron el desarrollo acelerado de los primeros sistemas informáticos de gestión masiva de datos. En el ámbito público, las administraciones comenzaron a informatizar registros poblacionales, sistemas de seguridad social y bases de datos fiscales. En el sector privado, las entidades financieras empezaron a utilizar los primeros sistemas automatizados de evaluación crediticia. Francia experimentó de manera particularmente inten-

2. Artículo 2. LOI n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. Disponible en: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000886460

sa estos cambios. El origen de esta disposición en la legislación francesa se encontraba en el sistema *GAMIN* (sistema de *gestión automatisée de médecine infantil*), un sistema de puntuación diseñado en 1970 por el Ministerio de Salud para prevenir discapacidades en los niños[3].

2. La Directiva europea de protección de datos de 1995

La experiencia acumulada por los Estados miembros de la UE durante los años 80 y principios de los 90 puso de manifiesto la necesidad de armonizar las regulaciones nacionales en materia de protección de datos. La Directiva 95/46/CE del Parlamento Europeo y del Consejo, de 24 de octubre de 1995, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos incorporó en su artículo 15 las previsiones sobre decisiones automatizadas de forma muy parecida al contemplado en la Ley francesa de 1978.

El artículo 15 de la Directiva introducía varios elementos innovadores respecto de su predecesor[4]. Primero, la prohibición de decisiones automatizadas se configuraba como un derecho de los titulares de los datos sobre los que se adoptaba la decisión. Segundo, se establecían excepciones que permitían este tipo de decisiones bajo determinadas condiciones y, se exigía el despliegue de ciertas garantías cuando se llevaran a cabo dichos tratamientos de datos. Por su parte, el artículo 12 de esta misma directiva reconocía el derecho de los interesados a conocer la lógica que había detrás de estos tratamientos totalmente automatizados.

A pesar del reconocimiento de este derecho en favor de los interesados, su aplicación práctica fue limitada. Varios factores contribuyeron a esta limitada aplicación. En primer lugar, la redacción del artículo 15 de la Directiva presentaba ambigüedades interpretativas. La determinación de cuándo una decisión tenía «efectos significativos» o se basaba «únicamente» en tratamiento automatizado generaba incertidumbre jurídica. En segundo lugar, el desarrollo tecnológico de los años 90 y 2000 aún no había alcanzado el grado de sofisticación que permitiera la implementación masiva de sistemas de decisiones automatiza-

3. Genicot, N. «Scoring the European citizen in the AI era», *Computer Law & Security Review: The International Journal of Technology Law and Practice,* 57, 2025,p.3.

4. Artículo 15 Directiva 95/46 Artículo 15. 1. *Los Estados miembros reconocerán a las personas el derecho a no verse sometidas a una decisión con efectos jurídicos sobre ellas o que les afecte de manera significativa, que se base únicamente en un tratamiento automatizado de datos destinado a evaluar determinados aspectos de su personalidad, como su rendimiento laboral, crédito, fiabilidad, conducta, etc. 2. Los Estados miembros permitirán, sin perjuicio de lo dispuesto en los demás artículos de la presente Directiva, que una persona pueda verse sometida a una de las decisiones contempladas en el apartado 1 cuando dicha decisión: a) se haya adoptado en el marco de la celebración o ejecución de un contrato, siempre que la petición de celebración o ejecución del contrato presentada por el interesado se haya satisfecho o que existan medidas apropiadas, como la posibilidad de defender su punto de vista, para la salvaguardia de su interés legítimo; o b) esté autorizada por una ley que establezca medidas que garanticen el interés legítimo del interesado.*

das precisos. Por último, la falta de sensibilización tanto de los responsables del tratamiento como de los propios interesados sobre estos derechos respecto de otros como el derecho de acceso o rectificación limitó su ejercicio efectivo. La cultura jurídica de la época no había interiorizado plenamente los riesgos asociados a la automatización de estas decisiones, que, aunque pudieran estar sucediendo en la práctica, su desconocimiento resultaba obvio.

3. El Reglamento Europeo de Protección de Datos y el avance de la inteligencia artificial

La entrada del siglo XXI trajo consigo transformaciones tecnológicas radicales que hicieron evidente la necesidad de modernizar el marco normativo europeo. El desarrollo de internet, la proliferación de dispositivos conectados, la emergencia de las redes sociales y el big data, crearon un ecosistema tecnológico completamente diferente al que había conocido la Directiva 95/46/CE.

Así, el considerando 71 del Reglamento General de Protección de datos ya reflejaba claramente las intenciones del legislador al desarrollar un marco claro de regulación sobre la elaboración de perfiles y las decisiones automatizadas, si bien, la regulación de estos tratamientos en el artículo 22 no era muy diferente al artículo 15 de la Directiva 95/46.

Pues bien, aunque en un momento inicial el artículo 22 pasó desapercibido, en los últimos años existe una tendencia clara de aplicación. Este cambio de dinámica obedece a varias razones que consideramos consecutivas las unas de las otras.

En primer lugar, el avance exponencial de la inteligencia artificial. Se han dado las condiciones necesarias para que esta tecnología que se teorizó inicialmente en los años 50 del siglo XIX tenga su aplicación práctica real en nuestros días[5]. Consecuencia de ello, empresas y organizaciones públicas han comenzado a utilizar estas herramientas debido a que cada vez son más precisas.

En segundo lugar, ha habido un interés creciente en la sociedad, los poderes públicos, así como en la academia por profundizar en los riesgos legales, sociales y éticos que puede generar el despliegue de estas herramientas. Ello ha permitido poner el foco de atención no solo en la inteligencia artificial, sino en otras herramientas informáticas deterministas que hasta la fecha habían pasado desapercibidas[6], o en la propia elaboración de perfiles que se viene haciendo desde hace tiempo en sectores como el bancario o el de seguros.

5. Se han unido tres factores claves: Disponibilidad masiva de datos, capacidad mayor y más eficiente de esos datos y mejores instalaciones para almacenar dichos datos. MondalL, B, «Artificial Intelligence: State of the Art», en: V. Balas; R. Kumar; R. SRIVASTAVA (eds), *Recent Trends and Advances in Artificial Intelligence and Internet of Things*. Springer, Cham, vol 172, 2020, pp. 389-425. Disponible en: https://doi.org/10.1007/978-3-030-32644-9_32

6. En el sector público español, los programas informáticos de VioGén o Riscanvi se llevan utilizando mucho antes de la entrada en vigor del artículo 22 del RGPD.

En tercer lugar, y a falta de una regulación específica que trate de afrontar los riesgos derivados del uso de la inteligencia artificial o algoritmos en el uso de toma de decisiones automatizadas, el artículo 22 se ha mostrado como una de las pocas garantías que resulta aplicable a algunos de los tratamientos de datos que están presentes en el ciclo de vida de los sistemas de IA. Muestra de ello lo ofrecen el aumento progresivo de resoluciones administrativas de las autoridades de protección de datos, así como resoluciones judiciales que estudian e interpretan el mentado artículo 22.

III. LAS DECISIONES AUTOMATIZADAS RELEVANTES

El artículo 22.1 del RGPD no se centra en todo tipo de decisiones, sino que pone el acento en las decisiones plenamente automatizadas que generan efectos significativos o similares. Son por tanto dos requisitos acumulativos los que deben estar presentes, la plena automatización del proceso decisorio y la relevancia de la decisión que se adopta.

1. Las decisiones plenamente automatizadas

Como es lógico, resulta necesario que todo el proceso del tratamiento de datos sea totalmente automatizado. Es decir, se ingresa el dato personal, el algoritmo lo procesa y éste último emite un resultado que automáticamente afecta a un individuo. Es por ello que cualquier tipo de procesamiento algorítmico, ya sea más o menos determinista entra dentro del ámbito de aplicación de esta norma. El tipo de tecnología que se utilice como hemos dicho previamente es irrelevante en este contexto siempre que el proceso decisorio se automatice por completo. En este sentido cabe destacar que no es necesario por tanto que la decisión automatizada requiera de un perfilado previo, sino que basta con que exista una plena automatización, sea con o sin dicho perfilado[7].

A) La supervisión real y significativa

Solo cuando exista una intervención humana real y significativa podremos considerar que no estamos ante una decisión plenamente automatizada. Existen diferentes autoridades que han publicado guías y directrices que ayudan a valorar cuándo la presencia de un humano en el proceso decisorio es real[8]. Estas

7. Future of Privacy Forum. *Automated Decision-Making Under the GDPR: Practical Cases from Courts and Data Protection Authorities*, 2022, p.8 Disponible en: https://fpf.org/wp-content/uploads/2022/05/FPF-ADM-Report-R2-singles.pdf?utm_source=chatgpt.com

8. Grupo Artículo 29. *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679*. 2018, p.23. AEPD. *Evaluación de la intervención humana en las decisiones automatizadas, 2024. Disponible en*::https://www.aepd.es/prensa-y-comunicacion/blog/evaluacion-de-la-intervencion-humana-en-las-decisiones-automatizadas. Infor-

guías se centran en el tipo de supervisión que se ha implementado durante el proceso decisorio o la formación, capacidad y autoridad de las personas que supervisan los resultados algorítmicos. Corresponde a los responsables del tratamiento justificar adecuadamente esa supervisión humana. Puede resultar muy conveniente también tomar como referencia las normas armonizadas que en su caso se están desarrollando actualmente por parte de los organismos de normalización europeos que tienen como objetivo aterrizar los requisitos exigidos a los sistemas de IA de alto riesgo regulados en el Reglamento de IA[9]. Así, uno de esos requisitos es el de la supervisión humana real[10]. Estas normas establecerán reglas más precisas para justificar dicha presencia humana[11].

B) El valor determinante del resultado algorítmico como decisión automatizada relevante

Recientemente, la Sentencia del Tribunal de Justicia de la Unión Europea (TJUE) de 7 de diciembre de 2023 en el caso SCHUFA ha establecido una interpretación amplia del concepto de decisión automatizada[12]. El Tribunal considera que constituye una decisión totalmente automatizada con efectos relevantes la generación por una agencia de información comercial de un valor de probabilidad automatizado sobre la capacidad de una persona para hacer frente a futuros compromisos de pago, cuando de ese valor de probabilidad dependa de manera determinante la decisión por la cual un tercero establezca, ejecute o ponga fin a una relación contractual con esa persona.

Es decir, este caso involucra tres actores, el primero, la agencia de información comercial que genera el scoring crediticio, un segundo, que es una organización que utiliza ese scoring para tomar la decisión (un banco por ejemplo), y, en tercer lugar, la persona afectada, en este caso el solicitante del crédito u

---

mation Commissioners Officer. *What does the UK GDPR say about automated decision-making and profiling?* Disponible en: :https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/automated-decision-making-and-profiling/what-does-the-uk-gdpr-say-about-automated-decision-making-and-profiling/#id2. Supervisor Europeo de Protección de datos.TechDispatch #2/2025 - Human Oversight of Automated Decision-Making, 2025. Disponible en: https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2025-09-23-techdispatch-22025-human-oversight-automated-making_en

9. Sobre las normas armonizadas y su papel en el cumplimiento normativo de la legislación europea véase: Comisión Europea. *Comunicación de la Comisión «Guía azul» sobre la aplicación de la normativa europea relativa a los productos*, 2022, p.49.

10. Artículo 14. Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.° 300/2008, (UE) n.° 167/2013, (UE) n.° 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial).

11. Radtke, «Human Oversight under the AI Act and its interplay with Art. 22 GDPR», en Raue/von Ungern-Sternberg/Kumkar/Rüfner(ed.), Artificial Intelligence and Fundamental RightsThe AI Act of the European Union and its implications for global technology regulation, Verein für Recht und Digitalisierung, Trier, 2025.

12. STJUE Asunto C-634/21Schufa de 7 de diciembre de 2023.

otro tipo de contrato. Para el TJUE, por tanto, lo relevante es que el valor generado algorítmicamente por la agencia condicione de manera decisiva la decisión final del tercero. De esta manera, ese valor por sí solo es una decisión totalmente automatizada relevante, si bien, queda condicionada a cómo de determinante sea respecto de la decisión que adopta la tercera entidad.

El TJUE justifica esta interpretación amplia para evitar la elusión del artículo 22 del RGPD, ya que, si se adoptara una interpretación restrictiva que considerase la generación del valor de probabilidad como mero acto preparatorio y únicamente el acto adoptado por el tercero como «decisión», se produciría una doble vulneración del RGPD. Así, por un lado, respecto de la agencia generadora del scoring, el afectado por la decisión no podría ejercer su derecho de acceso a la información específica prevista en el artículo 15.1.h) del RGPD, al no existir formalmente una «decisión automatizada» adoptada por dicha agencia. Por otro lado, por lo que se refiere a la entidad que adopta la decisión final, aunque el acto del tercero (por ejemplo, el banco) pudiera calificarse como decisión automatizada sometida al artículo 22.1 del RGPD, este tercero no podría facilitar la información específica exigida por el artículo 15.1.h), puesto que generalmente no dispone de ella. El banco conoce el resultado del scoring pero no la lógica subyacente, ni los factores ponderados ni los criterios utilizados por la agencia para generarlo[13].

Esta sentencia, si bien realiza una interpretación en favor de los interesados que pueden quedar desprotegidos por las operaciones presentes en este circuito específico, deja algunas cuestiones sin resolver. En primer lugar, el TJUE no clarifica en qué medida debe ser determinante el valor del scoring para considerarlo decisión automatizada, lo que llevará a analizar caso por caso cuándo existirá ese importancia del resultado algorítmico en la decisión final[14]. Además, en segundo lugar, la sentencia no aclara si esta doctrina del valor determinante es generalizable a otras situaciones con estructuras decisorias donde intervienen múltiples responsables del tratamiento.

Sobre esto último, la Agencia Española de Protección de Datos en una resolución muy extensa ha tomado en parte como referencia la sentencia del TJUE y ha considerado decisión totalmente automatizado un caso en el que también intervienen varios responsables del tratamiento[15]. En este supuesto también encontrábamos tres actores. Primero, la persona que solicita el bono social eléctrico, es decir, una reducción en la factura de electricidad. Segundo, la comercializadora eléctrica de referencia que recibe las solicitudes de los interesados. Tercero, el Ministerio para la Transformación Ecológica que tiene en su poder el programa informático que establece las personas que tienen derecho o no al bono social en función de las solicitudes que le haya facilitado la comercializadora. Emitido el resultado por el programa informático, el resultado es comunicado a la comercializadora para que en su caso ésta última aplique o

13. STJUE Asunto C-634/21 Schufa de 7 de diciembre de 2023. FJ.63.
14. Arroyo Amayuelas, E. «El *scoring* de Schufa», *InDret*, 3, 2024, p,13.
15. Agencia Española de Protección de Datos. Expediente N.º: EXP202211982

no el bono social[16]. Una vez más por tanto, el hecho de que haya más o menos responsables en el proceso decisorio es irrelevante si en tal proceso no existe intervención humana o existiendo, esta es irrelevante.

2. Los efectos jurídicos o significativamente similares. Los efectos relevantes.

El GT29 ha interpretado que una decisión produce efectos jurídicos cuando «afecte a los derechos o al estatuto jurídico del titular que se ve sometido a dicho tratamiento». Algunos ejemplos paradigmáticos incluyen la denegación de una prestación económica, la no obtención de un contrato, el despido de un trabajador o la concesión o denegación de un préstamo.

Respecto a las decisiones que «afectan significativamente de modo similar», estas abarcan aquellas donde, sin existir un cambio en las obligaciones o derechos jurídicos del particular, la decisión resulta de una importancia similar a las que producen efectos jurídicos. Por ejemplo, para el TJUE, en la Sentencia Schufa previamente analizada, el valor de probabilidad indicado por el algoritmo que se utiliza por un tercero para obtener un crédito se considera que genera estos efectos[17].

IV. BASES DE LEGITIMACIÓN DE LAS DECISIONES AUTOMATIZADAS RELEVANTES

1. El Principio de prohibición general y su fundamento

El artículo 22.1 del RGPD establece una prohibición general inicial sobre la toma de decisiones plenamente automatizadas relevantes[18]. Esta prohibición opera automáticamente, independientemente de que el interesado la invoque expresamente.

Este enfoque, a priori preventivo, contrasta con el régimen general del RGPD, donde la licitud del tratamiento se establece mediante la concurrencia de una base jurídica del artículo 6. En el caso del artículo 22, el legislador ha invertido la carga. Las decisiones automatizadas están prohibidas salvo que concurra una de las excepciones tasadas. Dicho lo cual, es innegable que las excepciones que habilitan a las decisiones automatizadas relevantes no dejan de ser algunas de las bases de legitimación que se mencionan en el artículo 6 del RGPD.

16. Agencia Española de Protección de Datos. Expediente N.º: EXP202211982, p.72.

17. STJUE Asunto C-634/21 Schufa de 7 de diciembre de 2023. FJ.50.

18. Como ha señalado el GT29, este precepto «contiene una prohibición general de llevar a cabo tratamientos basados en decisiones plenamente automatizadas relevantes». Grupo Artículo 29. *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679.* 2018, p.25.

Además de esa prohibición general, no podemos olvidar que este precepto se encuadra en la norma europea dentro de los apartados referidos a los derechos de los interesados, de ahí que las limitaciones a este derecho, y por tanto las excepciones aplicables que habilitan a la toma de decisiones automatizadas relevantes han de estar debidamente justificadas.

2. Las excepciones a las decisiones automatizadas relevantes

El artículo 22.2 del RGPD establece un sistema de excepciones taxativas que permiten levantar la prohibición general. Las tres excepciones establecidas responden a lógicas diferentes. Por un lado, la excepción contractual (artículo 22.2.a), la cual reconoce la necesesariedad del tratamiento algorítmico a la hora de formalizar o ejecutar ciertos contratos. Por otro lado, la excepción legal, en este caso, por medio de una norma nacional o una europea se puede autorizar este tratamiento. Esta excepción está especialmente diseñada para facilitar la toma de decisiones plenamente automatizada por parte de los poderes públicos. Finalmente, la tercera excepción se refiere al consentimiento explícito. En este caso se respeta la autonomía personal del interesado para que éste acepte ser sometido a estos tratamientos de plena automatización.

A) La excepción de necesidad contractual

La primera excepción contemplada en el artículo 22.2.a) del RGPD permite la toma de decisiones automatizadas cuando resulte necesaria para la celebración o ejecución de un contrato entre el interesado y el responsable del tratamiento. El requisito de necesidad que incorpora esta excepción no constituye una novedad aislada del artículo 22, sino que se integra en el marco más amplio del artículo 6.1.b) del RGPD, que establece la necesidad contractual como una de las bases generales de legitimación para el tratamiento de datos personales.

El concepto de necesidad contractual constituye un estándar jurídico indeterminado cuya concreción requiere una valoración casuística por parte del responsable. El Grupo de Trabajo del Artículo 29 ha precisado que el responsable del tratamiento debe poder demostrar que la decisión automatizada resulta necesaria, considerando si existe un método alternativo menos invasivo para la privacidad que permita alcanzar el mismo objetivo. En caso de que existan otros medios efectivos y menos intrusivos para lograr la finalidad contractual, el tratamiento automatizado no podrá considerarse «necesario» en el sentido del artículo 22.2.a).

El Grupo de Trabajo del Artículo 29 ha considerado necesario para el objeto del contrato el establecimiento de un sistema automatizado de filtrado de solicitudes en procesos selectivos con un volumen extremadamente elevado de candidaturas.[19]. La necesidad concurre por tanto en aquellos supuestos donde

19. El filtro inicial automatizado permitiría realizar un cribado preliminar de las solicitudes manifiestamente irrelevantes, resultando impracticable una revisión humana individualizada de la

la participación humana rutinaria puede resultar poco práctica o imposible debido a la magnitud de los datos tratados.

En el mismo sentido, la AEPD también ha considerado recientemente que para gestionar de forma adecuada el contrato que se formaliza entre el interesado que solicita el descuento del Bono Social electrónico y la compañía comercializadora de electricidad entra en juego la excepción de la necesariedad del contrato prevista en el artículo 22.2.a)[20].

La invocación de la excepción de necesidad contractual impone al responsable del tratamiento una carga argumentativa cualificada. No resulta por tanto suficiente una mera alegación genérica de conveniencia o eficiencia. El responsable debe evidenciar que la inclusión de la decisión automatizada no obedece a una simple preferencia organizativa, sino que responde a una necesidad objetiva vinculada a la naturaleza o ejecución del contrato.

Esta justificación puede articularse mediante diferentes líneas argumentativas. El responsable puede acreditar que el sistema automatizado alcanza al menos el mismo nivel de precisión que la intervención humana en la realización de las operaciones de que se trate, evitando así que la automatización redunde en un menoscabo de la calidad o fiabilidad del proceso decisorio. En otros supuestos, la justificación puede derivar de la propia imposibilidad material o económica de la intervención humana para realizar las operaciones necesarias. Por ejemplo, la retirada de contenidos ilícitos, o la alineación de modelos de inteligencia artificial generativa evidencian supuestos donde la necesidad de la automatización deriva de limitaciones estructurales que hacen impracticable o ineficaz la alternativa humana.

B) La habitación normativa de las decisiones automatizadas relevantes

La segunda excepción contemplada en el artículo 22.2.b) del RGPD habilita la adopción de decisiones automatizadas cuando estas estén autorizadas por el Derecho de la Unión o de los Estados miembros. Esta base de legitimación se orienta principalmente, aunque no exclusivamente, al sector público, donde el principio de legalidad exige en la mayoría de las ocasiones habilitación normativa expresa para la adopción de decisiones que afecten a los ciudadanos.

La aplicación de esta excepción requiere la concurrencia acumulativa de dos requisitos esenciales. En primer lugar, debe existir una norma de Derecho de la Unión Europea o del Estado miembro correspondiente que autorice expresamente al responsable del tratamiento a llevar a cabo la toma de decisiones automatizadas. En segundo lugar, esa misma norma habilitante debe contemplar medidas de garantía suficientes para la protección de los derechos e intereses de los afectados.

---

totalidad de las candidaturas. Grupo Artículo 29. *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679.* 2018, 26.

20. Agencia Española de Protección de Datos. Expediente N.º: EXP202211982, p.p 81 y 92.

La amplitud normativa respecto del contenido de las medidas adecuadas puede generar en muchos casos una inseguridad jurídica relevante, tanto para los Estados miembros que deben aprobar la legislación habilitante como para los responsables del tratamiento que deben aplicarla y los interesados que deben beneficiarse de las garantías. Esta indeterminación contrasta con la precisión con la que el RGPD regula otros aspectos del tratamiento de datos personales, sugiriendo que el legislador europeo optó deliberadamente por conferir un amplio margen de apreciación a los Estados miembros, esto es, poder definir los procesos decisorios automatizados de sus administraciones públicas.

Corresponde por tanto al legislador europeo o nacional establecer esas medidas de garantía adaptadas a cada derecho interno para la realidad que se pretenda regular. En este sentido, el TJUE recientemente ha aportado cierta claridad sobre este aspecto indicando que cuanto menos, las garantías que han de contemplarse para este tipo de tratamientos son las mismas que se prevén en el artículo 22.3. Estas son, el derecho a obtener intervención humana por parte del responsable, el derecho a expresar su punto de vista, y el derecho a impugnar la decisión[21]. Estas garantías están expresamente previstas en el RGPD para aquellas decisiones automatizadas relevantes que se legitimen a través de la necesariedad del contrato o del consentimiento explícito. Esta novedosa interpretación jurisprudencial resulta determinante porque establece un estándar mínimo armonizado aplicable en toda la Unión Europea, independientemente del contenido específico que cada Estado miembro incorpore en su legislación nacional[22].

Las garantías del artículo 22.3 del RGPD no constituyen, por tanto, meras recomendaciones o buenas prácticas, sino requisitos imperativos cuya ausencia en la normativa habilitante determinaría la inaplicabilidad de la excepción y, consecuentemente, la ilicitud de las decisiones automatizadas adoptadas al amparo de dicha normativa.

No obstante, la jurisprudencia del TJUE no cierra el debate sobre el contenido de las garantías adecuadas, sino que establece únicamente un umbral mínimo. Los Estados miembros conservan la facultad, y en muchos casos tendrán la obligación en virtud del principio de proporcionalidad, de establecer garantías adicionales más intensas cuando la naturaleza de las decisiones automatizadas o el contexto de su aplicación así lo requieran. Estas garantías complementarias podrían incluir evaluaciones de impacto obligatorias, auditorías algorítmicas periódicas, mecanismos de supervisión independiente, o requisitos específicos de explicabilidad y transparencia algorítmica que vayan más allá de las previsiones generales del RGPD.

21. STJUE Asunto C-634/21 Schufa de 7 de diciembre de 2023. FJ 65 y 66.

22. Es importante destacar que el TJUE también considera medidas necesarias las indicadas en el considerando 71 del RGPD. Estas son: obligación del responsable del tratamiento de utilizar procedimientos matemáticos o estadísticos adecuados, de aplicar las medidas técnicas y organizativas apropiadas para garantizar que se reduzca al máximo el riesgo de error y se corrijan errores, y de asegurar los datos personales de forma que se tengan en cuenta los posibles riesgos para los intereses y derechos del interesado e impedir, entre otras cosas, los efectos discriminatorios en las personas físicas. STJUE Asunto C-634/21 Schufa de 7 de diciembre de 2023. FJ 66.

C) El consentimiento explícito

La tercera excepción contemplada en el artículo 22.2.c) del RGPD permite la adopción de decisiones automatizadas cuando se fundamenten en el consentimiento explícito del interesado. Esta modalidad de consentimiento incorpora una cualificación reforzada respecto del consentimiento general previsto en el artículo 6.1.a) del RGPD como base ordinaria de legitimación para el tratamiento de datos personales. La exigencia de que el consentimiento sea «explícito» implica que no resulta suficiente una manifestación tácita o presunta de voluntad, sino que se requiere una declaración expresa e inequívoca del interesado autorizando específicamente su sometimiento a decisiones automatizadas.

Este reforzamiento del consentimiento responde a la especial trascendencia que revisten las decisiones automatizadas para los derechos e intereses del afectado. Mientras el consentimiento ordinario del artículo 6.1.a) puede manifestarse mediante una acción afirmativa clara, el consentimiento explícito exige una declaración escrita o verbal que no deje lugar a dudas sobre la voluntad del interesado.

La Agencia Española de Protección de Datos ha precisado que el consentimiento explícito en estos contextos no constituye una base de legitimación autosuficiente, sino que impone obligaciones adicionales al responsable del tratamiento. En particular, la AEPD ha considerado que el responsable debe ofrecer alternativas equivalentes y viables a la decisión automatizada, garantizando además que la elección del interesado de no ser sometido a dicha decisión automatizada no le genere perjuicio alguno[23].

El consentimiento explícito en este ámbito, así como toda la estructura del artículo 22 se construye sobre una premisa implícita de desconfianza hacia las decisiones automatizadas, configurándolas como una amenaza potencial para los derechos del interesado de la cual este debe ser protegido mediante ciertas garantías, además de bases de legitimación muy concretas y justificadas. Sin embargo, estas máquinas cada vez son más precisas, lo que puede llevar en un futuro a que las personas confiemos más en lo que dice el algoritmo respecto de lo indicado por los propios humanos. En este contexto, el consentimiento explícito del artículo 22.2.c) podría adquirir una dimensión diferente. No se trataría de que el interesado «acepte» ser sometido a una decisión automatizada a falta de alternativas, sino de que elija activamente la decisión automatizada porque la considera mejor para él.

V. GARANTÍAS DEL INTERESADO FRENTE A LAS DECISIONES AUTOMATIZADAS RELEVANTES

Una vez identificada la base de legitimación que habilita las decisiones automatizadas relevantes, resulta necesario analizar las garantías que prevé el artículo 22 para los supuestos en los que se lleve a cabo este tipo de tratamien-

23. AEPD. *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción*, febrero, 2020, p.28.

to. Así, el artículo 22 distingue entre las medidas aplicables a las decisiones basadas en ese consentimiento o la ejecución del contrato por un lado (art. 22.3), y las medidas aplicables que en su caso debe regular la norma europea o nacional que autorice dichas decisiones por otro (art.22.2b).

Como hemos dicho, pese a que el RGPD distingue entre un tipo de garantías u otras, el TJUE ya ha indicado la necesidad de homogenizar las garantías previstas para las tres excepciones que habilitan las decisiones automatizadas relevantes.

1. Las garantías expresas del artículo 22

El artículo 22.3 reconoce en favor de los interesados tres derechos concretos, esto son, la impugnación de la decisión automatizada, el derecho a obtener intervención humana tras la decisión y el derecho a que el interesado pueda expresar su punto de vista. Estas facultades no constituyen derechos aislados e independientes, sino manifestaciones específicas de un derecho general más amplio, este es, el derecho de audiencia o contradicción del interesado frente a la decisión automatizada que le afecta[24]. Bajo esta perspectiva integradora, el responsable del tratamiento debe establecer mecanismos adecuados que faciliten un trámite de audiencia donde el interesado pueda cumulativamente impugnar la decisión adoptada por el sistema, tener contacto efectivo con una persona perteneciente a la organización que utiliza el algoritmo, y expresar su punto de vista en relación con la decisión adoptada.

El despliegue efectivo de estas garantías no responde a una secuencia temporal predefinida ni a una forma procedimental única aplicable universalmente. El responsable del tratamiento debe adaptar la implementación de estas garantías al contexto específico en el que opera el sistema de decisiones automatizadas, considerando factores como la naturaleza de la decisión, su gravedad para el interesado, el sector de actividad, y las características del público afectado.

Por lo que se refiere al derecho a solicitar la intervención humana, éste se sitúa temporalmente tras la adopción de la decisión y a la producción de sus efectos sobre el interesado. A través de este derecho, un miembro cualificado de la organización que ha implementado el algoritmo puede valorar los posibles errores que el sistema haya podido generar, atendiendo a las circunstancias específicas del caso concreto y a la información adicional que el interesado pueda aportar.

La efectividad de este derecho depende críticamente de que la persona que realice la revisión posea capacidad real de modificar la decisión automatizada. Si la intervención humana se limita a verificar formalmente que el sistema ha funcionado correctamente sin poder alterar materialmente el resultado, o si sistemáticamente ratifica las decisiones algorítmicas sin análisis sustantivo, el

24. Palma Ortigosa, A. *Decisiones automatizadas y protección de datos. Especial atención a los sistemas de inteligencia artificial*, Dykinson, Madrid, 2022, p.286.

derecho queda vaciado de contenido. La supervisión humana debe ser significativa, no meramente simbólica, debiendo el supervisor disponer de la competencia, formación y autoridad necesarias para cuestionar y, en su caso, rectificar el output algorítmico. Resulta por tanto adecuado tomar como referencia las recomendaciones que previamente hemos indicado sobre la presencia humana significativa y real durante el proceso decisorio donde intervienen algoritmos[25].

2. Otras garantías aplicables a las decisiones automatizadas relevantes

Además de los derechos reconocidos en el artículo 22 en favor del interesado, el RGPD contempla otra serie de garantías presentes a lo largo de su articulado. Así, encontramos por un lado los deberes de información que corresponde al responsable del tratamiento sobre ciertos elementos relacionados con las decisiones automatizadas. Por otro lado, se reconoce el derecho de acceso que tiene el titular de los datos a acceder a cierta información sobre dichas decisiones automatizadas.

Tanto una como la otra garantía permiten al interesado conocer la existencia de esas decisiones automatizadas, así como a obtener información significativa sobre la lógica aplicada y las consecuencias previstas de éstas.

El TJUE recientemente ha indicado que por información significativa hay que entender toda información pertinente sobre el procedimiento y los principios relativos al uso, por medios automatizados, de datos personales con vistas a la obtención de un resultado específico[26].

A su vez, conviene destacar que en esta misma sentencia de 2025 el TJUE ha reconocido un derecho de explicación en favor del interesado a conocer el funcionamiento del mecanismo aplicado en la adopción de una decisión automatizada de la que ha sido objeto y sobre el resultado al que ha llegado dicha decisión[27].

La existencia de un derecho de explicación hasta ahora se había relacionado con las garantías previstas en el artículo 22.3 y el considerando 71 del RGPD, el cual expresamente lo reconocía. No obstante, la doctrina y las autoridades de protección de datos se encontraban divididas con relación a si realmente existía o no tal derecho debido a que los considerandos no forman parte del derecho vinculante de las normas europeas, sino que su función se centra en aclarar e interpretar las normas[28].

---

25. Ya hemos indicado previamente que para evitar la aplicación del artículo 22, necesariamente se deberá justificar adecamente que la presencia del humano durante el proceso decisorio es significativa.

26. STJUE Asunto C-203/22 de 27 de diciembre de 2025. FJ 43 y 50.

27. STJUE Asunto C-203/22 de 27 de diciembre de 2025. FJ 43 y 50.

28. En contra del derecho a la explicación. Wachter, Mittelstadt,B. & Floridi,L. «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation». *International Data Privacy Law*, Volume 7, Issue 2, 1 May 2017, p.p. 76 a 99. A favor del

Resulta por tanto interesante como el TJUE, a partir del derecho de acceso y del propio considerando 71 reconoce un derecho de explicación que el mismo tilda de genuino. En este sentido, el propio tribunal indica que a través de la explicación que se ha de facilitar mediante el ejercicio del derecho de acceso los interesados afectados por los procesos decisorios totalmente automatizados pueden ejercer de manera eficaz los derechos que le reconoce el artículo 22.3 de dicho Reglamento, a saber, el de expresar su punto de vista sobre esa decisión y el de impugnarla. Es decir, visto así, el derecho de acceso en relación con las decisiones automatizadas relevantes se convierte en un derecho instrumental para ejercer el resto de las garantías que también se prevén en este contexto de plena automatización.

Para terminar, el TJUE deja claro que, a pesar de que se puede limitar el acceso a cierta información que compromete los secretos comerciales o la propiedad intelectual presente en estos sistemas algorítmicos, el responsable del tratamiento nunca se puede negar a facilitar cierta información relativa al proceso decisorio y las razones de éste[29].

VI. EL ARTÍCULO 22 Y LOS MODELOS DE INTELIGENCIA ARTIFICIAL GENERATIVA

El artículo 22 del RGPD fue concebido en un contexto tecnológico donde predominaba como mucho la elaboración de perfiles cuyos resultados en determinadas ocasiones podían llegar a automatizarse. En ese tiempo las herramientas tecnológicas que dominaban se centraban esencialmente en sistemas de machine learning tradicionales, así como en algoritmos deterministas basados en patrones identificados en conjuntos de datos estructurados. Los modelos de clasificación, regresión o clustering que inspiraron la regulación operaban mediante lógicas relativamente lineales, esto es, input de datos, procesamiento algorítmico, output decisorio. Este paradigma permitía conceptualizar la «decisión automatizada» como un acto identificable temporalmente, con una lógica subyacente más o menos describible, y con un resultado específico atribuible al sistema. Es decir, la operatoria de la lógica del artículo 22 y todas las garantías que hemos explicado encajan en esta realidad.

A pesar de que estos sistemas de IA tradicionales o algoritmos determistas se siguen utilizando y se utilizarán en el futuro, la irrupción de modelos de inteligencia artificial generativa, particularmente los Large Language Models (LLMs) como GPT-4, Claude, Gemini o LLaMA, y su evolución hacia agentes autónomos capaces de planificar, usar todo tipo de herramientas y ejecutar tareas complejas, desafía radicalmente este marco conceptual. Estos sistemas no adoptan «decisiones» en el sentido tradicional que contempla el artículo 22,

derecho a la explicación: Goodman,B. & Flaxman,F. «EU Regulations on Algorithmic Decision-Making and a right to Explanation», 2016.

29. STJUE Asunto C-203/22 de 27 de diciembre de 2025. FJ 43 y 50.

sino que generan outputs lingüísticos, visuales o multimodales que pueden influir, condicionar o determinar decisiones humanas o acciones de otros sistemas sin que exista un momento decisorio claramente identificable ni una lógica subyacente plenamente explicable. Sin lugar a dudas que el funcionamiento de estos sistemas genera procedimientos totalmente automatizados, y, por tanto, de forma amplia, «decisiones», pero la pregunta sería si esas salidas, que se utilizan hoy día por todo tipo de empresas y administraciones públicas, pueden dar lugar a efectos relevantes como tal sobre un particular específico.

Se puede dar la paradoja por tanto de que el artículo 22 del RGPD, el cual está empezando a descubrirse ahora por parte de los operadores jurídicos, corra el riesgo de que se vea ampliamente superado por el desarrollo de una tecnología que avanza a pasos agigantados de forma prácticamente diaria.

VII. CONCLUSIONES

El artículo 22 del RGPD representa la culminación de una evolución normativa que se inicia con la Ley francesa de 1978 y continúa con la Directiva 95/46/CE. Sin embargo, solo en los últimos años este precepto ha adquirido verdadera relevancia práctica, impulsado por el desarrollo exponencial de la inteligencia artificial, la creciente sensibilización sobre los riesgos algorítmicos y la ausencia de regulación específica alternativa a la normativa sobre protección de datos.

El ámbito de aplicación del artículo 22 se define por dos requisitos acumulativos, estos son: la plena automatización del proceso decisorio y la generación de efectos jurídicos o significativamente similares sobre el interesado. Contrariamente a interpretaciones restrictivas, la elaboración de perfiles no constituye un elemento esencial para la aplicación del precepto, pudiendo existir decisiones totalmente automatizadas relevantes sin perfilado previo. La jurisprudencia del TJUE, especialmente la sentencia del caso SCHUFA (C-634/21), ha ampliado significativamente el concepto de decisión automatizada, estableciendo que la generación de un valor de probabilidad que condiciona de manera determinante la decisión adoptada por un tercero constituye por sí misma una decisión automatizada relevante. Esta interpretación extensiva responde a la necesidad de evitar la elusión del artículo 22 mediante estructuras decisorias fragmentadas entre múltiples responsables del tratamiento.

El precepto establece un sistema de prohibición general con excepciones tasadas que invierte la lógica habitual del RGPD. Las tres bases de legitimación habilitantes —necesidad contractual, autorización legal y consentimiento explícito— deben interpretarse restrictivamente y exigen justificación cualificada por parte del responsable del tratamiento. La reciente jurisprudencia del TJUE ha establecido que las garantías mínimas aplicables a todas las excepciones deben incluir los derechos del artículo 22.3, estos son: intervención humana significativa, expresión del punto de vista del interesado e impugnación de la decisión. Estas garantías no constituyen derechos aislados, sino manifestacio-

nes de un derecho general de audiencia o contradicción frente a la decisión automatizada.

El TJUE, en sentencia de 27 de febrero de 2025 (asunto C-203/22), ha reconocido un derecho genuino de explicación derivado del artículo 15.1.h) del RGPD, zanjando el debate doctrinal sobre su existencia. Este derecho tiene naturaleza instrumental, facilitando el ejercicio efectivo de las demás garantías.

No obstante, el artículo 22 fue concebido para sistemas algorítmicos tradicionales con lógicas relativamente lineales y decisiones identificables temporalmente. La irrupción de modelos de inteligencia artificial generativa y agentes autónomos desafía radicalmente este marco conceptual, generando outputs que pueden influir o determinar acciones sin un momento decisorio claramente identificable ni una lógica plenamente explicable. Esta evolución plantea interrogantes fundamentales sobre la aplicabilidad y suficiencia del artículo 22 para regular adecuadamente estos nuevos paradigmas de automatización.

BIBLIOGRAFÍA

AEPD. *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción*, febrero, 2020.

Arroyo Amayuelas, E. «El *scoring* de Schufa», *InDret*, 3, 2024.

Comisión Europea. *Comunicación de la Comisión «Guía azul» sobre la aplicación de la normativa europea relativa a los productos*, 2022.

Future of Privacy Forum. *Automated Decision-Making Under the GDPR: Practical Cases from Courts and Data Protection Authorities*, 2022.

Genicot, N. «Scoring the European citizen in the AI era», *Computer Law & Security Review: The International Journal of Technology Law and Practice,* 57, 2025.

Grupo Artículo 29. *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679.* 2018.

Goodman,B. & Flaxman,F. «EU Regulations on Algorithmic Decision-Making and a right to Explanation», 2016.

Huego Lora, A. «El uso de algoritmos y su impacto en los datos personales». *Revista de Derecho Administrativo,* 20, 2021.

Mondall, B, «Artificial Intelligence: State of the Art», en V. Balas; R. Kumar; R. Srivastava (eds), *Recent Trends and Advances in Artificial Intelligence and Internet of Things.* Springer, Cham, vol 172, 2020.

Supervisor Europeo de Protección de datos.TechDispatch #2/2025 - Human Oversight of Automated Decision-Making, 2025.

Palma Ortigosa, A. *Decisiones automatizadas y protección de datos. Especial atención a los sistemas de inteligencia artificial*, Dykinson, Madrid, 2022.

Radtke, «Human Oversight under the AI Act and its interplay with Art. 22 GDPR», en Raue/von Ungern-Sternberg/Kumkar/Rüfner(ed.), Artificial Intelligence and Fundamental RightsThe AI Act of the European Union

and its implications for global technology regulation, Verein für Recht und Digitalisierung, Trier, 2025.

WACHTER, MITTELSTADT,B. & FLORIDI,L. «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation». *International Data Privacy Law*, Volume 7, Issue 2, 1 May 2017.

La interacción entre la ley de servicios digitales, la carta de derechos fundamentales de la UE y los derechos fundamentales reconocidos en las constituciones española y portuguesa[1]

María Dolores Montero Caro
Profesora de Derecho constitucional
Universidad de Córdoba

SUMARIO: I. INTRODUCCIÓN. II. DERECHOS FUNDAMENTALES Y ENTORNO DIGITAL: FUNDAMENTOS TEÓRICOS. III. MARCO JURÍDICO-CONSTITUCIONAL APLICABLE. 1. Constitución española. 2. Constitución portuguesa. IV. A PROPÓSITO DE LA INCIDENCIA DEMOCRÁTICA DE LA INTELIGENCIA ARTIFICIAL. V. CONCLUSIONES. *BIBLIOGRAFÍA*.

1. Esta investigación se ha realizado en el marco de una estancia de investigación realizada en el Centro de Investigación en Derecho Público (CIDP) del Instituto de Ciencias Jurídico-Políticas de la Universidad de Lisboa (Portugal) en 2025, financiada por el proyecto de I+D+i *Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas* [2023-2026] (PID2022-136439OB-I00), financiado por MCIN/AEI/10.13039/501100011033/ y «FEDER Una manera de hacer Europa». Asimismo, este trabajo se inserta en la actividad investigadora del Grupo de Investigación SEJ-372 de la Junta de Andalucía «Democracia, Pluralismo y Ciudadanía», del que la autora forma parte.

I. INTRODUCCIÓN

Lo primero que debemos indicar es que la Ley de Servicios Digitales (*Digital Services Act*, en adelante, DSA)[2] marca un hito en la regulación europea de las plataformas en línea, con implicaciones directas sobre los derechos fundamentales, destacando en ese sentido su objetivo declarado, que no es otro que el de «garantizar un entorno en línea seguro y responsable», previniendo contenidos ilícitos y la desinformación, a la vez que protege la seguridad de los usuarios, sus derechos fundamentales y un espacio digital abierto y justo. En esencia, la DSA busca reequilibrar las responsabilidades de usuarios, plataformas y autoridades «en función de los valores europeos, situando a los ciudadanos en el centro», mediante normas claras y proporcionadas para proteger derechos y fomentar la innovación.

Ahora bien, la aplicación de este amplio marco normativo plantea interrogantes jurídico-constitucionales en los Estados miembros, por supuesto en países como España y Portugal, objeto de análisis del presente estudio y cuyas Constituciones consagran firmemente derechos como la libertad de expresión, la privacidad, la protección de datos personales, la igualdad y no discriminación, el debido proceso, así como el principio de proporcionalidad en las restricciones de derechos. A su vez, la Carta de Derechos Fundamentales de la UE (en adelante, CDFUE), que es jurídicamente vinculante para las instituciones de la Unión y los Estados en la aplicación del Derecho de la UE, establece estándares que la DSA debe respetar y que enmarcan su interpretación[3]

Así, la interacción de esta normativa, su estudio y análisis confluyen de manera necesaria para visualizar, desde un prisma constitucional, todas las relevantes afecciones de estos derechos fundamentales.

II. DERECHOS FUNDAMENTALES Y ENTORNO DIGITAL: FUNDAMENTOS TEÓRICOS

En las últimas décadas, la revolución digital ha desafiado las categorías jurídicas tradicionales de los derechos fundamentales, dando lugar a lo que algunos autores denominan «proceso de constitucionalización digital» o incluso la

2. Reglamento UE 2022/2065. «DOUE» núm. 277, de 27 de octubre de 2022, páginas 1 a 102 (102 págs.). DOUE-L-2022-81573.

3. Tal y como queda patente con la lectura del considerando 153 de la DSA, que reza lo siguiente: «El presente Reglamento respeta los derechos fundamentales reconocidos en la Carta y los derechos fundamentales que constituyen principios generales del Derecho de la Unión. En consecuencia, el presente Reglamento ha de interpretarse y aplicarse de conformidad con tales derechos fundamentales, incluida la libertad de expresión e información, y la libertad y el pluralismo de los medios de comunicación. En el ejercicio de las competencias establecidas en el presente Reglamento, todas las autoridades públicas implicadas deben alcanzar, en situaciones en las que los derechos fundamentales pertinentes entren en conflicto, un equilibrio justo entre los derechos afectados, de conformidad con el principio de proporcionalidad».

gestación de una «nueva constitución europea digital»[4]. Así, el rápido desarrollo de plataformas de redes sociales, buscadores, servicios en la nube y mercados en línea —servicios hoy esenciales en la vida cotidiana— ha amplificado la capacidad de ejercer derechos como la libertad de expresión y el acceso a la información, pero también ha creado nuevas vulnerabilidades para derechos como la intimidad, la protección de datos o la dignidad. Los mismos valores fundamentales subyacentes a las libertades clásicas deben protegerse en este entorno tecnológico, a menudo frente a actores privados con un poder sin precedentes de moderar contenidos y procesar datos personales.

Esta realidad ha impulsado el desarrollo de conceptos teóricos específicos, como el de «derechos digitales», y ha motivado la adaptación del Derecho constitucional y europeo[5]. En España, por ejemplo, se formuló en 2021 una Carta de Derechos Digitales[6] para orientar la protección de derechos en Internet, reconociendo principios de libertad de expresión en línea, privacidad digital, gobierno algorítmico responsable, etc. Por su parte, la Unión Europea ha respondido con un marco regulatorio integral —del cual la DSA forma parte junto con el Reglamento de Servicios Digitales de Mercados (en adelante, DMA) y propuestas sobre inteligencia artificial— que busca armonizar la salvaguarda de derechos fundamentales en la esfera digital a nivel continental. La DSA, en particular, se concibe como «un paso dentro de la nueva constitución digital europea», creando una estructura multinivel de supervisión para asegurar que los servicios digitales cumplan diligentemente sus obligaciones «especialmente el respeto de los derechos fundamentales» de los ciudadanos[7].

A este respecto, un elemento teórico central es entender cómo se aplican y limitan los derechos fundamentales en Internet. Por ello, la jurisprudencia y doctrina han enfatizado que Internet no es un «espacio ajeno» al Derecho, sino un ámbito donde rigen las mismas libertades y límites que offline, si bien con matices propios. En este sentido, el Tribunal Constitucional español, en su primera gran sentencia[8] sobre libertad de expresión en redes sociales[9], afirmó

4. Afonso, G. «Primeiras Notas sobre a Coordenação dos Serviços Digitais no Regulamento dos Serviços Digitais; em especial, o caso português», *e-Publica: public law journal*, vol. 11 núm. 3, 2024, pp. 71-102.

5. Para un estudio en profundidad sobre el funcionamiento de las redes sociales, la aplicación de IA para la toma de decisiones y su impacto en los derechos sociales y, en definitiva, en la democracia vid. Galdámez Morales, A. *De las tecnologías disruptivas y su ordenación jurídica: proceso, naturaleza y sistema constitucional de la comunicación digital*. Tesis doctoral, Universidad de Sevilla, 2024.

6. Pese a su carácter no normativo, tiene una influencia clara en el ajuste a la «digitalidad» que progresivamente va operando en el ordenamiento jurídico español. Para un análisis exhaustivo de la Carta vid. Cotino Hueso, L. (coord.). *La Carta de Derechos Digitales*, Tirant lo Blanch, Valencia, 2022.

7. Afonso, G. «Primeiras Notas sobre a Coordenação dos Serviços Digitais no Regulamento dos Serviços Digitais; em especial, o caso português», *op. cit.*

8. STC 8/2022, de 27 de enero de 2022.

9. Cotino Hueso, L. «La primera sentencia general del Tribunal Constitucional sobre la libertad de expresión e información en Internet: Seguimos pendientes de muchos temas clave para el futuro», en M.V. Álvarez Buján (coord.); P. Simón Castellano (dir.), *Evolución e interpretación del TC*

categóricamente que «no cabe duda de que las libertades de comunicación —libertad de información y libertad de expresión— también se ejercitan a través de las herramientas que facilita internet [en este sentido SSTEDH de 18 de diciembre de 2012, asunto *Ahmet Yildirim c. Turquía*, § 48, y de 10 de marzo de 2009, asunto *Times Newspapers Ltd. c. Reino Unido* (núms. 1 y 2), § 27], como lo son las redes sociales, siendo susceptibles de verse limitadas por el poder público allí donde se prevén también límites para el ejercicio de estas fuera del contexto de internet»[10]. Igualmente, consideró Internet como «una herramienta sin precedentes para el ejercicio de la libertad de expresión», cuya protección como medio forma parte del derecho[11]. Este reconocimiento implica que los derechos fundamentales tienen una dimensión tecnológica, por lo que proteger la libertad de expresión hoy implica proteger la arquitectura abierta de Internet y prevenir censuras u obstáculos indebidos en línea; proteger la privacidad supone controlar la vigilancia masiva y el flujo global de datos personales; garantizar la no discriminación incluye vigilar posibles sesgos algorítmicos en plataformas; asegurar el debido proceso exige introducir garantías de justicia procedimental en la moderación de contenidos (por ejemplo, que un usuario pueda ser oído y recurrir si su contenido es eliminado).

Otro concepto relevante es el de la «proporcionalidad digital»: dado el inmenso impacto que puede tener cualquier restricción en línea (por el alcance potencialmente global e inmediato de Internet), se requiere un escrutinio estricto de la necesidad y proporcionalidad de las medidas que limiten derechos. En este sentido el TC español ha señalado que las redes sociales multiplican los riesgos tanto para los derechos de la personalidad como para las libertades, por la «inmediatez y rapidez» de la difusión, debiendo el juicio de proporcionalidad considerar esas particularidades[12]. De igual forma, el Tribunal Constitucional portugués, al evaluar medidas de vigilancia digital, subraya que la mayor eficacia de una técnica intrusiva no la hace automáticamente constitucional, pues «no todo lo que resulte especialmente eficaz… quedará por ello justificado constitucionalmente» si sacrifica en exceso derechos básicos[13]. Así, la protección efectiva del derecho fundamental a la protección de datos personales requiere que

sobre derechos fundamentales y garantías procesales: cuestiones recientemente controvertidas: Análisis de sus últimos pronunciamientos más reseñables, Aranzadi, Cizur Menor (Navarra), 2023, pp. 45-69.

10. García Majado, P. «Libertades comunicativas y redes sociales: a propósito de la STC 8/2022, de 27 de enero de 2022», *Revista General de Derecho Constitucional*, núm. 37, 2022, p. 2.

11. La referencia es una conocida afirmación del Tribunal Europeo de Derechos Humanos (en adelante, TEDH) en la sentencia del caso Ahmet Yıldırım contra Turquía (2012), donde el TEDH afirmó explícitamente que «Internet se ha convertido hoy día en uno de los principales medios mediante los que las personas ejercen su derecho a la libertad de expresión e información [...], proporcionando una herramienta esencial para participar en actividades y discusiones sobre cuestiones políticas y asuntos de interés general». STEDH, núm. 3111/10, 18 de diciembre de 2012, párr. 54.

12. STC 8/2022, FJ 3º.

13. Tribunal Constitucional de Portugal. 19 de abril de 2022. *Acórdão n.º 268/2022*. Recuperado de https://www.tribunalconstitucional.pt/tc/acordaos/20220268.html

cualquier interferencia estatal cumpla estrictamente con el principio de proporcionalidad[14]. De este modo se colige que la teoría constitucional contemporánea postula que los estándares garantistas tradicionales (legalidad, necesidad, proporcionalidad, control judicial, etc.) deben aplicarse con igual o mayor vigor en el entorno digital, evitando que la novedad tecnológica debilite las salvaguardas de los individuos.

Como hemos apuntado supra, la DSA, desde su gestación, ha estado informada por estas consideraciones. De hecho, sus disposiciones generales enfatizan la primacía de los derechos fundamentales. El Reglamento declara que «respeta los derechos fundamentales reconocidos por la Carta» y que deberá interpretarse en consecuencia, incluida la libertad de expresión y de información, así como la libertad y pluralismo de los medios, debiendo las autoridades y plataformas lograr un justo equilibrio entre derechos en conflicto conforme al principio de proporcionalidad. De igual modo, la DSA obliga expresamente a las plataformas a respetar los derechos de los usuarios cuando apliquen sus términos de servicio y moderen contenidos, introduciendo obligaciones de diligencia, transparencia y no discriminación en esas decisiones[15].

Esta norma, de aplicación directa en los Estados miembros, supone una actualización profunda del marco jurídico del entorno digital. Su objetivo principal es aumentar la responsabilidad de las grandes plataformas en línea y garantizar una mayor transparencia en el manejo de contenidos. Entre sus disposiciones destaca la obligación de que las plataformas evalúen y mitiguen riesgos relacionados con la difusión de contenido ilícito o dañino, la protección de los derechos fundamentales —como la libertad de expresión, la privacidad y la no discriminación— y la preservación del pluralismo informativo. Además, la DSA promueve la creación de mecanismos de supervisión independientes y exige a las empresas tecnológicas una mayor rendición de cuentas ante las autoridades y los usuarios, buscando un equilibrio entre la libertad digital y la seguridad informativa.

Como señala Cotino, la DSA supone «un nuevo enfoque que permite garantizar al mismo tiempo la libertad de información y limitar bastante los peligros de que los gobiernos metan sus narices» en la moderación de contenidos, instaurando mecanismos de transparencia algorítmica sin precedentes (como el Centro de Transparencia Algorítmica de la UE) y creando vías de reclamación para los usuarios afectados por decisiones de las plataformas[16].

Por otra parte, el auge de las redes sociales como fuente principal de información ha favorecido la expansión de desinformación, lo que ha impulsado la aparición de plataformas de verificación o *fact-checkers*. Sin embargo, la lucha contra este fenómeno debe mantenerse vigilante para evitar que los esfuerzos por

14. *Ibídem*, Acórdão n.º 268/2022, 2022, párr. 68

15. Art. 14 DSA.

16. Cotino Hueso, L. «Entre bytes y democracia. Cómo la DSA de la UE armoniza la libertad de expresión y la lucha contra la desinformación», *OTROSÍ.: Revista del Colegio de Abogados de Madrid*, núm. 2, 2024, pp. 42-44.

frenar la manipulación se conviertan en herramientas de censura. En este contexto, la Unión Europea ha desarrollado diversas estrategias para hacer frente al fenómeno. Desde 2015, con las conclusiones del Consejo Europeo, se han impulsado iniciativas como la creación del grupo de expertos de alto nivel (HLEG) y el Plan de Acción contra la Desinformación, que estableció el Sistema de Alerta Rápida (RAS) para mejorar la coordinación entre los Estados miembros y detectar campañas de manipulación. Posteriormente, el Plan de Acción para la Democracia Europea, aprobado en 2020, reforzó la necesidad de proteger la integridad de los procesos democráticos frente a la influencia de la desinformación.

La regulación de la desinformación es un asunto complejo pues, aunque resulta esencial para proteger a la sociedad del impacto de las noticias falsas, también puede transformarse en un instrumento de control si se otorga a gobiernos o plataformas el poder de decidir qué información es veraz. Esta situación puede derivar en la censura de opiniones críticas bajo el pretexto de combatir la desinformación, lo que atentaría contra el derecho fundamental a la libertad de expresión. Por ello, la regulación debe diseñarse con equilibrio, garantizando la protección frente a la manipulación informativa sin restringir el pluralismo ni el debate democrático.

Estamos, por tanto, ante un intento regulatorio de armonizar el ecosistema digital con los valores constitucionales, cuyo éxito dependerá en gran medida de cómo se implemente y aplique esta norma en cada país, cuestión que exige atender al entramado constitucional interno.

III. MARCO JURÍDICO-CONSTITUCIONAL APLICABLE

El análisis de los desafíos constitucionales derivados de la DSA en España y Portugal requiere identificar, en primer lugar, los preceptos relevantes de ambas constituciones nacionales, así como las disposiciones específicas de la Carta de Derechos Fundamentales de la Unión Europea relacionadas con la libertad de expresión, la privacidad, la protección de datos personales, la no discriminación, el debido proceso y el principio de proporcionalidad. Este marco normativo constituye el estándar constitucional frente al cual deberá evaluarse la implementación y aplicación efectiva de la DSA.

Aunque en un primer momento en que el uso de internet no estaba generalizado se prefirió por un modelo cuyas únicas reglas a seguir eran las propias del comercio[17], pronto, dada la expansión y potencial de la red, se optó por una regulación amplia en la que, además de los Estados y de los organismos supranacionales, se dejaba un amplio margen reservado a la autorregulación por parte de las compañías tecnológicas. Es aquí, por ejemplo, donde encontramos

17. En un primer momento la moderación de contenidos en internet se regía por las reglas recogidas en la Directiva sobre Comercio Electrónico. (Directiva 2000/31/CE del Parlamento Europeo y del Consejo, de 8 de junio de 2000, relativa a determinados aspectos jurídicos de los servicios de la sociedad de la información, en particular el comercio electrónico en el mercado. DOCE, núm. 178, de 17 de julio de 2000).

la legitimación de las propias empresas para imponer su propio sistema de moderación de contenidos, siendo más rígidos por ejemplo en las plataformas TikTok y Meta (Instagram y Facebook) y más flexible en el caso de la red social X (antes Twitter).

Del mismo modo, la proliferación de las conocidas popularmente como *fake news*[18] o campañas de desinformación, muchas de ellas provocadas conscientemente por poderes políticos y económicos, se han convertido en una cuestión de seguridad nacional. Así lo pone de manifiesto, a nivel europeo, la aprobación en 2018 del Plan de Acción contra la Desinformación, la creación de un Sistema de Alerta Rápida (RAS) que detecte posibles campañas y actividades de desinformación y, más recientemente la aprobación de la DSA como norma directamente aplicable a los Estados miembros que supone un marco de actuación básico de las plataformas en línea, evaluando los riesgos sistémicos. Asimismo, en el ámbito estatal, España confirmó su preocupación por la desinformación, incluyendo las campañas de desinformación en la Estrategia de Seguridad Nacional de 2021[19]. A medida que la posibilidad de acceder a enormes volúmenes de información ha crecido de forma exponencial, también se han incrementado los peligros vinculados con la desinformación, las noticias falsas y la posverdad en el entorno digital. Frente a este escenario comunicativo actual, la moderación de contenidos en línea se ha vuelto un elemento clave para enfrentar dichos desórdenes[20].

1. Constitución española

La Constitución Española de 1978 (en adelante, CE) reconoce un amplio catálogo de derechos fundamentales, varios de los cuales resultan directamente involucrados por la regulación de los servicios digitales. En primer lugar, abordaremos lo relativo a la libertad de expresión e información (art. 20 CE). El art. 20.1 a) consagra el derecho a «expresar y difundir libremente los pensamientos, ideas y opiniones», por cualquier medio de difusión, y el art. 20.1 d) el derecho a comunicar y recibir información veraz por cualquier medio. Además, el art. 20.2 CE prohíbe la censura previa y el 20.5 CE exige que solo pueda acordarse el secuestro de publicaciones en virtud de resolución judicial. Esta configuración, fruto de la experiencia histórica, establece fuertes garantías contra intervencio-

18. Aunque el término más popular ha sido *fake news* o noticias falsas, en el ámbito del periodismo se rechaza firmemente esta expresión, ya que, si una información es falsa, no puede considerarse una noticia. Por más que lo aparente, carece de veracidad y no se sustenta en datos ni hechos comprobables. (Montero Caro, M.D., *Democracia en transición. Una agenda para su regeneración.* Dykinson, Madrid, 2023, p. 60)

19. Que ha sido ratificada y ampliada en la reunión del 28 de enero de 2025, del Consejo de Seguridad Nacional, en virtud del cual se aprobó el Acuerdo de elaboración de la Estrategia nacional contra las campañas de desinformación.

20. Sentí Navarro, C., «Desórdenes informativos en línea y la moderación de contenidos en la era digital», *Revista de Derecho Político.* Núm. 121, 2024, p. 227.

nes estatales arbitrarias sobre la libertad de expresión. En consecuencia, la jurisprudencia constitucional española ha desarrollado una doctrina protectora de la libre expresión como pilar del sistema democrático, aunque reconociendo sus límites en la protección de otros derechos (honor, intimidad, etc.). En el ámbito digital, como se indicó *supra*, el TC ha dejado claro que Internet y las redes sociales son también ámbitos protegidos por el art. 20 CE, debiendo las eventuales restricciones sujetarse a las exigentes condiciones constitucionales[21].

Respecto del derecho a la intimidad y privacidad (art. 18.1 CE), la Constitución garantiza el derecho a la intimidad personal y familiar y el secreto de las comunicaciones (art. 18.3 CE), protegiendo un ámbito privado inmune a intromisiones injustificadas. En contextos digitales, esto abarca la vida privada en línea, comunicaciones electrónicas (mensajería, emails) y la protección frente a vigilancias indiscriminadas. Así, el TC ha equiparado las comunicaciones telemáticas a las tradicionales a efectos de tutela del secreto[22]. Y, por su parte, el derecho a la intimidad se ve potencialmente afectado por las medidas de vigilancia de contenidos como pueden ser los escaneos automatizados de mensajes en busca de ilícitos o por la exposición pública de datos personales en plataformas; por tanto, cualquier restricción derivada de la DSA, como órdenes de retirar cierta información privada difundida ilícitamente, deberá valorar este derecho.

En cuanto al derecho fundamental a la protección de datos personales, la CE no menciona expresamente un «derecho a la protección de datos», pero el art. 18.4 CE dispone que «la ley limitará el uso de la informática para garantizar el honor y la intimidad personal y familiar de los ciudadanos y el pleno ejercicio de sus derechos». A partir de este mandato, la jurisprudencia (especialmente la STC 292/2000) ha reconocido un derecho fundamental autónomo a la protección de datos distinto del derecho a la intimidad y, según dicha sentencia, este derecho —implícito en el art. 18.4 CE e instrumentalizado mediante leyes orgánicas— otorga a los ciudadanos un «poder de disposición y control sobre sus datos personales (*habeas data*)», abarcando todo tipo de datos personales, incluso no íntimos, cuyo conocimiento o uso por terceros pueda afectar a sus derechos. Se trata de un derecho a imponer a terceros comportamientos (de abstención o diligencia) respecto a nuestros datos, de manera que ningún dato personal pueda ser tratado sin control. Este contenido se refleja actualmente en la legislación orgánica[23] y se ha visto fortalecido por el RGPD europeo, lo cual es especialmente relevante porque la protección de datos incide de lleno en la DSA: la regulación del procesamiento de datos de usuarios por las plataformas (perfiles, algoritmos de recomendación, publicidad dirigida) debe compaginarse con este derecho. España, al reconocer constitucionalmente la protección de datos, sitúa un listón

21. Cabe mencionar que en años recientes el TC ha empezado a pronunciarse sobre casos específicamente vinculados a Internet. Así encontramos la STC 13/2020 y la STC 93/2021, entre otras, que abordaron la colisión entre expresiones en redes sociales y derechos al honor, y la STC 8/2022 que fijó por primera vez principios generales sobre libertad de expresión en línea.

22. Exigiendo mandato judicial para acceder al contenido de correos electrónicos, como se observa en las SSTC 96/2012 y 170/2013.

23. Ley Orgánica 3/2018 de Protección de Datos y Garantía de Derechos Digitales.

alto ya que, por ejemplo, la utilización de datos personales para moderación algorítmica o para personalizar contenidos no puede vulnerar ese poder de control del individuo sobre la información que le concierne. Como dijo el TC, «el objeto del derecho de protección de datos no son solo los datos íntimos, sino cualquier dato personal —incluso público— cuya utilización sin consentimiento pueda suponer una amenaza para el individuo», incluyendo datos que permitan perfilar aspectos ideológicos, raciales, sexuales, etc. del sujeto. Esto enlaza con la preocupación por algoritmos discriminatorios o por la microsegmentación publicitaria, asuntos que la DSA aborda en parte prohibiendo el uso de datos sensibles para anuncios y la publicidad dirigida a menores[24].

En este sentido cabe destacar también la anulación por parte del TC del artículo 58. bis). 1. de la LOREG en virtud del cual se permitía a los partidos políticos recoger datos personales relacionados con las opiniones políticas de los ciudadanos. El TC consideró que la falta de garantías suficientes contravenía el art. 53.1. CE pues, tal y como acertadamente señala Jove Villares la ausencia de un marco claro de garantías ya podría suponer una vulneración del propio derecho fundamental a la protección de datos[25].

2. Constitución portuguesa

Al igual que la Constitución española, la Constitución de la República Portuguesa de 1976 (CRP, en adelante) consagra un catálogo de derechos fundamentales, aunque con ciertas diferencias importantes. En particular, el art. 37 de la CRP garantiza la libertad de expresión e información, afirmando que el ejercicio de estos derechos no puede ser impedido ni restringido por ningún tipo de censura previa. Esta prohibición de censura coincide con el modelo constitucional español, previsto en el art. 20.1 y 20.2 CE, si bien la Carta portuguesa añade explícitamente garantías como el derecho de réplica y rectificación para los afectados por informaciones inexactas. Asimismo, el art. 34 de la CRP consagra la inviolabilidad del domicilio y del secreto de las comunicaciones, reforzando la protección de la privacidad en términos análogos a los del art. 18 CE. Cabe destacar, además, que Portugal fue pionero en la constitucionalización de la protección de datos personales: el artículo 35 de la CRP reconoce a los ciudadanos derechos de acceso, rectificación y actualización de sus datos, y limita el tratamiento informático para salvaguardar derechos fundamentales, un precepto más detallado y avanzado que el escueto mandato del art. 18.4 de la Constitución Española, aun siendo la aprobación de esta posterior[26].

24. Art. 26 DSA.

25. Jove Villares, D. «La inconstitucional habilitación a los partidos políticos para recabar datos sobre opiniones políticas: comentario a la STC 76/2019, de 22 de mayo», *Revista Española de Derecho Constitucional,* núm. 121, 2021, pp. 303-331.

26. Ramiro Lozano, I. y Carrascosa López, V. «La protección de datos personales en la Península Ibérica» *Informática y derecho: Revista iberoamericana de derecho informático*, núm. 4, 1994 (Ejemplar dedicado a: III Congreso Iberoamericano de Informática y Derecho), pp. 247-260.

A semejanza de la realidad jurídica española, la jurisprudencia portuguesa ha desarrollado estos principios adaptándolos al entorno digital, enfatizando la necesidad de preservar estos derechos frente a los nuevos desafíos tecnológicos. El Tribunal Constitucional de Portugal ha subrayado que cualquier limitación a libertades como la de expresión debe respetar estrictos criterios de determinación y proporcionalidad, conforme al artículo 18 de la CRP pues, de lo contrario, se puede correr el riesgo de colisionar con las libertades informativas básicas.

En el plano supranacional, la aplicación de la DSA en Portugal ha requerido ajustes institucionales. Así, mediante el Decreto-Lei n.º 20-B/2024, de 16 de febrero, se designó a la Autoridade Nacional de Comunicações (ANACOM) como Coordinador de Servicios Digitales, con competencia general para velar por el cumplimiento del DSA, y se atribuyeron funciones específicas a la *Entidade Reguladora para a Comunicação Social* (ERC) en materia de contenidos de medios y a la Inspección General de las Actividades Culturales (IGAC) en materia de derechos de autor. Fuera de esta designación de autoridades, Portugal no ha aprobado por ahora legislación interna adicional de desarrollo, confiando en la aplicación directa del DSA y en la coordinación con las instituciones europeas para asegurar un entorno en línea seguro y respetuoso con los derechos fundamentales.

En 2021 se aprobó la Carta Portuguesa de Derechos Humanos en la Era Digital[27] cuyo artículo primero dejaba clara su intención de considerar los derechos y libertades constitucionales plenamente aplicables en el ámbito digital. En ella se regulan determinados aspectos relacionados con el entorno digital tales como: el derecho de acceso; la libertad de expresión y creación en el entorno digital; la garantía de acceso y uso; el derecho a la protección contra la desinformación; los derechos de reunión, manifestación, asociación y participación en el entorno digital; el derecho a la privacidad en el entorno digital y el uso de la inteligencia artificial y de robots, entre otros.

Como puede apreciarse, esta Carta, aprobada de manera paralela a la Carta de Derechos Digitales española, guarda numerosas similitudes con esta última. Ambas surgen en un contexto social y jurídico marcado por la preocupación ante el impacto del entorno digital en los derechos fundamentales, ya sea en su dimensión positiva —por ejemplo, al promover herramientas de participación ciudadana o el derecho de acceso a la información pública— o en su vertiente negativa —como la difusión de información falsa, la discriminación en el acceso a la información o la vulneración del derecho a la intimidad y a la protección de datos, entre otros—. Así, tanto la Carta lusa como la española se inspiran en derechos fundamentales existentes y reconocen nuevos derechos derivados de estos, como la identidad digital, la neutralidad de Internet, el derecho de acceso, la inclusión digital, o la responsabilidad ante Inteligencia Artificial.

Sin embargo, una de las diferencias más evidentes entre las dos Cartas es la que hace referencia a su carácter vinculante. Mientras que la Carta española no tiene carácter de ley y se presenta como una medida de carácter declarativa o propositiva; la norma portuguesa fue aprobada por medio de *la Lei núm. 27/2021,*

27. *Lei núm. 27/2021. Carta Portuguesa de Direitos Humaos na Era Digital.*

de 17 de maio, lo cual le dota de un carácter legal más formalizado que un mero texto de referencia. Precisamente, el carácter no normativo del texto español ha generado numerosas críticas en cuanto a su eficacia y, por ende, falta de ambición[28], al no disponer, por ejemplo, de un régimen sancionador propio.

La Carta portuguesa tampoco ha estado exenta de críticas tras su aprobación. Una de ellas se refiere a la excesiva pretensión reflejada en su propio título, que alude a los «derechos humanos» cuando, desde un punto de vista constitucional, habría sido más adecuado hablar de «derechos fundamentales», ya que solo estos se circunscriben al ordenamiento jurídico de un Estado[29]. También, existe un amplio consenso en lo que respecta a la problemática interpretativa que supone la creación de un nuevo derecho a la protección contra la desinformación al entrar en conflicto directo con el propio derecho a la libertad de expresión, pudiendo restringir este último[30]. Conviene tener presente que el art. 16 de la Constitución portuguesa permite el reconocimiento de derechos fundamentales a través de una ley ordinaria[31] y, tal y como hemos mencionado anteriormente, la propia Carta de Derechos Humanos en la Era Digital, es una ley. A raíz de esas críticas, se promovieron iniciativas parlamentarias y un proceso de revisión legislativa que culminó con la Ley n.º 15/2022, de 11 de agosto, mediante la cual se modificó la redacción del artículo 6.º y se revocaron expresamente sus números 2 a 6, manteniéndose únicamente el primero, relativo a la obligación del Estado de asegurar el cumplimiento del Plan Europeo de Acción contra la Desinformación. El Tribunal Constitucional, en su Acórdão n.º 66/2023, reconoció esta modificación y destacó que, tras la revocación, el artículo perdió buena parte de su eficacia normativa, al haber quedado sin las disposiciones que definían y operacionalizaban las medidas de combate a la desinformación[32].

IV. A PROPÓSITO DE LA INCIDENCIA DEMOCRÁTICA DE LA INTELIGENCIA ARTIFICIAL

Siguiendo la misma línea que supuso a comienzos de los 2000 la generalización del uso de internet, actualmente la reciente irrupción de la IA en la esfera pública ha sido recibida de forma positiva en lo relativo a la democratización de la comunicación e información potenciando, por ende, una mayor participación política de la ciudadanía. No obstante, casi desde un primer mo-

28. Arce Jiménez, C. *¿Una nueva ciudadanía para la era digital?* Dykinson, Madrid. 2022. p. 60.

29. Simões Barata, M. y Resende Alves D. «O advento dos direitos digitais em Portugal e na União Europeia», *FutureLaw.* Vol. V. coord. por Fábio da Silva Veiga, Paulo de Brito, 2024, p. 302.

30. Soares Farinho, D. «The Portuguese Charter of Human Rights in the Digital Age: a legal appraisal», *Revista Española de la Transparencia*. Núm. 13, 2021, pp. 85-105.

31. Moreno González, G. «Los derechos fundamentales consagrados en la Constitución no excluyen cualesquiera otros que resulten de las leyes y de las normas aplicables de Derecho internacional». *La Constitución portuguesa de 1976 y textos complementarios,* Athenaica. Sevilla. 2023. p. 107.

32. Tribunal Constitucional de Portugal. 2023. Acórdão n.º 66/2023, 2 de fevereiro. Lisboa: Tribunal Constitucional. Disponible en https://www.tribunalconstitucional.pt/tc/acordaos/20230066.html

mento se han advertido riesgos importantes para la integridad de los procesos democráticos a través de la configuración de algoritmos en las plataformas digitales que pueden llegar a incidir en aspectos tan cruciales como son los procesos electorales. El riesgo que presenta esta nueva tecnología presenta, en palabras de Castellanos, una dificultad añadida al ser, en muchas ocasiones, indetectable, ya que el ciudadano no se dará cuenta del proceso debido a la gran cantidad de datos proporcionados, lo que hará que la manera en que se sugiere la modulación de sus convicciones políticas sea sutil[33].

El resultado es un riesgo cierto de que la voluntad ciudadana no se refleje auténticamente en las urnas, al estar condicionada por estímulos calculados. De hecho, las estrategias de campaña basadas en *big data* e IA, combinadas con noticias falseadas o desinformación automatizada (p. ej., *deepfakes*), amenazan la integridad del proceso democrático. No es casual que el Reglamento de IA de la Unión Europea prohíba aquellos sistemas de IA que empleen técnicas subliminales o engañosamente manipuladoras para alterar el comportamiento de las personas menoscabando su capacidad de decisión informada o que, del mismo modo, la DSA imponga obligaciones a las grandes plataformas para transparentar sus algoritmos de recomendación y mitigar la difusión viral de desinformación y contenidos ilícitos.

El mencionado Reglamento de IA adopta un enfoque de regulación por niveles de riesgo, imponiendo estrictas exigencias de transparencia, trazabilidad y evaluación de riesgos a los sistemas de IA de alto riesgo. Su objetivo declarado es asegurar que la IA se desarrolle de forma «segura, fiable y ética, centrada en el ser humano», garantizando el respeto a los derechos fundamentales y a valores democráticos como el pluralismo. Por su parte, la ya comentada Carta de Derechos Digitales reconoce principios como la garantía de no discriminación algorítmica, la transparencia de los algoritmos y la protección frente a decisiones automatizadas sin supervisión humana, anticipando derechos que podrían consolidarse legislativamente en el futuro. Asimismo, la reforma del marco de servicios digitales en Europa (DSA) y otros instrumentos como el Reglamento General de Protección de Datos (RGPD) complementan esta red de garantías, estableciendo controles sobre el uso de datos personales y obligaciones de diligencia para las empresas tecnológicas en protección de la esfera pública.

Se hace necesario una interpretación renovada de los derechos fundamentales clásicos a la luz de la revolución digital: libertades como la de expresión, información o pensamiento deben revalorizarse considerando que hoy su ejercicio depende de entornos informacionales dominados por algoritmos. Por ejemplo, el derecho a la participación política, para ser efectivo, requiere que los ciudadanos puedan formarse una opinión libre de manipulación encubierta; de igual modo, podría argumentarse la existencia de un incipiente derecho a la veracidad informativa o a recibir información no distorsionada algorítmicamen-

33. Castellanos Claramunt, J., «La inteligencia artificial en el contexto jurídico. El Derecho como garante del desarrollo democrático», *Diálogos jurídicos: Anuario de la Facultad de Derecho de la Universidad de Oviedo*. Núm. 9, 2024. p. 91.

te, derivado de la combinación de derechos ya consagrados (participación, expresión, etc.). Igualmente, resulta crucial actualizar la legislación electoral y de publicidad política para introducir exigencias de transparencia algorítmica en las campañas.

V. CONCLUSIONES

No cabe duda de que las transformaciones tecnológicas han modificado profundamente las condiciones del ejercicio de los derechos fundamentales. La Unión Europea, consciente de esta realidad, ha optado por construir un marco normativo que intenta equilibrar el avance de la innovación tecnológica con la preservación de los valores constitucionales que sustentan su identidad. En este contexto, la Ley de Servicios Digitales (DSA) representa un punto de inflexión en la construcción de un constitucionalismo digital europeo, en tanto que actualiza las garantías de libertad, privacidad y pluralismo en un entorno dominado por grandes plataformas globales y sistemas algorítmicos de enorme poder estructural.

La DSA forma parte de un bloque normativo europeo junto con el Reglamento General de Protección de Datos, el Reglamento de Mercados Digitales y el Reglamento de Inteligencia Artificial cuyo objetivo principal radica en equilibrar la innovación digital con la protección de los derechos fundamentales[34]. Este conjunto de normas configura una arquitectura jurídica común orientada a asegurar que el desarrollo tecnológico se mantenga al servicio del ser humano y no a la inversa. Se trata, en definitiva, de trasladar al espacio digital los valores europeos sustentados en el respeto de la dignidad humana, la libertad y la igualdad, mediante instrumentos que garanticen transparencia, rendición de cuentas y responsabilidad en la gestión de la información.

No obstante, en un entorno digital caracterizado por la difuminación de las fronteras, se complica de manera significativa cualquier intento de regulación o control efectivo. Todo ello unido a la creciente dependencia europea de infraestructuras y servicios digitales controlados por potencias extracomunitarias, como estadounidenses o chinas, suscita un debate sobre la soberanía digital europea, entendida como la búsqueda de una autonomía tecnológica y normativa que permita al continente afirmar su posición internacional y salvaguardar sus valores democráticos en la era digital[35]. En este sentido, la DSA no puede entenderse solo como una norma de mercado o de protección de consumidores, sino como un instrumento de afirmación política y constitucional de la Unión Europea en un contexto de competencia global por el control de los flujos de información y de los estándares éticos y jurídicos de la tecnología.

34. Barrero Artiguez, Á. «El Reglamento Europeo de Servicios Digitales y la Defensa de la Democracia». *Revista de Derecho Político*. Núm. 122, 2025, p. 318.

35. Robles Carrillo, M. «La articulación de la soberanía digital en el marco de la Unión Europea». *Revista de Derecho Comunitario Europeo*. Núm. 75, 2023, p. 151.

Desde el punto de vista constitucional, el análisis comparado entre España y Portugal revela la existencia de una base común de derechos y principios que permiten integrar la DSA en los respectivos ordenamientos sin rupturas sistémicas. Ambos países comparten una concepción garantista de los derechos fundamentales y han emprendido esfuerzos paralelos por adaptar sus marcos jurídicos a los desafíos del entorno digital, ya sea mediante la formulación de Cartas de Derechos Digitales o la designación de autoridades nacionales competentes. No obstante, la eficacia real de estas medidas dependerá en buena medida de la capacidad institucional de cada Estado para supervisar el cumplimiento de las obligaciones impuestas a las plataformas y asegurar que la protección de los derechos fundamentales se mantenga como eje central de la transformación digital.

El equilibrio entre la libertad de expresión y la lucha contra la desinformación constituye una de las cuestiones más complejas del debate actual. Las normas europeas, y en particular la DSA, tratan de evitar tanto la inacción frente a contenidos ilícitos o manipuladores como el riesgo de instaurar mecanismos de censura. La jurisprudencia constitucional española y portuguesa ha subrayado que toda limitación a la libertad de expresión debe ser estrictamente proporcional, clara y necesaria en una sociedad democrática. Ello implica que las medidas de moderación de contenidos o de control de información, ya provengan del Estado o de actores privados, deben estar sometidas a garantías de transparencia y revisión judicial efectiva. En última instancia, preservar el pluralismo informativo exige asegurar que los ciudadanos puedan acceder a una esfera pública digital libre, diversa y veraz.

Asimismo, la irrupción de la inteligencia artificial ha introducido nuevas tensiones vinculadas con la posible vulneración de derechos fundamentales. La posibilidad de que los algoritmos condicionen la opinión pública, reproduzcan sesgos discriminatorios o alteren los procesos democráticos obliga a reinterpretar los derechos clásicos a la luz de los riesgos tecnológicos. La DSA y el Reglamento de Inteligencia Artificial, en su interacción con el RGPD, apuntan hacia un nuevo paradigma de «gobernanza algorítmica responsable», en el que la transparencia, la trazabilidad y la supervisión humana se conviertan en garantías estructurales de los derechos fundamentales. De este modo, la dimensión tecnológica de la libertad, la privacidad y la igualdad se consolida como un elemento esencial del constitucionalismo europeo.

El análisis comparado demuestra, además, que tanto el ordenamiento español como el portugués comparten una concepción dinámica de la proporcionalidad, capaz de adaptarse a los retos de la sociedad digital. En ambos sistemas, los tribunales constitucionales han insistido en que la eficacia de una medida no puede justificar su constitucionalidad si compromete en exceso los derechos individuales. Este enfoque se alinea con la filosofía subyacente a la DSA, que busca conjugar la protección frente a los abusos tecnológicos con el respeto al núcleo esencial de las libertades públicas.

Por último, cabe destacar que la plena efectividad del modelo europeo dependerá no solo de la adopción formal de las normas, sino de su aplicación

práctica y del compromiso de los actores públicos y privados con los principios democráticos que las inspiran. El éxito de la DSA y de la estrategia europea de regulación digital no se medirá únicamente en términos de cumplimiento técnico, sino en la capacidad de consolidar un espacio digital europeo que promueva la confianza ciudadana, garantice la integridad de la información y fortalezca la participación democrática.

En definitiva, el proceso de digitalización ofrece a Europa una oportunidad singular para reafirmar su identidad constitucional en la era tecnológica, postulándose además como un sistema garantista en comparación con otras regulaciones como la estadounidense (más liberal) o la China (menos garantista). Frente a un escenario global marcado por la concentración de poder en manos de grandes corporaciones tecnológicas y por la expansión de modelos autoritarios de control digital, la Unión Europea ha optado por un camino propio: el de un humanismo digital basado en el Derecho. En ese horizonte, la DSA no solo regula los servicios digitales, sino que actúa como vehículo de actualización del constitucionalismo europeo, proyectando en el ciberespacio los valores que históricamente han definido al continente. Del mismo modo, la integración de la inteligencia artificial en las estructuras democráticas debe realizarse bajo el amparo del Estado de Derecho, garantizando que su desarrollo y aplicación sirvan para promover los valores constitucionales fundamentales —participación, igualdad y pluralismo— y no para erosionarlos mediante prácticas de control o manipulación masiva.

La consolidación de una verdadera soberanía digital europea exigirá, por tanto, mantener la coherencia entre la innovación tecnológica, la autonomía normativa y la protección efectiva de los derechos fundamentales, garantizando que la revolución digital no erosione, sino que refuerce, las bases democráticas de nuestras sociedades.

BIBLIOGRAFÍA

AFONSO, G. «Primeiras Notas sobre a Coordenação dos Serviços Digitais no Regulamento dos Serviços Digitais; em especial, o caso português», *e-Publica: public law journal*, vol. 11 núm. 3, 2024, pp. 71-102.

ARCE JIMÉNEZ, C. *¿Una nueva ciudadanía para la era digital?* Dykinson, Madrid. 2022.

BARATA, M.S y ALVES, D.R. «O advento dos direitos digitais em Portugal e na União Europeia», *FutureLaw*. Vol. V. coord. por Fábio da Silva Veiga, Paulo de Brito, 2024, pp. 298-307.

BARRERO ARTIGUEZ, Á. «El Reglamento Europeo de Servicios Digitales y la Defensa de la Democracia», *Revista de Derecho Político*. Núm. 122, 2025, pp. 295-326.

CASTELLANOS CLARAMUNT, J., «La inteligencia artificial en el contexto jurídico. El Derecho como garante del desarrollo democrático», *Diálogos jurídicos:*

Anuario de la Facultad de Derecho de la Universidad de Oviedo. Núm. 9, 2024, pp. 85-101.

COTINO HUESO, L. «Entre bytes y democracia. Cómo la DSA de la UE armoniza la libertad de expresión y la lucha contra la desinformación», *OTROSÍ.: Revista del Colegio de Abogados de Madrid*, núm. 2, 2024, pp. 42-44.

COTINO HUESO, L. «La primera sentencia general del Tribunal Constitucional sobre la libertad de expresión e información en Internet: Seguimos pendientes de muchos temas clave para el futuro», en M.V. Álvarez Buján (coord.); P. Simón Castellano (dir.), *Evolución e interpretación del TC sobre derechos fundamentales y garantías procesales: cuestiones recientemente controvertidas: Análisis de sus últimos pronunciamientos más reseñables*, Aranzadi, Cizur Menor (Navarra), 2023, pp. 45-69.

COTINO HUESO, L. (coord.). *La Carta de Derechos Digitales*, Tirant lo Blanch, Valencia, 2022.

GALDÁMEZ MORALES, A. *De las tecnologías disruptivas y su ordenación jurídica: proceso, naturaleza y sistema constitucional de la comunicación digital*. Tesis doctoral, Universidad de Sevilla, 2024.

GARCÍA MAJADO, P. «Libertades comunicativas y redes sociales: a propósito de la STC 8/2022, de 27 de enero de 2022», *Revista General de Derecho Constitucional*, núm. 37, 2022.

JOVE VILLARES, D. «La inconstitucional habilitación a los partidos políticos para recabar datos sobre opiniones políticas: comentario a la STC 76/2019, de 22 de mayo», *Revista Española de Derecho Constitucional,* núm. 121, 2021, pp. 303-331.

MONTERO CARO, M.D., *Democracia en transición. Una agenda para su regeneración*. Dykinson, Madrid, 2023.

MORENO GONZÁLEZ, G. «Los derechos fundamentales consagrados en la Constitución no excluyen cualesquiera otros que resulten de las leyes y de las normas aplicables de Derecho internacional». *La Constitución portuguesa de 1976 y textos complementarios*, Athenaica. Sevilla. 2023.

RAMIRO LOZANO, I. y Carrascosa López, V. «La protección de datos personales en la Península Ibérica» *Informática y derecho: Revista iberoamericana de derecho informático*, núm. 4, 1994 (Ejemplar dedicado a: III Congreso Iberoamericano de Informática y Derecho), pp. 247-260.

ROBLES CARRILLO, M. «La articulación de la soberanía digital en el marco de la Unión Europea», *Revista de Derecho Comunitario Europeo*. Núm. 75, 2023, pp. 133-171.

SENTÍ NAVARRO, C., «Desórdenes informativos en línea y la moderación de contenidos en la era digital», *Revista de Derecho Político*. Núm. 121, 2024, pp. 203-233.

ESPECIFICACIONES DE SEGURIDAD, MODELOS VERIFICABLES Y CONTROL JURÍDICO: HACIA UNA IA CONFIABLE EN LAS ADMINISTRACIONES PÚBLICAS[1]

Pere Simón Castellano
Profesor Titular de Derecho constitucional
Universidad Internacional de la Rioja - UNIR

SUMARIO: I. INTRODUCCIÓN. II. ESPECIFICACIONES DE SEGURIDAD EN IA: CONCEPTO Y FUNDAMENTO NORMATIVO. 1. Concepto y alcance. 2. Fundamento normativo. III. ESTÁNDARES TÉCNICOS Y LEGALES DE VERIFICABILIDAD DE LOS MODELOS DE IA. 1. Verificabilidad, explicabilidad y transparencia de los algoritmos públicos. 2. Estándares técnicos. 3. Estándares jurídicos. 4. Mecanismos de certificación y evaluación de conformidad. IV. MODELOS DE SUPERVISIÓN Y RESPONSABILIDAD JURÍDICA EN EL SECTOR PÚBLICO. 1. Supervisión humana e institucional. 2. Control democrático y jurisdiccional. 3. Responsabilidad jurídica y rendición de cuentas. V. CONCLUSIONES. *BIBLIOGRAFÍA*.

I. INTRODUCCIÓN

La incorporación de sistemas de inteligencia artificial (IA) en el sector público ha abierto enormes posibilidades de mejora en la eficiencia y calidad de los servicios administrativos. Sin embargo, también ha suscitado preocupaciones sobre los riesgos que estas tecnologías pueden entrañar para los derechos de

1. Esta investigación se ha realizado en el marco del proyecto de I+D+i *Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas* [2023-2026] (PID2022-136439OB-I00), financiado por MCIN/AEI/10.13039/501100011033/ y «FEDER Una manera de hacer Europa».

los ciudadanos y los principios del Estado de derecho[2]. La confiabilidad de la IA en las Administraciones Públicas se ha convertido así en un objetivo esencial, impulsando el desarrollo de marcos jurídicos específicos y requisitos técnicos rigurosos a nivel nacional, europeo e internacional.

La Unión Europea, en particular, ha buscado liderar este ámbito mediante un enfoque de regulación basado en el riesgo, plasmado en el reciente Reglamento (UE) 2024/1689 (en adelante, RIA[3]), cuyo propósito declarado es promover una IA *centrada en el ser humano y fiable* a la vez que se garantiza un elevado nivel de protección de la salud, la seguridad y los derechos fundamentales. En paralelo, instrumentos internacionales como el Convenio del Consejo de Europa sobre Inteligencia Artificial[4] reafirman la necesidad de que el diseño, desarrollo y uso de sistemas de IA respeten la democracia, el Estado de derecho y los derechos humanos, imponiendo obligaciones de transparencia, evaluación de riesgos y supervisión para mitigar posibles consecuencias negativas[5].

La confiabilidad de la IA en contextos públicos exige una aproximación multidisciplinar que conjugue soluciones técnicas robustas con garantías y controles jurídicos efectivos[6]. Desde el punto de vista técnico, surgen conceptos como las especificaciones formales de seguridad y los modelos verificables, orientados a dotar a los sistemas automatizados de garantías cuantitativas de comportamiento seguro. Estos enfoques buscan emular en la IA de alto riesgo los rigurosos estándares de certificación ya conocidos en sectores críticos tradicionales, tales como el transporte aéreo, la energía nuclear o los dispositivos médicos. Contextos en los que se exige probar preventivamente la seguridad de un sistema antes de autorizar su despliegue.

Por su parte, desde el prisma jurídico, se plantean requisitos de transparencia algorítmica, responsabilidad, supervisión humana y posibilidad de revisión de las decisiones automatizadas, todo ello para asegurar que la introducción de

2. Véanse al respecto L. Cotino Hueso, «Cómo abordar jurídicamente el impacto de la inteligencia artificial en los derechos fundamentales», en M. E. Casas Baamonde (Dir.) y D. Pérez del Prado (Coord.), *Derecho y tecnologías*, Madrid, Fundación Ramón Areces, 2025, pp. 123-176; R. Bustos Gisbert, «El constitucionalista europeo ante la inteligencia artificial: reflexiones metodológicas de un recién llegado», *Revista Española de Derecho Constitucional,* 131, 2024, pp. 146-178.

3. Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y se modifican diversos reglamentos y directivas (Ley de IA). Diario Oficial de la Unión Europea núm. 1689, de 12 de julio de 2024, p. 1-144.

4. Convenio marco sobre inteligencia artificial, los derechos humanos, la democracia y el Estado de Derecho (Convenio sobre la IA). Consejo de Europa, Estrasburgo, 17 de mayo de 2024.

5. Véase al respecto E. Chaveli Donet, «La evaluación de impacto de derechos fundamentales por quienes despliegan sistemas de inteligencia artificial en el Reglamento», L. Cotino Hueso y P. Simón Castellano (dirs.), *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Cizur Menor, Aranzadi, 2024, pp. 495-533.

6. Véase sobre este extremo 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*, 58, 2019, pp. 82-96.

la IA no socave los derechos de los ciudadanos a una buena administración[7], a la protección de sus datos personales ni a la tutela judicial efectiva.

La normativa vigente —desde el Reglamento General de Protección de Datos (RGPD) hasta el aludido RIA, pasando por disposiciones nacionales como el artículo 41 de la Ley 40/2015 en España o el artículo 23 de la Ley 15/2022 de igualdad de trato y no discriminación— establecen un entramado de garantías destinadas a encauzar el uso de algoritmos por parte de la Administración.

Asimismo, la jurisprudencia empieza a ofrecer pautas y límites: por ejemplo, una sentencia pionera del Tribunal de Distrito de La Haya anuló en 2020 un sistema automatizado antifraude por falta de transparencia y por impedir la defensa efectiva de los ciudadanos afectados, al entender que dicho algoritmo violaba desproporcionadamente el derecho a la vida privada protegido por el artículo 8 del Convenio Europeo de Derechos Humanos[8]. Más reciente es la resolución del caso BOSCO por parte del Tribunal Supremo de España, que recuerda que la transparencia es muy importante también cuando se emplean sistemas informáticos en la toma de decisiones automatizadas por parte de las Administraciones Públicas[9]. Este y otros precedentes evidencian la necesidad de mecanismos de control jurídico que complementen las soluciones técnicas, de modo que las Administraciones puedan beneficiarse de la IA sin menoscabar los derechos ni la confianza del público.

En este capítulo se abordan, de forma integrada, estos tres ejes fundamentales para una IA confiable en el sector público: (i) las especificaciones de seguridad que deben guiar el diseño de sistemas de IA seguros, (ii) los modelos verificables y métodos formales que permiten comprobar el cumplimiento de dichas especificaciones, y (iii) el control jurídico y la regulación que enmarcan y supervisan el uso de la IA en las Administraciones Públicas.

A lo largo del texto se expondrá cómo las técnicas de verificación formal proporcionan garantías adicionales frente a los riesgos de la IA, por qué las

7. Véase al respecto J. Ponce Solé, *El Reglamento de Inteligencia Artificial de la Unión Europea de 2024, el derecho a una buena administración digital y su control judicial en España*, Marcial Pons, Madrid, 2024.

8. Tribunal de Distrito de La Haya, sentencia de 5 de febrero de 2020 (caso SyRI), declarando ilegal un sistema algorítmico de perfilado de riesgo de fraude social por vulnerar el derecho a la vida privada al carecer de la transparencia y el equilibrio justo exigible: el algoritmo no hacía público su modelo ni informaba a los ciudadanos afectados, impidiendo verificar su diseño y dificultando el derecho de defensa. Se considera la primera sentencia europea que anula una decisión automatizada en la esfera pública por violar derechos fundamentales. Véase L. Cotino Hueso, «"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», *La Ley privacidad*, 4, 2020.

9. Nos referimos a la Sentencia del Tribunal Supremo núm. 1119/2025, de 11 de septiembre, de la Sala de lo Contencioso-Administrativo (Sección 3ª), que ha anulado las sentencias previas del Juzgado Central de lo Contencioso-Administrativo y de la Audiencia Nacional. El Tribunal Supremo reconoce a la fundación CIVIO el derecho a acceder al código fuente del algoritmo BOSCO, una aplicación informática (cuya clasificación como sistema o modelo de IA es controvertida por la doctrina) mediante la que la Administración indica a los comercializadores de energía eléctrica si un consumidor reúne, o no, los requisitos legales (familiares y de renta) que permiten beneficiarse del bono social.

normas jurídicas emergentes (europeas e internas) demandan cada vez más dichas garantías, y cómo ambas dimensiones —la técnica y la legal— pueden articularse para avanzar *hacia una IA confiable en las Administraciones.*

II. ESPECIFICACIONES DE SEGURIDAD EN IA: CONCEPTO Y FUNDAMENTO NORMATIVO

1. Concepto y alcance

En el desarrollo de sistemas de inteligencia artificial de alto impacto —muy especialmente aquellos empleados en la automatización de decisiones administrativas— resulta crucial definir especificaciones de seguridad claras y exigentes desde la fase de diseño. Una especificación de seguridad constituye, en este contexto, una descripción formal de las condiciones de funcionamiento seguro y aceptable del sistema de IA, mediante criterios objetivos y verificables que delimitan los resultados permitidos y aquellos considerados prohibidos por entrañar un riesgo inaceptable. Dicho de otro modo, la especificación fija el umbral de tolerancia al riesgo.

En un ejemplo práctico, hablando de un sistema de IA más básico como podría ser uno basado en IA simbólica[10] como podría catalogarse en el mejor de los casos BOSCO, un sistema de evaluación automatizada de solicitudes de prestación social debería garantizar que ninguna persona que cumpla con los requisitos legales sea injustamente rechazada (previniendo falsos negativos), o que la tasa de error no supere un margen previamente definido con un nivel de confianza estadística determinado. A diferencia de las meras directrices éticas o códigos de buenas prácticas, las especificaciones de seguridad suelen adoptar una formulación lógica o matemática rigurosa, susceptible de verificación formal.

10. La llamada *IA simbólica* o *Good Old-Fashioned AI* (GOFAI) hace referencia al enfoque clásico de la IA basado en la representación explícita del conocimiento mediante símbolos y reglas lógicas, en contraposición a los modelos conexionistas actuales como las redes neuronales profundas. En este paradigma, los sistemas de IA manipulan símbolos siguiendo estructuras formales —como árboles de decisión o lenguajes de programación lógicos— para imitar el razonamiento humano. Aunque hoy ha perdido centralidad frente al aprendizaje automático, la IA simbólica sigue siendo relevante en entornos donde la trazabilidad, la interpretabilidad y la seguridad jurídica de las decisiones resultan esenciales. La IA simbólica es menos potente, segura o robusta, pero es más fácil de entender o comprender por parte de los seres humanos, al tratarse de una IA determinista, aunque a veces resulte complejo, costoso o difícil hacer un esfuerzo para encontrar las razones lógicas del razonamiento, probabilidad o propuesta efectuada por el sistema. Véase S. J. Russell y P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, New Jersey, 1995. El sistema BOSCO, por ejemplo, por parte de algunos autores ha sido calificado como una IA muy simple o básica, que estaría enmarcada dentro del paraguas de la IA simbólica, lo que justificaría el tratamiento que se le ha dado en la ya citada Sentencia del Tribunal Supremo núm. 1119/2025, de 11 de septiembre.

La idea de dotar a la IA de especificaciones de seguridad hunde sus raíces en la ingeniería de sistemas críticos, donde resulta impensable autorizar el despliegue de un dispositivo sin criterios exhaustivos de seguridad[11]. En el ámbito administrativo, la analogía es evidente: las decisiones automatizadas pueden incidir de manera directa y grave en derechos fundamentales, lo que obliga a definir ex ante los requisitos técnicos y jurídicos que aseguren su legitimidad.

Estas especificaciones abarcan varias dimensiones, que van desde criterios funcionales, que definen lo que el sistema debe o no debe hacer en cada supuesto; límites de rendimiento cuantitativo, como márgenes de error tolerables o umbrales de fiabilidad estadística; o requerimientos jurídicos, orientados a garantizar principios como la no discriminación, la proporcionalidad o la protección de la privacidad.

Una especificación de seguridad integral para un algoritmo público podría establecer, verbigracia, que *la decisión no podrá basarse en variables sensibles como la raza o la religión del interesado* (garantía de no discriminación), o que *ante la mínima incertidumbre sobre un caso que pueda perjudicar a un ciudadano, la decisión se derivará a un funcionario humano* (garantía pro persona en caso de duda). Estos elementos complementan la pura lógica de negocio del sistema o modelo de IA, con cláusulas de salvaguarda destinadas a proteger derechos. En efecto, tal como sugiere la literatura sobre IA confiable, las especificaciones de seguridad deben entrelazar requisitos técnicos con valores jurídicos, de forma que la conformidad del sistema con la ley pueda ser evaluada en términos verificables[12].

De este modo, las especificaciones actúan como un contrato técnico-jurídico entre el diseño algorítmico y los valores constitucionales, sirviendo como puente entre ingeniería y derecho. Su correcta formulación constituye la condición de posibilidad para posteriores ejercicios de verificación y validación. Una vez definidas, abren la puerta a la aplicación de métodos rigurosos de validación, como veremos en el siguiente apartado. Sin una buena especificación, no es posible

11. Parte de la doctrina ha definido la necesidad de someter sistemas de IA a estándares rigurosos de seguridad, argumentando que es plausible que pronto existan sistemas de IA al menos tan críticos en términos de seguridad, lo que significa que deberían adherirse a estándares de seguridad igual de estrictos. Véase D. Dalrymple, J. Skalse, Y. Bengio, S. Russell, M. Tegmark, S. Seshia, S. Omohundro, C. Szegedy, A. Abate, J. Halpern, C. Barrett, D. Zhao, B. Goldhaber y N. Ammann, «Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems», *Technical Report UCB/EECS-2024-45*, Departamento de Ingeniería Eléctrica y Ciencias de la Computación, Universidad de California, Berkeley, 4 de mayo de 2024, pp. 1-30.

12. De nuevo, el polémico caso BOSCO es un buen ejemplo, puesto que el Consejo de Transparencia de España exigió las especificaciones de funcionamiento y las pruebas de validación del algoritmo, equivalentes a una suerte de auditoría técnica de su fiabilidad y adecuación a la normativa, en lugar de centrase en la apertura o acceso íntegro al código fuente. Este requerimiento refleja correctamente cómo las autoridades deberían exigir que las Administraciones definan y documenten las reglas de decisión de sus algoritmos, esto es, sus especificaciones, para posibilitar un control público de su correcto diseño. Véase al respecto 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*, 1, 2020, pp. 223-269.

garantizar valores o derechos y cualquier evaluación de seguridad será vaga o incompleta. Por ello, organismos reguladores y guías oficiales recomiendan fuertemente dedicar esfuerzos a esta fase inicial de determinación de requisitos.

2. Fundamento normativo

El fundamento normativo de las especificaciones de seguridad en IA se encuentra, en primer lugar, en el RIA, que clasifica numerosas aplicaciones públicas como sistemas de alto riesgo y exige a los proveedores documentar ex ante las medidas técnicas y organizativas de mitigación destinadas a garantizar la confiabilidad de sus sistemas: robustez, exactitud, ciberseguridad y ausencia de sesgos indebidos. Este marco normativo consagra el principio de *security by design*, obligando a que la seguridad sea incorporada desde el diseño inicial y no como un añadido posterior. No es de extrañar tampoco que se exija a todo el sector público, cuando puedan ser considerados responsables del despliegue de sistemas y modelos de IA, a realizar evaluaciones de impacto en los derechos fundamentales, de forma previa a la implementación y uso del sistema, tratándose de una obligación desigual o descafeinada para el sector privado[13].

Además, organismos europeos de supervisión, como la futura Oficina Europea de IA o nacionales, como la Agencia Estatal de Supervisión de la IA (AESIA), tienen previsto emitir directrices específicas para la clasificación de riesgos y la determinación de especificaciones técnicas adecuadas, consolidando así la exigencia de que toda autoridad u operador público pueda demostrar la seguridad jurídica y técnica de sus algoritmos antes de su despliegue.

El Consejo de Europa, por su parte, ha establecido un marco complementario con la Convención Marco sobre Inteligencia Artificial, los Derechos Humanos, la Democracia y el Estado de Derecho, cuyo artículo 16 impone la obligación de evaluar y mitigar de forma continua los riesgos e impactos adversos de los sistemas de IA. Esta previsión refuerza la idea de que la seguridad debe entenderse en sentido amplio, no solo como fiabilidad técnica sino también como garantía material de los derechos fundamentales.

Por todo ello, puedo afirmarse sin temor a caer en error que tanto el legislador de la Unión Europea como las instancias paneuropeas han sentado las bases normativas para que las especificaciones de seguridad de la IA pública constituyan un requisito jurídico vinculante, y no un mero estándar de buena práctica. Ello confirma que la cultura de la prevención y la anticipación de riesgos ha de convertirse en la piedra angular de una IA confiable en el sector público.

13. Al respecto véase P. Simón Castellano, «L'avaluació d'impacte en els drets fonamentals en l'ús d'intel·ligència artificial en el sector públic: models i metodologies en perspectiva comparada», *Revista Catalana de Dret Públic*, 71, 2025; P. Simón Castellano, *La evaluación de impacto algorítmico en los derechos fundamentales*, Aranzadi, Cizur Menor, 2023; E. Chaveli Donet, «La evaluación de impacto de derechos fundamentales… *ob. cit.*

III. ESTÁNDARES TÉCNICOS Y LEGALES DE VERIFICABILIDAD DE LOS MODELOS DE IA

1. Verificabilidad, explicabilidad y transparencia de los algoritmos públicos

En el contexto de los algoritmos públicos, la verificabilidad, la explicabilidad y la transparencia se consideran pilares de la confianza y la rendición de cuentas. Transparencia implica que se conozca cómo funciona el algoritmo o, al menos, cuáles son sus criterios lógicos y datos utilizados; la explicabilidad alude a la capacidad de ofrecer a los afectados una justificación comprensible de cada decisión individual tomada por la IA; y la verificabilidad supone la posibilidad de comprobar formalmente que el algoritmo se comporta conforme a ciertas reglas o especificaciones en todos los casos previstos[14].

Estos tres conceptos están relacionados, pero no son intercambiables: un sistema puede ser transparente (mostrar su código o reglas) y aun así ser difícil de explicar a un lego, o podría ofrecer explicaciones de sus resultados sin que por ello esté matemáticamente garantizado que nunca violará una norma. En las aplicaciones públicas, en función del contexto de uso, lograr las tres dimensiones puede resultar complejo, aunque es especialmente importante que en algunos supuestos se cumplan dado que están en juego derechos de la ciudadanía, principios de igualdad ante la ley y la legitimidad de la acción estatal.

La verificabilidad aporta un nivel de garantía adicional al complementar la transparencia y la explicabilidad. Un algoritmo público transparente permite el escrutinio externo, por ejemplo, mediante auditorías independientes, y la explicabilidad asegura que cada decisión pueda ser motivada en términos comprensibles. Sin embargo, solo la verificabilidad, a vueltas vinculada con la transparencia como propiedad, podría proporcionar una certeza casi absoluta de que el sistema siempre respetará ciertos requisitos críticos como, por ejemplo, no denegar nunca un beneficio social a quien legalmente le corresponda. Expertos en seguridad de sistemas o modelos de IA sostienen que las pruebas empíricas y la interpretabilidad humana, si bien son útiles, no aportan las sólidas garantías de seguridad que sí puede ofrecer la verificación formal[15].

De este modo, examinar el algoritmo y hacer *testing* o comprobaciones sobre casos de ejemplo puede generar confianza, pero no significa que se consiga probar exhaustivamente la ausencia de errores o sesgos ocultos. En cambio, la verificación formal, siempre y cuando sea viable, permite demostrar con rigor matemático que ciertas situaciones indeseables no ocurrirán, cubriendo así lagunas que escaparían a otros métodos.

14. Sobre las propiedades y subpropiedades que pueden considerarse garantías jurídicas véase el trabajo P. Simón Castellano, «Taxonomía de las garantías jurídicas en el empleo de los sistemas de inteligencia artificial», *Revista de Derecho Político*, 117, 2023, pp. 153-196.

15. Véanse P. Simón Castellano, «Taxonomía de las garantías jurídicas… *ob. cit*; D. Dalrymple, et al., «Towards Guaranteed Safe AI… *ob. cit.*

Cabe enfatizar que las tres aproximaciones no se excluyen, sino que se refuerzan mutuamente. Una administración pública responsable debería aspirar a que sus sistemas algorítmicos sean transparentes y explicables y, en la medida de lo posible, formalmente verificables en cuanto a su conformidad con las leyes y valores públicos. De hecho, un planteamiento holístico abogaría por una estrategia en cartera que combine auditorías externas, publicación de información sobre el funcionamiento del algoritmo, mecanismos de explicación caso a caso, y técnicas formales de verificación en los aspectos más críticos.

Un ejemplo ilustrativo proviene de la experiencia comparada, el caso SyRI ya citado anteriormente, que muestra como en 2020 un tribunal neerlandés anuló un sistema de IA usado para detectar fraude en prestaciones sociales, en parte porque no cumple las exigencias de proporcionalidad, carece de transparencia y vulnera el respeto a la vida privada, lo que hacía imposible verificar cómo está diseñado el árbol de decisión que utiliza el algoritmo. El caso nos sirve como botón de muestra que subraya que, sin transparencia y verificabilidad, las personas afectadas quedan indefensas y los derechos fundamentales pueden verse comprometidos. Por tanto, en el ámbito público, la legitimidad del uso de IA requiere no solo que las decisiones automatizadas estén debidamente motivadas, sino idealmente que se pruebe que dichas decisiones no violarán las normas sustantivas aplicables, por ejemplo, las garantías legales de no discriminación o los criterios objetivos fijados en una política pública.

Por ello, interesa señalar que verificabilidad, explicabilidad y transparencia forman un triángulo virtuoso para los algoritmos públicos. La transparencia permite supervisión y auditabilidad; la explicabilidad garantiza el derecho del ciudadano a conocer las razones de una decisión; y la verificación formal brinda la máxima confianza en que el sistema se atiene a reglas fijadas de antemano. Las recientes tendencias tanto técnicas como jurídicas, que detallaremos a continuación, van encaminadas a fomentar estas cualidades de forma conjunta, de modo que la adopción de IA en el sector público no menoscabe la seguridad jurídica ni la confianza ciudadana en las instituciones.

2. Estándares técnicos

Desde el punto de vista técnico, lograr IA verificable implica incorporar métodos de validación formal al ciclo de diseño y despliegue de los algoritmos. En términos generales, un enfoque de IA verificable (en la literatura anglosajona, Guaranteed Safe AI o GS-AI) se basa en tres componentes fundamentales[16]. El primero, el modelo formal del entorno (*world model*), una representación matemática o lógica del mundo o contexto en el que opera la IA, que describe con precisión las variables relevantes y cómo las acciones del sistema afectan el estado de ese entorno. Este modelo abstrae las condiciones de operación y

16. Seguimos aquí la descripción ofrecida en D. Dalrymple, et al., «Towards Guaranteed Safe AI… *ob. cit.*

los posibles escenarios (por ejemplo, las reglas de un dominio, límites físicos, supuestos sobre los datos de entrada, etc.).

En segundo lugar, el componente de la especificación formal de seguridad, que es un conjunto de condiciones, invariantes o propiedades lógicas que definen el comportamiento aceptable del sistema en términos de sus efectos o resultados. Equivale a traducir los requisitos de seguridad o legales a un lenguaje matemático no ambiguo. Por ejemplo, una especificación podría ser «si el solicitante cumple todos los requisitos legales, entonces la IA debe conceder la prestación» (garantía de no exclusión indebida).

El tercer componente es el verificador o el procedimiento de verificación, esto es, el mecanismo algorítmico capaz de analizar el modelo del sistema y demostrar, idealmente con validez matemática, que el sistema cumple la especificación de seguridad en todos los casos previstos. El verificador produce típicamente una prueba o certificado auditable de que, bajo las hipótesis del modelo formal, no existe recorrido de ejecución del programa que viole las condiciones de seguridad definidas.

Con estos tres elementos, en teoría, se puede obtener una prueba de corrección del sistema de IA respecto a las propiedades críticas que nos interesan. Parte de la doctrina y, muy especialmente aquellos que proponen el marco GS-AI, describen que, mediante la interacción rigurosa entre el modelo del mundo, la especificación formal y el verificador, es posible equipar a los algoritmos con garantías cuantitativas de seguridad de alta confianza[17]. Así, a diferencia de la validación tradicional por pruebas experimentales que solo cubren muestras finitas de casos, la verificación formal aspira a cubrir casi todos los escenarios posibles, proporcionando una certeza mucho mayor sobre el comportamiento del sistema.

Para ilustrar cómo opera este enfoque, retomemos el ejemplo simplificado de un sistema automático de concesión de prestaciones sociales, aprovechando la cercanía temporal y la polémica en torno al caso BOSCO ya citado anteriormente. Se construiría un modelo formal del mundo que incluya las reglas legales de elegibilidad y el formato de los datos de entrada (solicitudes de ciudadanos). La especificación formal de seguridad establecería algo así como «nunca denegar una prestación a alguien que cumpla todos los requisitos legales», especificando de forma más concreta cuales son los requisitos. El verificador, en último lugar, analizaría el código o la lógica de decisión de la IA para comprobar que para cualquier combinación posible de datos de entrada que satisfaga los requisitos, la salida del sistema es efectivamente la concesión de la prestación. Si se encuentra un caso en que alguien elegible sería denegado, el verificador lo reportaría como contraejemplo; si no, emitiría un certificado de cumplimiento. En este sentido, la verificación formal actúa como un *control de calidad automatizado* capaz de escudriñar exhaustivamente la lógica de la IA.

Es importante señalar que las técnicas de verificación formal no son nuevas en informática, puesto que en sistemas de software convencionales (determinis-

17. De nuevo, véase D. Dalrymple, et al., «Towards Guaranteed Safe AI... *ob. cit.*

tas), especialmente en entornos críticos (aeroespacial, sanidad, control industrial), se aplican desde hace décadas métodos como el *model checking* para la exploración exhaustiva de estados o las demostraciones formales asistidas por ordenador. Sin embargo, su aplicación a sistemas de IA avanzada plantea desafíos particulares. Muchos algoritmos de IA modernos se basan en modelos de aprendizaje automático (p. ej. redes neuronales profundas entrenadas con grandes volúmenes de datos) que no siguen una lógica determinista explícita escrita por programadores, sino que contienen patrones estadísticos complejos. Esta caja negra estadística dificulta extraer garantías formales tradicionales, porque el espacio de posibles entradas es enorme y las relaciones internas no son fácilmente expresables en lógica proposicional o de primer orden. Con todo, en la medida que se trate de meros sistemas automáticos de IA simbólica, como es el caso del ejemplo de árbol de decisión simple para concesión de prestaciones sociales, la verificación formal sí que reúne todo el potencial para actuar como una verdadera garantía jurídica.

Además, interesa destacar que la investigación reciente ha logrado progresos notables en verificación de modelos de aprendizaje automático. Por un lado, se exploran métodos para imponer a priori restricciones durante el entrenamiento de redes neuronales, de modo que ciertas propiedades deseables queden garantizadas por construcción (por ejemplo, añadir términos en la función de costo que penalicen la falta de equidad, logrando modelos intrínsecamente más justos). Por otro lado, han surgido analizadores formales a posteriori que toman una red neuronal entrenada y examinan su estructura (sus pesos y funciones de activación) buscando demostrar propiedades o encontrar contraejemplos. Por ejemplo, existen herramientas capaces de verificar que un clasificador de imágenes no cometerá ciertos errores graves dentro de un rango de variación de los pixeles de entrada[18]. Este tipo de verificación se ha aplicado en dominios como la conducción autónoma y la aviación, en los que se verificaron propiedades de seguridad en redes neuronales de un sistema de evasión de colisiones aéreas (ACAS Xu), logrando probar matemáticamente que, bajo ciertas condiciones de incertidumbre, el controlador autónomo siempre evitará la colisión entre dos aeronaves[19]. Estos resultados demuestran que la verificación de redes neuronales es posible al menos para escenarios acotados, si bien revelan también los límites actuales de escalabilidad[20].

Los principales obstáculos técnicos señalados en la literatura para extender la verificación formal a IA complejas se pueden clasificar, además, teniendo en

18. Véase S. Bak, «Verifying Neural Networks to Avoid Air-to-Air Collisions», AI Innovation Institute, Stony Brook University, 2024, disponible en https://ai.stonybrook.edu/about-us/News/verifying-neural-networks-avoid-air-air-collisions

19. *Ibidem*.

20. En el caso citado, solo 5 de 45 redes pudieron ser verificadas exhaustivamente con las herramientas disponibles. El propio Bak afirma que su estudio evidenció tanto un resultado positivo —que el análisis formal de redes es factible— como un resultado negativo —grandes redes o sistemas con dinámicas muy complejas aún quedan fuera del alcance de estas técnicas por razones computacionales—. S. Bak, «Verifying Neural Networks… *ob. cit.*

cuenta los componentes que tres componentes que integran la verificabilidad. En relación con el modelado del mundo, el desafío es construir un modelo formal del entorno suficientemente fiel a la realidad, pero manejable computacionalmente. Si el modelo es demasiado simple, las garantías obtenidas podrían no transferirse al sistema real; si es demasiado detallado, la verificación se vuelve intratable. Una estrategia para abordar esto es usar modelos conservadores que acoten el peor caso (*worst-case scenario*), garantizando seguridad bajo condiciones incluso más adversas que las reales, aunque a costa de cierto conservadurismo.

Por lo que se refiere a las especificaciones formales, hay que tener presente que traducir requerimientos complejos (muchas veces provenientes de normativa jurídica o de consideraciones éticas) a fórmulas lógicas sin ambigüedad puede resultar complejo y requiere de equipos de trabajo multi y interdisciplinares. Muchas nociones legales (como «justicia» o «igualdad») son difíciles de formalizar. Se propone empezar por invariantes críticas y mucho más concretas y objetivas (por ejemplo, prohibir absolutamente la discriminación directa por raza o género en una decisión algorítmica) e ir refinando con propiedades adicionales. La especificación puede enriquecerse gradualmente, pero debe mantenerse coherente y verificable.

En cuanto a la escalabilidad de la verificación, los algoritmos de IA pueden tener millones de parámetros o incontables estados posibles. Los métodos formales clásicos sufren explosión combinatoria al analizar sistemas complejos. Para enfrentar esto, se investigan técnicas híbridas que combinen análisis estático inteligente con pruebas aleatorias dirigidas (*fuzzing* guiado) para cubrir más espacio de estados. También se estudia emplear la propia IA como aliada, a través de meta-verificadores impulsados por aprendizaje automático que aprendan a encontrar más eficientemente contraejemplos o zonas problemáticas en el modelo.

Adicionalmente, desde la ingeniería de software se sugiere diseñar la arquitectura de los sistemas de IA de forma modular para facilitar la verificación. Por ejemplo, dividir un sistema complejo en módulos o submodelos más simples, de tal manera que cada uno pueda certificarse por separado respecto a ciertas propiedades (divide y vencerás).

Otra práctica es incorporar contratos en el código, esto es, aserciones lógicas que el propio programa verifica durante la ejecución o en tiempo de compilación, asegurando que no se violen condiciones predefinidas. Incluso se han propuesto arquitecturas de IA con un módulo verificador interno actuando como filtro final de las decisiones. En este esquema, el sistema de IA genera una decisión candidata, por ejemplo, conceder o denegar un permiso, y antes de emitirse al mundo real, la decisión es validada por el módulo verificador que comprueba que no vulnere ninguna regla de seguridad o normativa. Solo si pasa esta comprobación, la decisión se ejecuta; si no, se rechaza o envía a revisión humana. Esto emula un doble control de calidad automatizado. Por ejemplo, imaginemos una IA de un ente municipal que propone la resolución de un expediente urbanístico: un verificador integrado podría chequear que la resolu-

ción propuesta no contradiga ninguna ordenanza ni exceda las atribuciones del órgano. Por ejemplo, en el caso de propuesta de sanción, verificar que el importe esté dentro de los límites legales para ese caso y que sea proporcional. Este doble chequeo algorítmico reduciría drásticamente la probabilidad de errores graves o decisiones arbitrarias, incrementando la confianza en el sistema.

Así las cosas, desde el punto de vista técnico la verificación formal en IA consiste en adaptar y extender las herramientas matemáticas de la informática teórica para cubrir modelos basados en datos y aprendizaje. Aún persisten retos formidables —modelar el mundo, formalizar la ética y escalar los algoritmos— pero las líneas de investigación descritas muestran caminos prometedores. Un punto fundamental para tener en cuenta es que la verificación formal no sustituye a otras medidas tradicionales de validación, tales como las pruebas empíricas o la interpretabilidad, sino que las complementan. De hecho, en sistemas críticos se busca una aproximación integral, es decir, se siguen haciendo pruebas con datos reales, se analiza la interpretabilidad del modelo (por ejemplo, técnicas XAI para extraer explicaciones), pero además se añade una capa de verificación formal para aquellos requisitos que no admiten excepciones. Probar un modelo únicamente con simulaciones o ejemplos deja infinitos casos sin cubrir, y nunca seremos capaces de garantizar que no haya errores, mientras que los métodos formales intentan justamente cubrir ese infinito mediante razonamiento matemático sobre conjuntos de estado. El futuro pasa entonces por los estándares técnicos de vanguardia, específicos para alcanzar una IA segura i confiable, como si se tratara de una garantía matemática que debe convivir las comprobaciones y las explicaciones que se exigen como garantía jurídica, elevando en definitiva el nivel de confianza justificable en sistemas cada vez más autónomos y complejos.

3. Estándares jurídicos

Paralelamente a los avances técnicos, el marco jurídico se está adaptando para exigir mayores garantías en los sistemas de IA, especialmente en aquellos de alto riesgo o usados en el sector público. Diversas normas ya consagran principios de transparencia, explicabilidad y control humano que inciden en la verificabilidad de los modelos de IA[21].

Por un lado, el Derecho administrativo tradicional impone la motivación adecuada de los actos administrativos. Esto significa que cualquier decisión que afecte derechos de los ciudadanos debe estar fundada en razones comprensibles y verificables[22].

21. Véanse J. Ponce Solé, *El Reglamento de Inteligencia Artificial… ob. cit;* A. Cerrillo Martínez y C. I. Velasco Rico, «La transparencia algorítmica», *Working Papers DIGITAPIA*, 1 (3), 2025, pp. 25-41.

22. Véase A. Cerrillo Martínez, «La transparencia de los algoritmos que utilizan las administraciones públicas», *Anuario de Transparencia Local*, 3, 2020, pp. 41-78.

El uso del tratamiento automatizado para la toma de decisiones solamente está permitido[23] cuando (i) la decisión basada en el algoritmo es necesaria para celebrar o ejecutar un contrato con la persona cuyos datos haya tratado a través del algoritmo (por ejemplo, una solicitud de préstamo por internet); (ii) una ley concreta permite el uso de algoritmos y ofrece garantías adecuadas para salvaguardar los derechos, las libertades y los intereses legítimos de la persona (por ejemplo, la legislación contra la evasión de impuestos); (iii) la persona ha consentido explícitamente una decisión basada en el algoritmo. Sin embargo, la decisión automatizada debe salvaguardar los derechos, las libertades y los intereses legítimos de la persona aplicando las garantías adecuadas. Salvo cuando este tipo de toma de decisiones se base en una ley, en los otros dos casos se deberá informar a la persona, como mínimo, de (i) la lógica aplicada en el proceso de toma de decisiones, (ii) su derecho a obtener intervención humana, (iii) las posibles consecuencias del tratamiento, y (iv) su derecho a impugnar la decisión. Por tanto, deberá establecer los requisitos de procedimiento obligatorios para que la persona pueda expresar su punto de vista e impugnar la decisión.

Este criterio jurídico básicamente prohíbe en la UE, por razones de Estado de Derecho, las decisiones administrativas puramente algorítmicas que carezcan de transparencia y explicabilidad. Así, se refuerza la idea de que cualquier algoritmo público debe poder ser auditado y entendido por humanos, al menos a posteriori, para que sus resultados sean legalmente válidos.

En España, la Ley 40/2015 de Régimen Jurídico del Sector Público ya define la *actuación administrativa automatizada* y exige que se predefinan de antemano los criterios aplicables y que se identifique un responsable de la calidad y seguridad del algoritmo. De hecho, el artículo 41.2 de la citada Ley 40/2015 va más allá al exigir que en las actuaciones automatizadas se establezcan previamente las especificaciones, programación, mantenimiento, supervisión y control de calidad, y, en su caso, auditoría del sistema de información y de su código fuente, designando un responsable a efectos de impugnación.

El artículo 41.2 de la Ley 40/2015 sienta una base legal indubitada para la verificabilidad: la Administración debe ser capaz de demostrar que su algoritmo decide conforme a las reglas fijadas y esas reglas, además, deben ser públicas si bien, por lo que se refiere al grado de publicidad o accesibilidad, lo mínimo es que deben ser auditables.

Asimismo, la normativa de protección de datos personales y el art. 22 del RGPD reconocen al individuo el derecho a no ser objeto de decisiones totalmente automatizadas que produzcan efectos jurídicos adversos, salvo ciertas excepciones, y en todo caso a obtener una explicación sobre la lógica empleada. Todo ello configura un entorno legal en el que la opacidad algorítmica está crecientemente vedada, y donde surge la necesidad de mecanismos formales de verificación que permitan a los operadores demostrar el cumplimiento de tales obligaciones.

23. Seguimos aquí las Directrices del CEPD sobre decisiones individuales automatizadas y perfilado a los efectos del RGPD.

El desarrollo más significativo, no obstante, ha sido la aprobación del RIA, que establece un marco jurídico uniforme sobre IA en Europa, con un enfoque basado en el riesgo. Los sistemas de IA se clasifican en prohibidos (riesgo inaceptable), de alto riesgo, de riesgo limitado o mínimo.

Los sistemas de alto riesgo (que incluyen muchos sistemas de interés público, como algoritmos de puntuación crediticia, selección de personal, gestión de beneficios sociales, herramientas policiales, sanitarias, de transporte autónomo, etc.) están sujetos a requisitos legales estrictos antes de su puesta en marcha y ejecución. Entre estos requisitos destacan la gestión de riesgos[24] durante todo el ciclo de vida, altos estándares de calidad de datos, medidas de seguridad, robustez y precisión, trazabilidad en los procesos de diseño (registro de eventos, datos de entrenamiento, etc.), transparencia hacia usuarios (incluso obligando a notificar que se trata de IA) y aseguramiento de supervisión humana adecuada[25]. Antes de comercializar o desplegar un sistema de IA de alto riesgo, el proveedor deberá realizar una evaluación de conformidad que certifique que cumple todos esos requisitos.

En esencia, el RIA exige prueba y validación rigurosa de los sistemas de IA de alto riesgo antes de su uso real, lo cual impulsa implícitamente el uso de técnicas avanzadas de verificación. Si un algoritmo debe ser seguro, robusto, trazable y conforme a la normativa, el desarrollador se verá motivado a emplear herramientas formales que le den garantías fuertes de no violación de requisitos, por ejemplo, verificar que el error del modelo está acotado por debajo de un umbral, o que ciertas condiciones nunca desencadenan una respuesta prohibida.

De hecho, el RIA prevé que se elaboren estándares técnicos armonizados (a través de organismos de estandarización como CEN-CENELEC a nivel europeo, e ISO/IEC a nivel internacional) que detallen cómo cumplir tales requisitos. La adopción de estos estándares dará presunción de conformidad con la ley, es decir, si un sistema de IA se desarrolla siguiendo las normas técnicas X, Y, Z (por ejemplo, una norma sobre gestión del riesgo de sesgo en algoritmos), se presumirá que cumple las obligaciones correspondientes del RIA. Es previsible, entonces, que dichos estándares incorporen métodos formales de verificación al menos para ciertos tipos de sistemas críticos o por cuestiones sectoriales (IA médica, conducción autónoma, control industrial o justicia[26]), precisamente para asegurar las propiedades de seguridad exigidas. La tendencia regulatoria empuja por tanto a que la verificabilidad deje de ser opcional y pase a formar parte del estado del arte obligatorio en IA de alto riesgo.

---

24. Sobre esta cuestión, véase P. Simón Castellano, «Los sistemas de gestión de riesgos como obligación específica para los sistemas de inteligencia artificial de alto riesgo en el artículo 9 del Reglamento», en L. Cotino Hueso y P. Simón Castellano (dirs.), *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Cizur Menor, Aranzadi, 2024, pp. 535-564.

25. Véase A. Obregón Fernández y G. Lazcoz Moratinos, «La supervisión humana de los sistemas de inteligencia artificial de alto riesgo. Aportaciones desde el Derecho Internacional Humanitario y el Derecho de la Unión Europea», *Revista electrónica de estudios internacionales*, 42, 2021.

26. Véase J. Castellanos, «Garanzie giuridiche contro l'intelligenza artificiale. Possibilità e limiti della Cyberjustice», *i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale*, 13 (1). Numero Speciale AI and Justice, 2020, pp. 1-19.

Un signo de esta evolución es la reciente publicación de la norma ISO/IEC 42001:2023 (adoptada en Europa como EN ISO 42001 y en España como UNE-ISO 42001:2025). Se trata del primer estándar internacional enfocado en establecer un sistema de gestión de IA en las organizaciones, análogo a normas de calidad como ISO 9001 o de seguridad de sistemas como ISO 27001 pero orientado a IA. Dicha norma enfatiza principios de gobernanza, transparencia, ética y seguridad en el uso de IA. En particular, subraya la importancia de la trazabilidad —registrar y poder seguir el rastro de cómo funciona y ha sido entrenado el modelo— y la fiabilidad del algoritmo, incluyendo su verificación. Aunque ISO 42001 no prescribe técnicas específicas, sí establece que la organización debe implementar procesos para gestionar riesgos de la IA, garantizar la calidad de los datos y de los modelos, y asegurar propiedades como la equidad, la no discriminación y la explicabilidad a lo largo del ciclo de vida. En suma, el ecosistema jurídico-técnico se está alineando en exigir tanto *accountability* (rendición de cuentas) como evidencias formales de confiabilidad.

Para los sistemas de IA utilizados en el sector público, estos estándares jurídicos implicarán seguramente controles más rigurosos. Por ejemplo, como ya se ha indicado anteriormente, el RIA obliga a las administraciones que utilicen sistemas algorítmicos en ámbitos de alto riesgo (educación, empleo, servicios esenciales, justicia, etc.) a realizar evaluaciones de impacto en derechos fundamentales y a asegurar la intervención humana cuando proceda.

Es de esperar también un incremento de auditorías algorítmicas externas y posiblemente la exigencia de certificaciones especiales antes de autorizar el uso de ciertos algoritmos públicos. En este contexto, la verificación formal puede convertirse en una herramienta valiosa para cumplir con el principio de legalidad: si una ley prohíbe discriminación por sexo en cierta decisión automatizada, una verificación formal que demuestre la ausencia de tal discriminación en el modelo brindaría un nivel de garantía muy sólido de cumplimiento de la ley y, a su vez, defensa jurídica en caso de impugnación. Del mismo modo, el principio de motivación de las decisiones administrativas podría operacionalizarse exigiendo que el sistema sea capaz de generar una traza explicativa en cada caso; dicha propiedad podría incorporarse a la especificación formal del modelo y verificarse matemáticamente antes de desplegarlo. De este modo, los estándares legales emergentes tanto en la UE como en los Estados miembros están creando un entorno en el que la verificabilidad de la IA ya no es solo una aspiración técnica, sino un requisito ligado al cumplimiento normativo y al respeto de los derechos fundamentales.

4. Mecanismos de certificación y evaluación de conformidad

Dada la conjunción de altos estándares técnicos y exigencias legales, en el mercado internacional se están desarrollando mecanismos para certificar y evaluar la conformidad de los sistemas de IA de forma sistemática. En la práctica, esto significa establecer procesos parecidos a los de certificación de productos

tradicionales (por ejemplo, el marcado CE) pero aplicados a algoritmos y modelos de IA, incluyendo sus datos, su desempeño y su impacto[27].

El RIA introduce de lleno esta idea: todo sistema de IA de alto riesgo deberá someterse a una evaluación de conformidad antes de su puesta en servicio. En la mayoría de los casos, será el propio proveedor del sistema quien realice internamente esta evaluación siguiendo procedimientos aprobados, de manera similar a como un fabricante de maquinaria hace una evaluación interna según estándares y emite una declaración CE. En ciertos casos especialmente sensibles, podría requerirse la participación de un organismo notificado independiente, como podría ser una entidad de certificación acreditada, que audite el sistema[28]. La evaluación de conformidad abarcará la revisión de la documentación técnica del sistema de IA, incluyendo sus especificaciones, el registro de su proceso de entrenamiento, los datos utilizados, los resultados de pruebas de desempeño (precisión, tasas de error, etc.), y las salvaguardas implementadas para mitigar riesgos.

Aquí es donde la verificabilidad formal puede jugar un rol clave: si el desarrollador ha aplicado métodos formales para garantizar ciertas propiedades —tales como la ausencia de resultados inseguros fuera de un rango, o el cumplimiento de una regla normativa en todos los casos—, podrá aportar el certificado o prueba formal correspondiente al dosier técnico. Es esperable que los evaluadores de conformidad valoren muy positivamente —incluso lleguen a exigir en algunos ámbitos— la presentación de evidencias formales de seguridad o equidad algorítmica.

Un elemento fundamental de estos mecanismos son los estándares armonizados mencionados[29]. Actualmente, se trabaja en comités europeos e internacionales para desarrollar normas que cubran requisitos como la gestión del riesgo de IA, la calidad de datos de entrenamiento, la medición de sesgos, la explicabilidad, etc. Cuando tales normas estén disponibles, los proveedores podrán voluntariamente aplicarlas; al hacerlo, obtendrán una presunción de conformidad con las obligaciones legales correspondientes. Por ejemplo, si se aprueba un estándar CEN sobre «Requisitos de calidad del dato en sistemas de IA sanitarios» y un fabricante lo sigue al desarrollar su algoritmo médico, se presumirá que cumple el artículo del RIA que exige gestión de riesgo en datos. De igual modo, podemos imaginar futuros estándares sobre verificación formal de redes neuronales en automoción, y seguirlos podría implicar automáticamente que se satisfacen los requisitos de seguridad funcional. La presencia de estas normas hará más objetiva y transparente la certificación, evitando que cada empresa o administración tenga que improvisar cómo demostrar cumplimiento.

Además de la evaluación previa a la puesta en servicio, el RIA exige un sistema de seguimiento post-mercado: los proveedores deberán monitorear el

27. Sobre esta cuestión, véase A. Palma Ortigosa, «La evaluación de la conformidad en el diseño y producción de sistemas basados en inteligencia artificial en el contexto del «Nuevo Marco Legislativo», en L. Cotino Hueso y P. Simón Castellano (dirs.), *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Cizur Menor, Aranzadi, 2024, pp. 427-446.

28. Ibidem.

29. Ibidem.

funcionamiento de sus IA una vez desplegadas, recoger informes de fallos o incidentes y tomar medidas correctivas si aparecen problemas[30]. También en esta fase la verificabilidad puede ayudar, por ejemplo, incorporando monitores en tiempo de ejecución que detecten desviaciones inesperadas del comportamiento previsto (*runtime verification*).

A nivel voluntario, se espera emergerán sellos o certificaciones de calidad en IA. Por ejemplo, la norma ISO/IEC 42001 antes citada es certificable y las organizaciones ya pueden auditarse y obtener un certificado de que su sistema de gestión de IA cumple con los requisitos de la norma. También a nivel sectorial podrían verse esquemas de certificación y no es descabellado imaginar por ejemplo un sello europeo de «IA fiable en sanidad» que avale que un determinado software clínico ha pasado por validaciones independientes rigurosas. En la contratación pública, probablemente se requerirá o valorará que los sistemas proveídos tengan tales certificaciones o evaluaciones externas de conformidad.

En España, una iniciativa innovadora es la creación de un entorno controlado de pruebas o *sandbox regulatorio* para IA, mediante el Real Decreto 817/2023, de 8 de noviembre, que establece un entorno controlado de pruebas para el ensayo del cumplimiento de la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial. Este *sandbox* permite a empresas y administraciones voluntarias ensayar el cumplimiento del RIA en proyectos piloto supervisados, obteniendo guías prácticas sobre cómo lograr la conformidad. Se simulan procesos de evaluación de conformidad con expertos, lo que facilita identificar qué herramientas, incluidas las formales, son eficaces para asegurar que una IA pasa el examen regulatorio sin incurrir en riesgos no aceptable o tolerables desde la óptica del principio de proporcionalidad. Todo apunta a que la cultura de la certificación algorítmica irá así en aumento.

Los mecanismos de certificación y evaluación de conformidad son el vehículo mediante el cual las exigencias técnicas y jurídicas aterrizarán en la práctica. Un sistema de IA de alto riesgo no podrá simplemente autodeclararse seguro; deberá demostrar su seguridad y cumplimiento normativo ante auditores internos o externos de manera estructurada. La verificación formal, cuando esté al alcance, proporcionará un as bajo la manga para esa demostración, pues nada convence más a un evaluador que una garantía matemática acompañada de una trazabilidad completa del desarrollo. En perspectiva, la verificabilidad pasará de los laboratorios académicos a incorporarse en el arsenal obligatorio de herramientas para una IA confiable tanto en contextos públicos como privados. Reguladores y estándares la están amparando, operadores y desarrolladores están adquiriendo experiencia en su uso, y la sociedad en su conjunto se beneficiará de sistemas de IA más seguros, justos y transparentes gracias a ello.

30. Sobre esta cuestión, véase I. Salazar y M. A. Liébanas, «Vigilancia poscomercialización en los sistemas de inteligencia artificial de alto riesgo en el Reglamento: Descripción, medidas y casos de uso», en L. Cotino Hueso y P. Simón Castellano (dirs.), *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Cizur Menor, Aranzadi, 2024, pp. 743-754.

Convergemos así hacia un modelo de IA responsable donde *lo que no se pueda probar, no se pueda desplegar*, al menos, en entornos sensibles, cumpliendo el adagio de que, en tecnología y derecho, la confianza solo puede cimentarse sobre evidencia sólida.

IV. MODELOS DE SUPERVISIÓN Y RESPONSABILIDAD JURÍDICA EN EL SECTOR PÚBLICO

El despliegue de sistemas y modelos de IA en la Administración pública plantea importantes retos jurídicos en materia de supervisión y responsabilidad. A continuación, se analizan los mecanismos vigentes y las perspectivas de futuro para garantizar, por un lado, una supervisión humana e institucional adecuada de la IA, y por otro, un efectivo control democrático y jurisdiccional, así como la responsabilidad jurídica y rendición de cuentas de estos sistemas en el sector público, con especial referencia al marco europeo y español.

1. Supervisión humana e institucional

Un principio central en la gobernanza de la IA es asegurar que su uso esté sujeto a control humano y organizativo. La normativa europea emergente consagra la necesidad de una supervisión humana sobre los sistemas de IA de alto riesgo e impone para dichos sistemas controles estrictos, entre ellos la obligatoriedad de la supervisión humana durante su funcionamiento[31]. Esto significa que, por diseño, las decisiones algorítmicas críticas no pueden quedar totalmente en manos de máquinas, por lo que debe haber operadores humanos capaces de vigilar, interpretar y, en su caso, corregir o anular la actuación de la IA y sus eventuales alucinaciones. Esta fórmula llamada *human-in-the-loop* garantiza que se respeten los criterios jurídicos y éticos, evitando una automatización irreflexiva que vulnere derechos fundamentales.

En el ámbito español, el marco jurídico ya reconocía el papel del ser humano en las decisiones automatizadas. Como se ha indicado, la Ley 40/2015 exige que todo acto administrativo automatizado designe previamente una autoridad competente para la programación, mantenimiento, supervisión y control de calidad del sistema, e incluso su auditoría, indicando expresamente qué órgano será responsable a efectos de impugnación. Una previsión legal que asegura que detrás de cada algoritmo público haya responsables identificables encargados de su correcto funcionamiento y de responder jurídicamente por él. De igual modo, la Ley 15/2022 prohíbe la discriminación algorítmica y exige garantizar la transparencia, explicabilidad y supervisión humana de los sistemas de IA que puedan incidir en derechos o causar impactos significativos. De esta forma, en la práctica administrativa diaria, la normativa española subraya que la última

31. Véase A. Obregón Fernández y G. Lazcoz Moratinos, «La supervisión humana… *op. cit.*

palabra y comprensión de la decisión deben corresponder a personas, no a cajas negras algorítmicas.

Junto con la supervisión humana individual, se desarrollan mecanismos institucionales de control. La Oficina europea de la IA y la AESIA tiene atribuidas funciones de supervisión amplias, para garantizar que tanto las empresas como las administraciones públicas cumplen la normativa, así como para evaluar y clasificar los sistemas de IA según su riesgo, certificar algoritmos y sancionar infracciones, así como asesorar y guiar en buenas prácticas. El RIA, como se ha indicado anteriormente, exige que las administraciones que quieran usar IA de alto riesgo realicen una evaluación de impacto en derechos fundamentales antes de implantarlas, identificando riesgos y medidas de mitigación. Esta obligación de evaluación previa, inédita hasta ahora, intensifica los principios de transparencia y responsabilidad en el sector público. Así como la necesidad de reportar los informes y resultados de las evaluaciones de impacto a las autoridades de control.

Por todo ello, de cara al futuro, es de esperar que en la práctica se produzca un reforzamiento de la supervisión humana e institucional. Doctrinas como la reserva de humanidad sostienen que ciertas decisiones administrativas —especialmente las discrecionales, que requieren empatía y ponderación de valores— no deben delegarse por completo en algoritmos, sino quedar reservadas a los humanos[32]. Tal principio podría consagrarse legalmente para garantizar que ámbitos sensibles (justicia, sanciones, prestaciones sociales, etc.) cuenten siempre con intervención humana en última instancia. Adicionalmente, los marcos internacionales apuntan a supervisiones independientes: el reciente y ya citado Convenio Marco del Consejo de Europa sobre IA obliga a cada Estado parte a establecer mecanismos independientes de supervisión que vigilen el cumplimiento del tratado, sensibilicen a la ciudadanía y promuevan el debate público informado sobre el uso de la IA. En suma, en el ámbito de las Administraciones, la combinación de controles humanos individuales (funcionarios preparados para monitorizar algoritmos) y estructuras institucionales especializadas (agencias nacionales y comités internacionales) será clave para una IA pública ética, segura y alineada con el Estado de derecho.

2. Control democrático y jurisdiccional

El uso de IA en el sector público debe someterse ineludiblemente al control democrático, por parte de la sociedad y sus representantes, y al control jurisdiccional, por parte de los tribunales, para asegurar que la automatización no socave los principios de transparencia, participación, legalidad y tutela judicial efectiva propios de un Estado de Derecho.

Un primer pilar es la transparencia algorítmica hacia la ciudadanía. Democracia y opacidad son conceptos antagónicos: los administrados tienen derecho

32. J. Ponce Solé, *El Reglamento de Inteligencia Artificial... ob. cit.*

a saber que una decisión que les afecta ha sido tomada o apoyada por un algoritmo, y cómo funciona este.

En España, el principio de transparencia algorítmica se ha concretado en normas pioneras ya señaladas anteriormente, como el artículo 41 de la Ley 40/2015 o el artículo 23 de la Ley 15/2022 de igualdad de trato y no discriminación. Esto permite un escrutinio público ex ante de la presencia de IA en la gestión pública. Además, la sociedad civil y los medios pueden utilizar las leyes de acceso a la información para fiscalizar algoritmos públicos.

Un hito reciente es el de ya citado caso BOSCO que ha resuelto la sentencia del Tribunal Supremo español de 17 de septiembre de 2025. Con independencia de la valoración jurídica del fallo, este consagra que la opacidad tecnológica no puede justificar la negación de información, ni la propiedad intelectual ni el secreto comercial ni consideraciones de seguridad pueden prevalecer absolutamente sobre el interés público en saber cómo decide una IA gubernamental que toma decisiones o ayuda a tomar decisiones que producen efectos sobre los administrados. En términos prácticos, esta jurisprudencia abre la puerta a una mayor fiscalización democrática de la IA, ya que solo con transparencia puede haber debate público y control externo.

Junto a la transparencia, el control democrático se ejerce vía marco legal y parlamentario. Las Cortes y parlamentos deben establecer límites claros al uso de IA en la esfera pública mediante leyes que definan qué está permitido y qué prohibido. El propio RIA es fruto de un proceso democrático y contiene salvaguardas alineadas con valores de la UE. A nivel nacional, y a la espera de la aprobación del anteproyecto de Ley para el buen uso y la gobernanza de la IA, ya se empieza a hablar de la reserva de ley algorítmica. Por analogía a la reserva de ley en materia de derechos fundamentales, ciertos usos de IA altamente intrusivos o discrecionales deberían requerir una ley formal que los habilite y regule sus condiciones. De este modo, se evita que decisiones de gran impacto se introduzcan vía meras decisiones administrativas o contratos con proveedores tecnológicos, eludiendo el debate legislativo.

Por último, el control jurisdiccional garantiza que las personas afectadas por decisiones automatizadas dispongan de recursos efectivos ante los tribunales. En el ordenamiento español y europeo, ninguna decisión de la Administración —sea tomada por un funcionario o por un algoritmo— está exenta de revisión judicial. El artículo 41.2 de la Ley 40/2015, antes citado, obliga a designar un órgano responsable para impugnaciones en caso de actos automatizados, de forma que el ciudadano siempre tiene un sujeto público al que demandar o recurrir. Los jueces, a su vez, están empezando a desarrollar doctrinas para enfrentarse a las cajas negras algorítmicas y, como efecto de la sentencia del TS en el caso BOSCO, pueden requerir a la Administración la aportación, dependiendo de las circunstancias de cada caso concreto, de los criterios de decisión del algoritmo en un proceso o incluso su código íntegro, para verificar si cumple la legalidad. Si la Administración se negara, se pondría en entredicho su acto por falta de motivación o indefensión del ciudadano. La última garantía frente a posibles abusos o errores de la IA pública reside en jueces indepen-

dientes capaces de escrutar la tecnología a la luz del Derecho. Este control jurisdiccional, junto con la fiscalización democrática, actúa como contrapeso imprescindible para que la IA se utilice al servicio de la ciudadanía y no al margen de ella.

3. Responsabilidad jurídica y rendición de cuentas.

La introducción de sistemas de IA en la gestión pública no diluye ni elimina las responsabilidades legales; por el contrario, obliga a redefinir cómo se atribuyen las responsabilidades y se garantiza la rendición de cuentas cuando algo sale mal o se vulneran derechos. En el estado actual del Derecho, la responsabilidad jurídica por los actos de IA en el sector público recae sobre las personas jurídicas y físicas involucradas —nunca sobre la máquina—, siguiendo el principio general de que los sistemas de IA son herramientas bajo control humano.

Las Administraciones públicas, como entes de derecho, responden patrimonialmente de los daños que causen sus servicios, con base en el artículo 106.2 de la Constitución y las leyes de procedimiento administrativo. Esto significa que si un algoritmo utilizado por una Administración comete un error perjudicial (por ejemplo, deniega indebidamente una prestación o produce un sesgo discriminatorio en un proceso selectivo), el afectado tiene derecho a reclamar una reparación económica al organismo público responsable, igual que lo haría ante un error humano.

El uso de IA no es excusa para eludir esta responsabilidad; al contrario, la Administración debe extremar cautelas, puesto que delegar decisiones en sistemas complejos conlleva el deber de diligencia en su selección, supervisión y mantenimiento. De nuevo, el caso BOSCO es un buen botón de muestra, en el que Alto Tribunal señaló que la transparencia y escrutinio público de algoritmos incentiva a la Administración a extremar las cautelas de seguridad en el diseño y control del programa, permitiendo además detectar y corregir vulnerabilidades a tiempo. Se configura, así, como un estándar de responsabilidad proactiva: las Administraciones deben anticipar los posibles riesgos de la IA (mediante auditorías, evaluaciones de impacto, etc.) y están obligadas a responder si dichos riesgos se materializan en daños.

En el contexto público, más allá de la responsabilidad civil, cobra relevancia esta noción de rendición de cuentas o *accountability*. Este concepto abarca la idea de que los gestores públicos deben dar cuenta y asumir las consecuencias de sus decisiones automatizadas ante instancias de control político, social y legal. La rendición de cuentas también implica que las Administraciones establezcan procedimientos internos de control: por ejemplo, comités éticos o comités algorítmicos que revisen periódicamente los sistemas, canales de reclamación específicos para ciudadanos que sospechen errores algorítmicos, y planes de contingencia para suspender o corregir una IA si empieza a arrojar resultados injustos.

Hacia el futuro, es previsible que la cultura de la rendición de cuentas algorítmica se afiance. El desarrollo de estándares de auditoría algorítmica independientes permitirá evaluar la equidad y fiabilidad de los sistemas de IA públicos de forma regular

V. CONCLUSIONES

Primera. Especificaciones de seguridad como puente técnico-jurídico.

A lo largo del trabajo hemos analizado como un aspecto fundamental que se establezcan especificaciones de seguridad que integren desde el diseño técnico de la IA los requisitos legales y garantías jurídicas exigibles. Estas especificaciones actúan como puente entre la ingeniería del sistema y las normas, traduciendo obligaciones legales en controles técnicos concretos. En este sentido, estándares recientes como ISO/IEC 42001:2023 enfatizan la gestión del ciclo de vida de las IA incorporando requisitos legales desde la fase de diseño y desarrollo, en línea con las nuevas regulaciones europeas. Gracias a ello, se refuerza la conformidad por diseño, asegurando que las soluciones de IA públicas nacen respetando la normativa vigente y los derechos fundamentales.

Segunda. Modelos verificables y validación formal como pilares de confianza.

Una IA confiable en la Administración requiere modelos verificables, sujetos a validación formal rigurosa antes de su despliegue. La verificación formal y las pruebas exhaustivas permiten demostrar que el algoritmo cumple las propiedades de seguridad, exactitud y fiabilidad esperadas, lo cual es indispensable para su auditoría técnica y jurídica. De hecho, el marco regulatorio europeo ya subraya estos extremos y el RIA exige para los sistemas de alto riesgo una documentación técnica detallada, registro de actividades, trazabilidad de decisiones y evaluaciones continuas, incluyendo auditorías independientes de los modelos. Tales medidas evidencian que la validación formal de los algoritmos, verificando matemáticamente su robustez y ausencia de sesgos o fallos, es un pilar esencial para garantizar sistemas de IA seguros y auditables en el sector público.

Tercera. Estándares técnicos y marcos legales armonizados.

La construcción de una IA fiable pasa por combinar estándares tecnológicos sólidos con una regulación jurídica uniforme. En el plano técnico, la adopción de normas internacionales como la ISO/IEC 42001:2023 proporciona un marco común para gestionar los riesgos y la calidad de los algoritmos. Al mismo tiempo, instrumentos jurídicos supranacionales como el RIA y el Convenio Marco del Consejo de Europa sobre IA están estableciendo principios

y obligaciones parcialmente homogéneas que deberán seguir todos los actores públicos y privados. La existencia de estos referentes armonizados —tanto técnicos (ISO, NIST, IEEE, etc.) como legales— resulta crucial para evitar lagunas o divergencias entre jurisdicciones, facilitando un ecosistema donde los sistemas de IA en las Administraciones Públicas respeten estándares de seguridad, ética y derechos humanos de forma coherente a nivel internacional.

Cuarta. Supervisión humana, control democrático y jurisdiccional.

La confiabilidad de la IA pública demanda mecanismos robustos de supervisión por parte de personas e instituciones. No basta con la automatización técnica: se requiere la intervención humana significativa que monitorice el funcionamiento del algoritmo y pueda intervenir ante errores o desviaciones. En esta línea, el RIA impone obligaciones estrictas de gestión de riesgos, transparencia algorítmica y presencia de supervisores humanos para los sistemas de alto riesgo, reconociendo que el factor humano es irreemplazable para garantizar el respeto a los valores constitucionales. Junto a la supervisión interna, es imprescindible un control democrático e institucional, esto es, órganos independientes que evalúen periódicamente estos sistemas, así como la participación y escrutinio público que eviten una caja negra burocrática. Igualmente, debe asegurarse el control jurisdiccional efectivo de la IA administrativa —por vía de recursos ante tribunales—, de forma que existan remedios legales cuando un sistema automatizado vulnere derechos o tome decisiones arbitrarias. Solo con supervisión humana e institucional constante, sumada al control democrático y judicial, puede garantizarse que la IA actúe al servicio del Estado de Derecho y no en su contra.

Quinta. Hacia la certificación, transparencia y rendición de cuentas.

De cara al futuro inmediato, las propuestas convergen en reforzar la certificación, la transparencia y la rendición de cuentas de los sistemas algorítmicos en el sector público. Se plantea establecer esquemas de certificación independiente de algoritmos (similares a certificaciones de calidad), de modo que antes de su adopción oficial se acredite su fiabilidad, seguridad y cumplimiento normativo. En este aspecto, la propia ISO/IEC 42001 prevé un sistema voluntario pero certificable de gestión de IA, lo que marca el inicio de marcos globales para auditar y certificar algoritmos.

Asimismo, las iniciativas legislativas más recientes insisten en la transparencia: obligar a las administraciones a divulgar información significativa sobre el funcionamiento de sus IA (por ejemplo, a través de registros de algoritmos o evaluaciones de impacto algorítmico públicas) para que puedan ser escrutadas. Con ello se promueve una mayor rendición de cuentas, exigiendo identificar responsables de cada sistema y permitiendo auditorías externas periódicas. Avanzar hacia IA auditables, transparentes y certificadas, con obligaciones de explicabilidad y responsabilidades bien definidas, es clave para fortalecer la

confianza ciudadana y asegurar que las Administraciones Públicas puedan ser plenamente responsables de las decisiones automatizadas que tomen en el ejercicio de sus funciones.

BIBLIOGRAFÍA

BAK, S. «Verifying Neural Networks to Avoid Air-to-Air Collisions», *AI Innovation Institute, Stony Brook University*, 2024, disponible en https://ai.stonybrook.edu/about-us/News/verifying-neural-networks-avoid-air-air-collisions

BOIX PALOP, A. «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*, 1, 2020, pp. 223-269.

BUSTOS GISBERT, R. «El constitucionalista europeo ante la inteligencia artificial: reflexiones metodológicas de un recién llegado», *Revista Española de Derecho Constitucional*, 131, 2024, pp. 146-178.

CASTELLANOS, J. «Garanzie giuridiche contro l'intelligenza artificiale. Possibilità e limiti della Cyberjustice», *i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale*, 13 (1). Numero Speciale AI and Justice, 2020, pp. 1-19.

CHAVELI DONET, E. «La evaluación de impacto de derechos fundamentales por quienes despliegan sistemas de inteligencia artificial en el Reglamento», en L. Cotino Hueso y P. Simón Castellano (dirs.), *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Cizur Menor, Aranzadi, 2024, pp. 495-533.

CERRILLO MARTÍNEZ, A. «La transparencia de los algoritmos que utilizan las administraciones públicas», *Anuario de Transparencia Local*, 3, 2020, pp. 41-78.

CERRILLO MARTÍNEZ, A. y VELASCO RICO, C. I. «La transparencia algorítmica», *Working Papers DIGITAPIA*, 1 (3), 2025, pp. 25-41.

COTINO HUESO, L. «"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», *La Ley privacidad*, 4, 2020.

COTINO HUESO, L. y SIMÓN CASTELLANO, P. (dirs.), *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Cizur Menor, Aranzadi, 2024.

COTINO HUESO, L. «Cómo abordar jurídicamente el impacto de la inteligencia artificial en los derechos fundamentales», en M. E. Casas Baamonde (Dir.) y D. Pérez del Prado (Coord.), *Derecho y tecnologías*, Madrid, Fundación Ramón Areces, 2025, pp. 123-176.

DALRYMPLE, D., J. SKALSE, Y. BENGIO, S. RUSSELL, M. TEGMARK, S. SESHIA, S. Omohundro, C. Szegedy, A. Abate, J. Halpern, C. Barrett, D. Zhao, B. Goldhaber y N. Ammann, «Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems», *Technical Report UCB/EECS-*

2024-45, Departamento de Ingeniería Eléctrica y Ciencias de la Computación, Universidad de California, Berkeley, 4 de mayo de 2024, pp. 1-30.

Obregón Fernández, A. y Lazcoz Moratinos, G. «La supervisión humana de los sistemas de inteligencia artificial de alto riesgo. Aportaciones desde el Derecho Internacional Humanitario y el Derecho de la Unión Europea», *Revista electrónica de estudios internacionales*, 42, 2021.

Palma Ortigosa, A. «La evaluación de la conformidad en el diseño y producción de sistemas basados en inteligencia artificial en el contexto del «Nuevo Marco Legislativo», en L. Cotino Hueso y P. Simón Castellano (dirs.), *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Cizur Menor, Aranzadi, 2024, pp. 427-446.

Ponce Solé, J. *El Reglamento de Inteligencia Artificial de la Unión Europea de 2024, el derecho a una buena administración digital y su control judicial en España*, Marcial Pons, Madrid, 2024.

Russell S. J. y Norvig, P. *Artificial Intelligence: A Modern Approach*, Prentice Hall, New Jersey, 1995.

Salazar, I. y Liébanas, M. A. «Vigilancia poscomercialización en los sistemas de inteligencia artificial de alto riesgo en el Reglamento: Descripción, medidas y casos de uso», en L. Cotino Hueso y P. Simón Castellano (dirs.), *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Cizur Menor, Aranzadi, 2024, pp. 743-754.

Simón Castellano, P. *La evaluación de impacto algorítmico en los derechos fundamentales*, Aranzadi, Cizur Menor, 2023;

Simón Castellano, P. «Taxonomía de las garantías jurídicas en el empleo de los sistemas de inteligencia artificial», *Revista de Derecho Político*, 117, 2023, pp. 153-196.

Simón Castellano, P. «Los sistemas de gestión de riesgos como obligación específica para los sistemas de inteligencia artificial de alto riesgo en el artículo 9 del Reglamento», en L. Cotino Hueso y P. Simón Castellano (dirs.), *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Cizur Menor, Aranzadi, 2024, pp. 535-564.

Simón Castellano, P. «L'avaluació d'impacte en els drets fonamentals en l'ús d'intel·ligència artificial en el sector públic: models i metodologies en perspectiva comparada», *Revista Catalana de Dret Públic*, 71, 2025.

Valero Torrijos, J. «Las garantías jurídicas de la inteligencia artificial en la actividad administrativa desde la perspectiva de la buena administración», *Revista catalana de dret públic*, 58, 2019, pp. 82-96.

Transparencia algorítmica y seguridad digital en la administración pública: construcción de un régimen operativo de compatibilización en los sistemas de inteligencia artificial del estado colombiano como referente iberoamericano

Marco Emilio Sánchez Acevedo[1]
Universidad Católica de Colombia

SUMARIO: I. INTRODUCCIÓN. II. TENSIONES CONCEPTUALES Y PRÁCTICAS. 1. La transparencia algorítmica. 2. Facetas activa y pasiva de la transparencia algorítmica (art. 7 de la directiva; jurisprudencia constitucional). 3. Seguridad digital como condición de confianza pública (referencia a la ley 1581/2012 y estándares internacionales como UNESCO). 4. El dilema del código fuente: entre el derecho de acceso (art. 74 CP + ley 1712/2014) y la protección frente a riesgos (art. 17 y 19 directiva, test de daño). III. MECANISMOS DE

1. El presente capítulo de libro de investigación es resultado del trabajo adelantado por el autor dentro del proyecto de investigación «*Análisis de datos para la toma de decisiones espaciales e inteligencia artificial para la administración»,* vinculado al grupo de investigación Derecho Público y Tecnologías, categoría Colciencias A1, Universidad Católica de Colombia durante el año 2024 -2025. De igual forma, resultado de la colaboración con el grupo de investigación Régimen jurídico constitucional de las libertades, el gobierno abierto y el uso de las nuevas tecnologías (UVEG- GIUV2016-270), así como del proyecto nacional proyecto de I+D+i Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas [2023-2026] (PID2022-136439OB-I00), financiado por MCIN/AEI/10.13039/501100011033/ y «FEDER Una manera de hacer Europa».

COMPATIBILIZACIÓN. 1. Explicabilidad y lenguaje claro. 2. Análisis de impacto y auditorías. 3. Canales de revisión y objeción. 4. Protección de datos y ciberseguridad por diseño. 5. Test de daño y ponderación de riesgos en la transparencia algorítmica. IV. CRITERIOS E INDICADORES DEL CUADRO OPERATIVO DE COMPATIBILIZACIÓN. V. DISCUSIÓN CRÍTICA. VI. CONCLUSIONES.

I. INTRODUCCIÓN

El despliegue de sistemas de decisión automatizada (SDA) y de sistemas de inteligencia artificial (IA) en la administración pública intensifica la exigencia de control democrático sobre su diseño, uso y efectos. La Sentencia T-067 de 2025 expedida por la Corte Constitucional Colombiana reconoce la «transparencia algorítmica» (en sus facetas activa y pasiva) como elemento del «núcleo esencial del derecho de acceso a la información pública», y ordena la expedición de «estándares mínimos sobre transparencia algorítmica» para las entidades estatales. En otras palabras, la Corte Constitucional dejó claro que ya no basta con que la administración responda solicitudes de información de manera formal: ahora, en la era digital, debe garantizar que los ciudadanos comprendan cómo funcionan los algoritmos y sistemas de IA que impactan directamente sus vidas. Esto exige explicar de forma accesible —pero sin perder precisión técnica— cómo se procesan datos sensibles, cómo se llega a una decisión y cuáles son los riesgos asociados. La transparencia deja de ser un ideal abstracto y se convierte en una herramienta práctica para fortalecer la confianza democrática frente al uso de tecnologías de inteligencia artificial en lo público.

En ejecución de ese mandato, la «Directiva Conjunta No. 007 de 30 de septiembre de 2025» expedida por Procuraduría General de la Nación y Defensoría del Pueblo (con apoyo técnico de la Agencia Nacional Digital) fija «estándares mínimos de transparencia algorítmica aplicables a los sujetos obligados en el artículo 5 de la Ley 1712 de 2014», desarrollando lo dispuesto por la Corte Constitucional. Esta Directiva representa, en términos prácticos, una hoja de ruta para todas las entidades públicas y privadas que ejercen funciones públicas. Establece qué información debe publicarse, cómo debe explicarse y bajo qué condiciones puede limitarse el acceso. Al hacerlo, no solo desarrolla la jurisprudencia de la Corte, sino que traduce principios generales en obligaciones operativas, imponiendo plazos, canales de publicación y contenidos mínimos que no pueden obviarse.

El núcleo del problema consiste en compatibilizar, de un lado, las obligaciones de «transparencia activa y pasiva» (arts. 7 y 8 de la Directiva, que ordenan la «publicación clara, accesible y periódica de información relevante sobre el diseño, propósito, funcionamiento y efectos de los sistemas») y, de otro, las garantías de «seguridad digital» e integridad de sistemas y datos, incluidas las medidas frente a riesgos de explotación y las reservas legalmente autorizadas.

Aquí se ubica la tensión central: la ciudadanía tiene derecho a entender cómo un sistema de IA asigna un subsidio, predice un riesgo o decide sobre el acceso a un servicio; al mismo tiempo, revelar en exceso la arquitectura o el

código de esos sistemas podría abrir la puerta a ataques informáticos, manipulación maliciosa o filtraciones de datos sensibles. La cuestión, entonces, no es si debe haber transparencia, sino hasta qué punto y en qué condiciones puede garantizarse sin debilitar la seguridad del Estado ni la protección de derechos fundamentales.

Desde la transparencia, la Directiva exige contenidos mínimos de divulgación: «nombre del sistema, objetivo, función institucional, estado, tratamiento de datos personales, tipo de datos utilizados, desarrollador, contacto para objeciones y contacto para más información» (art. 9). Esto significa que un ciudadano no necesita ser experto en programación para acceder a información básica sobre el sistema que le afecta. Por ejemplo, si se trata de un algoritmo o sistema de IA que prioriza pacientes en la atención hospitalaria, la entidad debe informar claramente su finalidad, qué tipo de datos procesa (biométricos, clínicos, georreferenciación), en qué etapa se encuentra (piloto o implementación) y a quién se puede acudir en caso de inconformidad. Estos requisitos buscan materializar la transparencia como práctica cotidiana y no como declaración abstracta.

Desde la seguridad digital, el régimen establece un esquema escalonado: en caso de solicitudes sobre «código fuente», la entidad debe evaluar «caso a caso, con herramientas como el test de daño, si para conseguir una mayor transparencia algorítmica significativa opta por publicar dicho código o, en su lugar, proporciona una explicación significativa en lenguaje claro» (art. 17). Si se entrega el código, deben cumplirse «condiciones de entrega» (art. 18) orientadas a prevenir daños a la «seguridad informática o a la integridad de sistemas o datos». Este equilibrio supone un reto: entregar el código fuente permite auditorías independientes, pero también puede exponer vulnerabilidades críticas si cae en manos equivocadas. Por ello, la Directiva privilegia la entrega de explicaciones significativas, como detallar los criterios que usa el sistema para clasificar perfiles o los parámetros que influyen en una decisión, con lo cual se garantiza control democrático sin comprometer la seguridad informática. Cuando excepcionalmente se entrega código, este debe ir acompañado de actas de entrega y cláusulas de no comercialización ni uso indebido.

La «negativa de acceso» solo procede cuando exista «causal de reserva legítima, expresamente prevista en los artículos 18 y 19 de la Ley 1712 de 2014», y el sujeto obligado demuestre, mediante «test de daño», que el perjuicio probable y específico excede el interés público del acceso (art. 19 de la Directiva).

Esto implica que las entidades no pueden negar información de forma discrecional, sino que deben justificar con pruebas que publicar ciertos datos causaría un daño real, probable y específico (por ejemplo, comprometer operaciones de seguridad nacional). Además, la restricción nunca puede ser más amplia de lo estrictamente necesario: si es posible entregar información parcial, anonimizada o adaptada en lenguaje claro, la entidad está obligada a hacerlo antes de negar por completo el acceso.

En este contexto, la investigación se estructura con base en la metodología cualitativa propuesta por Taylor y Bogdan, orientada a comprender fenómenos

sociales desde la perspectiva de los actores y a través de un análisis inductivo de la realidad. Esta metodología permite interpretar cómo los marcos normativos, jurisprudenciales y técnicos sobre transparencia y seguridad digital son apropiados y aplicados por las entidades estatales, generando hallazgos que no se limitan a la letra de la norma, sino que incorporan las dinámicas prácticas de su implementación.

De acuerdo con esta perspectiva, se privilegia una aproximación analítica que conecta las disposiciones jurídicas con su contexto de aplicación, observando tanto los discursos institucionales como las tensiones concretas que surgen en la gestión de sistemas algorítmicos e inteligencia artificial. Esto hace posible no solo describir la normativa vigente, sino también identificar los puntos de fricción y las vías de armonización entre transparencia algorítmica y seguridad digital, en clave de gobernanza pública.

La pregunta orientadora que guiará este capítulo es: ¿De qué manera puede compatibilizarse el derecho a la transparencia algorítmica con la seguridad digital en el uso de sistemas de decisión automatizada y de inteligencia artificial por parte de las autoridades públicas, a la luz de la Sentencia T-067 de 2025 de la Corte Constitucional Colombiana y de los estándares fijados por la Procuraduría General de la Nación y la Defensoría del Pueblo? Esta pregunta articula dos dimensiones que tradicionalmente se perciben como opuestas. La tarea del capítulo será demostrar que transparencia y seguridad no son polos irreconciliables, sino elementos que, al integrarse adecuadamente, se refuerzan mutuamente: mayor transparencia contribuye a detectar sesgos y vulnerabilidades, y una seguridad digital sólida asegura la confiabilidad de los procesos transparentes. El presente capítulo tiene como propósito desarrollar un marco analítico y operativo sobre la transparencia algorítmica en los sistemas de decisión automatizada (SDA) e inteligencia artificial (IA), atendiendo a su configuración normativa y jurisprudencial reciente. En particular, se examinará el alcance del deber de transparencia —tanto en su faceta activa como pasiva— y su articulación con el derecho fundamental de acceso a la información pública, conforme a la Sentencia T-067/25 y a la Directiva 007/25. Desde esta base conceptual, el capítulo avanzará hacia la identificación de criterios técnicos y jurídicos que permitan decidir, en casos concretos, entre la publicación del código fuente o la provisión de explicaciones significativas, aplicando la metodología del test de daño y considerando las condiciones de entrega, confidencialidad y proporcionalidad de la información divulgada.

De manera integrada, el análisis incorporará las obligaciones de transparencia con las salvaguardas de seguridad digital y protección de datos personales, atendiendo a elementos como la trazabilidad de decisiones algorítmicas, la gestión de riesgos, el uso de lenguaje claro y la actualización permanente del ciclo de vida de los sistemas (arts. 5, 9, 13 y 22 de la Directiva). Finalmente, el capítulo concluirá con la formulación de un cuadro operativo de compatibilización, compuesto por criterios, pruebas y medidas aplicables tanto ex ante —a través de análisis de impacto algorítmico— como ex post, mediante auditorías, reportes y mecanismos de actualización (arts. 14, 20–21 y 28). Esta estructura permi-

tirá avanzar de lo conceptual a lo práctico, consolidando un enfoque progresivo y aplicable para la gestión pública de sistemas algorítmicos transparentes, seguros y responsables.

II. TENSIONES CONCEPTUALES Y PRÁCTICAS

El análisis de la transparencia algorítmica y la seguridad digital exige identificar, antes que los mecanismos de compatibilización, las tensiones estructurales que subyacen a su relación. Estas tensiones no se limitan a un conflicto entre apertura informativa y reserva tecnológica; reflejan la compleja interacción entre principios constitucionales, obligaciones internacionales y realidades técnicas de la administración digital contemporánea. La Sentencia T-067 de 2025 y la Directiva Conjunta 007 de 2025 introducen en el ordenamiento colombiano un modelo de transparencia algorítmica significativa, fundado en la accesibilidad, la explicabilidad y la máxima divulgación, pero al mismo tiempo subordinado al principio de seguridad digital consagrado en el Decreto 767 de 2022 como elemento esencial de la Política de Gobierno Digital. De ahí que el objeto de esta sección sea desentrañar, con base en estos marcos normativos, los puntos de fricción y complementariedad entre transparencia y seguridad, distinguiendo sus conceptos, facetas y alcances jurídicos.

En este sentido, los apartados que siguen examinan cuatro ejes críticos de la relación: primero, la definición normativa y jurisprudencial de la transparencia algorítmica como principio constitucional emergente; segundo, la dualidad entre transparencia activa y pasiva como dimensiones complementarias del derecho de acceso a la información; tercero, la seguridad digital como condición de confianza pública, que articula la protección de datos, la resiliencia tecnológica y la legalidad del uso de sistemas automatizados; y finalmente, el dilema del código fuente, donde se materializa la tensión entre la apertura informativa y la preservación de los intereses públicos y privados que concurren en el entorno digital. El propósito de esta sección es, por tanto, establecer las bases conceptuales para los mecanismos de compatibilización que serán desarrollados en la siguiente parte del capítulo.

1. La transparencia algorítmica

La Sentencia T-067 de 2025 definió la «transparencia algorítmica» como «la disponibilidad de información sobre sistemas de algoritmos que permite conocer su operación y valorar su rendimiento», destacando que su finalidad constitucional es «democratizar el funcionamiento interno de un sistema de toma de decisión automatizado, para que sea entendible por quienes se ven afectados por su puesta en marcha y operación». En concordancia, el artículo 2.h de la Directiva Conjunta 007 de 2025 precisó que «lo que se busca con la transparencia algorítmica es que el público en general pueda comprender cómo los siste-

mas de toma de decisiones automatizadas (SDA) procesan los datos que capturan y cómo toman decisiones que afectan la vida de las personas».

Estas definiciones, en el plano jurídico y normativo, sitúan la transparencia algorítmica como una condición del derecho fundamental de acceso a la información pública (art. 74 CP y Ley 1712 de 2014), pero con un alcance adaptado a los retos de la inteligencia artificial y los sistemas automatizados. No se trata solamente de divulgar datos sobre la existencia de un sistema, sino de garantizar que las personas puedan comprender su funcionamiento, evaluar su legitimidad y, cuando corresponda, desafiar sus decisiones. Además, la Directiva 007 de 2025 refuerza que la transparencia algorítmica no se limita a los SDA, sino que se extiende también a los sistemas de inteligencia artificial (IA) que procesan datos y generan decisiones con efectos jurídicos o materiales sobre la ciudadanía (art. 6). Esto enlaza directamente con el objeto de estudio del presente capítulo: examinar cómo se armonizan las exigencias de transparencia con las salvaguardas de seguridad digital en el ámbito público, en un contexto donde tanto SDA como IA se convierten en instrumentos de gestión estatal.

En términos más claros, hablar de transparencia algorítmica significa exigir al Estado que explique cómo funcionan las «máquinas de decidir» que afectan la vida cotidiana de los ciudadanos. Si un algoritmo determina el orden en el que se atienden pacientes en un hospital, o si un sistema de IA proyecta quién tiene más probabilidad de acceder a un subsidio, la persona afectada no puede quedarse con una explicación vaga del tipo «el sistema decidió así». La transparencia exige detallar, en un lenguaje comprensible, qué variables se usaron, cómo se procesaron y por qué la decisión resultó de esa manera.

Este concepto conecta de manera inmediata con la idea central, que es identificar los mecanismos jurídicos, técnicos y procedimentales que permitan compatibilizar transparencia y seguridad digital. De hecho, la definición misma de transparencia algorítmica establece la base para el análisis: si el fin último es que la ciudadanía comprenda y controle cómo funcionan los sistemas de decisión automatizada e inteligencia artificial, entonces el desafío es asegurar que ese acceso a la información no comprometa la confidencialidad, integridad y disponibilidad de los sistemas ni exponga vulnerabilidades que puedan ser explotadas.

En consecuencia, la transparencia algorítmica constituye tanto el punto de partida conceptual como el marco de referencia práctico. Es la brújula que orienta la discusión y al mismo tiempo el terreno sobre el que se despliegan las tensiones: ¿cómo lograr explicabilidad y apertura sin sacrificar la seguridad digital de los sistemas? ¿Cómo democratizar la comprensión de los algoritmos sin debilitar la capacidad del Estado de proteger los datos y procesos que gestiona? Estas preguntas derivadas de la definición son las que guiarán los apartados siguientes, donde se analizarán las facetas activa y pasiva de la transparencia, el papel de la seguridad digital como condición de confianza pública y el dilema del acceso al código fuente.

2. Facetas activa y pasiva de la transparencia algorítmica (art. 7 de la Directiva; jurisprudencia constitucional)

La Sentencia T-067 de 2025 estableció que la transparencia algorítmica posee dos dimensiones complementarias: una faceta activa y una faceta pasiva, ambas integradas en el núcleo esencial del derecho fundamental de acceso a la información pública. En palabras de la Corte, «la transparencia significativa tiene dos elementos implícitos que son la accesibilidad y la explicabilidad», lo cual supone que las autoridades no solo deben publicar información de manera proactiva, sino también garantizar que esta sea comprensible y verificable por la ciudadanía.

De manera concordante, el artículo 7 de la Directiva Conjunta 007 de 2025 dispone expresamente que «la transparencia algorítmica en su faceta activa y pasiva es un componente del núcleo esencial del derecho de acceso a la información pública», y ordena a los sujetos obligados que «garanticen que la ciudadanía pueda acceder, de manera clara, oportuna y comprensible, a la información relacionada con los sistemas algorítmicos utilizados en procesos administrativos o de toma de decisiones que puedan afectar positiva o negativamente sus derechos».

La faceta activa implica el deber de divulgar, de forma proactiva, clara y accesible, información sobre los sistemas de decisión automatizada y de inteligencia artificial, sin necesidad de que exista una solicitud previa. Este deber se concreta en los artículos 8 y 9 de la Directiva, que establecen los «estándares de información sobre los sistemas algorítmicos utilizados por el Estado» y los contenidos mínimos de divulgación, tales como el nombre del sistema, su objetivo, la función institucional que cumple, el tipo de datos que emplea y los canales de contacto. Así, por ejemplo, en el sector salud, un hospital público que utilice un sistema de IA para priorizar pacientes en listas de espera debe publicar de manera proactiva la finalidad del algoritmo, los tipos de datos que procesa (biométricos, clínicos, georreferenciación) y los criterios generales de priorización. De esa forma, los ciudadanos pueden conocer cómo se decide el orden de atención sin tener que interponer una solicitud, materializando la transparencia activa.

Por su parte, la faceta pasiva garantiza que toda persona pueda solicitar y recibir información clara, veraz, completa y comprensible sobre los sistemas algorítmicos utilizados por el Estado, conforme a lo dispuesto en el artículo 16 de la Directiva, y que las entidades respondan oportunamente dentro de los plazos de la Ley 1755 de 2015. Este componente se refuerza con el principio de máxima divulgación, que impone la obligación de facilitar el acceso salvo en los casos excepcionales expresamente previstos en la Constitución o la ley. En el ámbito de la justicia, por ejemplo, si un ciudadano sospecha que un sistema automatizado utilizado por una entidad judicial influye en la asignación de procesos o en la gestión de despachos, puede ejercer su derecho a solicitar información sobre ese sistema: cómo se diseñó, qué variables considera y qué mecanismos de supervisión humana incorpora. La entidad, a su vez, tiene la

obligación de responder en términos comprensibles y en lenguaje claro, garantizando así la transparencia pasiva.

De igual manera, en el sector de los subsidios sociales, si una persona considera que un sistema de IA ha denegado injustamente su acceso a un beneficio estatal puede presentar una solicitud de información para conocer los criterios y parámetros utilizados en la decisión. La administración debe responder con información verificable —por ejemplo, que el sistema pondera ingreso, composición familiar y residencia—, asegurando que el ciudadano pueda cuestionar o impugnar la decisión con base en información suficiente.

Finalmente, la doble naturaleza de la transparencia cobra especial relevancia en el campo de la seguridad pública, donde las autoridades pueden utilizar sistemas predictivos o de vigilancia inteligente. En estos casos, la transparencia activa exige publicar las finalidades, límites y protocolos de uso de dichos sistemas; mientras que la transparencia pasiva protege el derecho de las personas o comunidades vigiladas a obtener información sobre los criterios generales de operación, siempre que su divulgación no comprometa la seguridad nacional ni la integridad de las operaciones.

En términos más sencillos, la faceta activa de la transparencia equivale a «mostrar sin que se lo pidan», mientras que la faceta pasiva representa el derecho ciudadano a «preguntar y recibir respuestas completas». La primera evita la opacidad estructural al obligar a las entidades a publicar información sobre sus algoritmos e inteligencias artificiales antes de que alguien sospeche o reclame; la segunda garantiza que cualquier persona, sin importar su nivel técnico, pueda solicitar explicaciones o detalles adicionales cuando un sistema automatizado le afecta directamente.

Ambas facetas son necesarias para equilibrar la relación entre el Estado y la ciudadanía en entornos gobernados por algoritmos. Si la administración solo actuara de forma reactiva, la transparencia sería meramente defensiva; y si solo divulgara información sin atender las solicitudes concretas de los ciudadanos, perdería su sentido participativo y democrático. Por eso la Directiva y la jurisprudencia constitucional construyen un modelo de transparencia dual, que combina la publicación rutinaria de información con la posibilidad de escrutinio y diálogo ciudadano. En efecto, clarificar el alcance de la transparencia algorítmica (objetivo a) exige entender que ambas facetas son complementarias: la activa materializa la rendición de cuentas institucional, y la pasiva habilita el control ciudadano y judicial de los sistemas de IA y SDA. Asimismo, el análisis de esta estructura anticipa los dilemas que se examinarán en los apartados siguientes, especialmente en relación con la seguridad digital: cuanta más información se publica (faceta activa), mayor es el riesgo potencial de exposición; y cuanto más se restringe la información (limitación de la faceta pasiva), menor es la legitimidad del uso de algoritmos en la función pública.

3. Seguridad digital como condición de confianza pública (referencia a la Ley 1581/2012 y estándares internacionales como UNESCO)

La seguridad digital no solo constituye un requisito técnico indispensable para la operación confiable de los sistemas algorítmicos y de inteligencia artificial (IA), sino que, de acuerdo con el Decreto 767 de 2022, expedido el 16 de mayo y mediante el cual «se establecen los lineamientos generales de la Política de Gobierno Digital y se subroga el Capítulo 1 del Título 9 de la Parte 2 del Libro 2 del Decreto 1078 de 2015», se erige como uno de los ejes estructurales del modelo de gobernanza digital del Estado colombiano. Este decreto consolida la visión del Estado sobre el uso de las Tecnologías de la Información y las Comunicaciones (TIC) no como un fin en sí mismo, sino como un medio para garantizar derechos, generar valor público y fortalecer la confianza ciudadana en la administración digital.

El artículo 2.2.9.1.1.1 define la Política de Gobierno Digital como una estrategia orientada al «uso y aprovechamiento de las TIC con el objetivo de impactar positivamente la calidad de vida de los ciudadanos, promoviendo la generación de valor público a través de la transformación digital del Estado, de manera proactiva, confiable, articulada y colaborativa, y permitir el ejercicio de los derechos de los usuarios del ciberespacio». Esta formulación, de naturaleza reglamentaria, convierte la seguridad digital en una condición de validez para el ejercicio de los derechos digitales: solo un entorno confiable puede garantizar el libre y seguro uso de los servicios públicos digitales, el acceso a la información y la interacción en el ciberespacio.

A su vez, el artículo 2.2.9.1.1.3 establece los principios rectores de la Política, entre los cuales la Confianza ocupa un lugar central. Este principio impone a los sujetos obligados la responsabilidad de cumplir con todas las disposiciones que aseguren la garantía de la seguridad digital, la protección de datos y la transparencia pública, de manera que la transformación digital se oriente a satisfacer expectativas ciudadanas sin sacrificar la integridad institucional. Junto con la confianza, la norma introduce el principio de Legalidad tecnológica, que exige que el uso de las TIC, los sistemas algorítmicos y las soluciones de IA se ajusten estrictamente a la Constitución, la ley y los reglamentos, asegurando que la tecnología opere dentro del Estado de Derecho. De igual modo, el principio de Resiliencia tecnológica obliga a prevenir y gestionar los riesgos que puedan afectar la continuidad y la disponibilidad de los servicios digitales, estableciendo mecanismos de prevención, recuperación y continuidad operativa ante interrupciones o incidentes. Finalmente, los principios de Proactividad y Prospectiva tecnológica refuerzan el deber de anticipar amenazas, identificar tecnologías emergentes seguras y desarrollar capacidades institucionales para sostener una administración digital robusta y adaptable.

El mismo Decreto, en su artículo 2.2.9.1.2.1, al definir la estructura de la Política de Gobierno Digital, consagra la seguridad y privacidad de la información como uno de los habilitadores obligatorios que deben implementar las

entidades públicas. Este habilitador busca que las instituciones desarrollen capacidades orientadas a preservar la confidencialidad, integridad, disponibilidad y privacidad de los datos en todos los procesos, servicios, sistemas de información e infraestructuras bajo su control. En el diseño del sistema, la seguridad deja de ser una función auxiliar para convertirse en una obligación transversal y permanente, vinculada a la buena administración (artículo 209 de la Constitución Política) y a los principios de publicidad y transparencia previstos en la Ley 1712 de 2014. De esta manera, el Decreto 767 de 2022 no solo consolida un marco operativo de seguridad, sino que lo juridifica, al conferirle un carácter vinculante y sistemático dentro del ordenamiento administrativo colombiano.

La seguridad digital, en este contexto, deja de ser un aspecto puramente técnico o dependiente de la voluntad institucional, para convertirse en un principio estructural del derecho administrativo digital. Esto significa que la obligación de garantizar la seguridad no deriva únicamente de consideraciones tecnológicas, sino del propio mandato constitucional de proteger derechos fundamentales como la intimidad, el debido proceso y la protección de datos personales. Así, la seguridad se erige como una condición jurídica para el ejercicio legítimo y seguro de la transparencia pública. La transparencia sin medidas de seguridad puede degenerar en vulnerabilidad, mientras que la seguridad sin transparencia puede convertirse en opacidad; por ello, el modelo de gobernanza digital colombiano busca un equilibrio regulado entre ambos valores.

Este marco reglamentario se complementa con otras normas de rango legal y técnico. La Ley 1581 de 2012, sobre protección de datos personales, impone a los responsables del tratamiento el deber de implementar «medidas técnicas, humanas y administrativas necesarias para otorgar seguridad a los registros, evitando su adulteración, pérdida, consulta, uso o acceso no autorizado o fraudulento». Por su parte, la Directiva Conjunta 007 de 2025, en su artículo 5.g, reitera que todo desarrollo, adquisición, uso o implementación de sistemas algorítmicos deberá «garantizar el cumplimiento de la normativa vigente en materia de protección de datos personales». Finalmente, los estándares internacionales —en especial, la Recomendación sobre la Ética de la Inteligencia Artificial de la UNESCO (2021) y los Principios de la OCDE sobre Inteligencia Artificial (2024)— coinciden en que la seguridad, robustez y trazabilidad de los sistemas constituyen la base de su legitimidad democrática y de su confiabilidad social.

El resultado de esta convergencia normativa es un bloque reglamentario integrado que articula la seguridad digital, la protección de datos personales y la transparencia algorítmica bajo un mismo principio de confianza pública digital. En este bloque, la seguridad digital funciona como una condición habilitante de la transparencia significativa: garantiza que la información divulgada pueda ser comprendida y auditada sin poner en riesgo los bienes jurídicos protegidos. De este modo, el Estado puede promover la apertura informativa y el control ciudadano sin comprometer la integridad de sus sistemas ni la confidencialidad de los datos que procesan.

Esta interacción se observa claramente en los entornos de aplicación. En el ámbito financiero y fiscal, los algoritmos de detección de fraude o de evasión

tributaria deben ser explicables y auditables, pero la revelación de sus parámetros o ponderaciones específicas podría facilitar su manipulación; por ello, se divulga la lógica general del modelo y los criterios de riesgo, preservando su arquitectura técnica bajo medidas de seguridad. En el sector justicia, los sistemas de priorización o reparto judicial requieren trazabilidad y registro de auditorías, pero las configuraciones técnicas de los sistemas deben mantenerse protegidas para evitar alteraciones que comprometan el debido proceso. En el campo sanitario, los modelos predictivos de inteligencia artificial permiten informar a los ciudadanos sobre sus criterios de análisis (variables de riesgo, umbrales de priorización, niveles de confiabilidad), garantizando al mismo tiempo la confidencialidad médica y la integridad de las bases de datos clínicas. Finalmente, en los sistemas de gestión territorial y seguridad ciudadana, la divulgación de la finalidad, fuentes de datos y mecanismos de supervisión de los sistemas de videovigilancia o analítica predictiva materializa la transparencia, mientras que la reserva de protocolos operativos preserva la seguridad nacional y la protección de las personas.

Por tanto, la seguridad digital se configura como un principio operativo y jurídico de compatibilización, que permite armonizar el derecho de acceso a la información con los deberes de prevención de riesgo y continuidad de los servicios digitales. En el marco del presente estudio, representa el punto de convergencia entre los objetivos (b) y (c): de un lado, condiciona la decisión entre publicar el código fuente o suministrar una explicación significativa, determinando el alcance de las reservas mediante el test de daño; y de otro, viabiliza la integración de las obligaciones de transparencia con las salvaguardas de seguridad digital y protección de datos, asegurando que la apertura informativa se realice dentro de un régimen de legalidad tecnológica, control y responsabilidad.

En síntesis, dentro del ordenamiento jurídico colombiano, la seguridad digital es una obligación de naturaleza reglamentaria y un principio de orden público derivado del Decreto 767 de 2022, que vincula a todas las entidades de la administración pública y a los particulares que ejercen funciones administrativas. Su cumplimiento no solo garantiza la estabilidad técnica de los sistemas de decisión automatizada y de inteligencia artificial, sino también la legitimidad constitucional del derecho de acceso a la información pública en la era digital. Constituye, por tanto, la columna vertebral de una transparencia algorítmica segura, responsable y sostenible.

4. El dilema del código fuente: entre el derecho de acceso (art. 74 CP + Ley 1712/2014) y la protección frente a riesgos (art. 17 y 19 Directiva, test de daño)

Uno de los puntos más sensibles de la transparencia algorítmica se encuentra en la tensión entre el derecho fundamental de acceso a la información pública —consagrado en el artículo 74 de la Constitución Política y desarrollado por la Ley 1712 de 2014— y la necesidad de preservar la seguridad digital, la integridad del sistema y los derechos de propiedad intelectual que pueden con-

currir sobre el código fuente de los sistemas algorítmicos e inteligencia artificial utilizados por el Estado. Este equilibrio constituye el núcleo del denominado «dilema del código fuente», que no solo plantea una cuestión técnica, sino una verdadera disyuntiva jurídica entre transparencia y protección.

La Sentencia T-067 de 2025 abordó este dilema de manera directa. En sus considerandos 79 y 104, la Corte Constitucional estableció que los sujetos obligados deberán «evaluar, caso a caso, con herramientas como el test de daño, si para conseguir una mayor transparencia algorítmica significativa optan por publicar el código fuente o, en su lugar, proporcionan una explicación significativa… en lenguaje claro sobre cómo el sistema toma una decisión». De esta forma, el tribunal reconoció que la publicación del código no puede concebirse como una obligación automática ni como un secreto absoluto: debe resolverse mediante una ponderación entre el interés público en la información y el riesgo cierto y específico que su divulgación entrañe.

El desarrollo reglamentario de este principio se encuentra en los artículos 17, 18 y 19 de la Directiva Conjunta 007 de 2025, los cuales delimitan las condiciones en que las entidades públicas pueden entregar o negar el código fuente de los sistemas que emplean. Según el artículo 17, ante solicitudes expresas relacionadas con el código, el sujeto obligado debe determinar si el propósito de transparencia puede alcanzarse mediante la entrega de una «explicación significativa», suficiente para que las personas entiendan la lógica, criterios y efectos del sistema. El artículo 18 dispone que, cuando excepcionalmente se decida entregar el código, debe suscribirse un acta de entrega que establezca compromisos claros de uso responsable, incluyendo la prohibición de su utilización con fines comerciales o maliciosos, la obligación de no reproducirlo ni distribuirlo, y el deber de preservar la integridad del sistema. Finalmente, el artículo 19 regula las excepciones y obliga a realizar un test de daño, a través del cual la entidad debe demostrar que conceder el acceso produciría un daño presente o futuro, probable y específico sobre un bien constitucionalmente protegido, y que ese daño supera el interés público del acceso.

Desde una perspectiva técnica y jurídica integrada, el dilema del código fuente enfrenta tres planos normativos que deben armonizarse. En primer lugar, el plano constitucional, que consagra el principio de publicidad como regla general y las excepciones como interpretaciones estrictas sujetas a motivación suficiente. En segundo lugar, el plano legal, en el que la Ley 1712 de 2014 impone a las entidades la carga de justificar toda reserva, y el artículo 28 de dicha ley exige demostrar que la negación del acceso busca proteger un fin legítimo y proporcional. Y, en tercer lugar, el plano técnico-reglamentario, que reconoce que el código fuente constituye, en muchos casos, un activo estratégico de ciberseguridad cuya divulgación indiscriminada puede generar vulnerabilidades, violaciones de derechos de autor o riesgos de manipulación de datos.

En la práctica administrativa, las entidades enfrentan escenarios complejos. Cuando se trata de sistemas de inteligencia artificial de alto impacto, como los que gestionan beneficios sociales, decisiones judiciales o diagnósticos médicos, la entrega del código fuente podría exponer modelos entrenados con grandes

volúmenes de datos personales o revelar arquitecturas cuya explotación pondría en riesgo la infraestructura pública. En estos casos, la opción más razonable, y jurídicamente adecuada, consiste en ofrecer una explicación significativa que describa la lógica decisional, los criterios ponderados y las medidas de mitigación de sesgos, acompañada de información verificable sobre el proceso de entrenamiento y evaluación del sistema. Por el contrario, cuando el algoritmo tiene un impacto limitado o se encuentra basado en reglas explícitas, como los sistemas de priorización de trámites o la automatización de procedimientos rutinarios, el acceso al código puede concederse de manera controlada, firmando el acta de entrega prevista en el artículo 18 de la Directiva y aplicando cláusulas de confidencialidad.

Este equilibrio dinámico exige que las autoridades adopten una metodología de ponderación documentada, en la que el test de daño opere como instrumento de garantía constitucional. El test obliga a demostrar que la restricción al acceso: (i) persigue un objetivo legítimo, (ii) se encuentra expresamente prevista en la ley, (iii) previene un daño probable y específico, y (iv) ese daño es mayor que el beneficio público de divulgar la información. La aplicación sistemática de este test no solo fortalece la transparencia significativa, sino que permite evidenciar el cumplimiento del principio de legalidad tecnológica previsto en el Decreto 767 de 2022, asegurando que toda decisión sobre acceso a información tecnológica se adopte dentro de los parámetros del Estado de Derecho digital.

El dilema del código fuente también pone de relieve la necesidad de articular los principios de seguridad digital y rendición de cuentas. La transparencia no se materializa mediante la exposición absoluta del software, sino mediante la generación de confianza institucional basada en mecanismos de auditabilidad y explicabilidad. Por ello, la Directiva 007/2025 adopta el concepto de «transparencia significativa», que privilegia la comprensión sobre la simple apertura del código. Se trata de una evolución conceptual que traslada el foco del objeto técnico (el código) al resultado normativo (el entendimiento ciudadano de la decisión automatizada).

Desde el punto de vista del objeto de estudio de este capítulo, el dilema del código fuente sintetiza la tensión central entre transparencia y seguridad digital. Si el Estado opta por divulgar sin control, expone la infraestructura digital y reduce la resiliencia del sistema; si reserva toda la información, vulnera el derecho de los ciudadanos a conocer y cuestionar las decisiones algorítmicas que los afectan. La clave está en el diseño de mecanismos de compatibilización ex ante y ex post, tales como los análisis de impacto algorítmico, las auditorías técnicas, la trazabilidad documental y la entrega de explicaciones significativas, que permitan cumplir el mandato constitucional de publicidad sin renunciar a la protección de los activos digitales.

En síntesis, el código fuente constituye el punto de encuentro entre el derecho de acceso a la información pública y el deber de garantizar la seguridad digital del Estado. Su tratamiento debe regirse por criterios de proporcionalidad, razonabilidad y finalidad legítima, siguiendo la metodología establecida por la

Sentencia T-067 de 2025 y la Directiva Conjunta 007 de 2025. En el marco del Decreto 767 de 2022, este equilibrio adquiere una dimensión estructural: la seguridad digital, al ser parte de la Política de Gobierno Digital, no limita la transparencia, sino que la hace jurídicamente posible. Solo a través de este equilibrio normativo y técnico puede consolidarse una transparencia algorítmica responsable, capaz de conjugar apertura informativa, protección de derechos y preservación de la infraestructura digital del Estado colombiano.

III. MECANISMOS DE COMPATIBILIZACIÓN

El tránsito desde el plano de los principios hacia el terreno de la acción administrativa exige identificar los instrumentos concretos mediante los cuales puede alcanzarse una compatibilización efectiva entre la transparencia algorítmica y la seguridad digital. Dicha compatibilización no debe entenderse como una concesión recíproca de derechos ni como un equilibrio abstracto entre valores en tensión, sino como la configuración de un régimen técnico-jurídico operativo, sustentado en mecanismos verificables de prevención, gestión y control. Bajo este enfoque, los mecanismos de compatibilización son herramientas institucionales y normativas que permiten al Estado cumplir simultáneamente con su obligación de transparencia y su deber de protección, garantizando que la apertura informativa se ejerza dentro de los límites del debido proceso tecnológico, la protección de datos personales y la seguridad de la infraestructura digital.

Esta compatibilización se construye sobre dos dimensiones complementarias: una dimensión ex ante, de carácter preventivo, orientada a asegurar que la transparencia y la seguridad se integren desde la concepción y el diseño de los sistemas de decisión automatizada e inteligencia artificial; y una dimensión ex post, de naturaleza correctiva y de rendición de cuentas, destinada a garantizar la trazabilidad, la auditabilidad y la revisión permanente de los sistemas ya desplegados.

En el ordenamiento colombiano, esta doble dimensión se encuentra positivizada en la Directiva Conjunta 007 de 2025, particularmente en sus artículos 5, 12, 13, 14, 17 y 19, complementados por la Ley 1581 de 2012 (protección de datos personales), la Ley 1712 de 2014 (transparencia y acceso a la información pública), la Ley 1755 de 2015 (derecho de petición) y el Decreto 767 de 2022 (Política de Gobierno Digital). En conjunto, estas disposiciones configuran un entramado normativo coherente que permite derivar cinco mecanismos de compatibilización esenciales, que serán analizados en los apartados siguientes.

El primero de ellos es la explicabilidad y el lenguaje claro, previstos en los artículos 5.c y 5.d de la Directiva 007 de 2025 y vinculados al artículo 22 del Reglamento General de Protección de Datos (RGPD) de la Unión Europea. Este mecanismo traduce el principio de transparencia en un deber de inteligibilidad: los sistemas deben ser capaces de ofrecer explicaciones verificables y comprensibles sobre la lógica y los criterios que sustentan sus resultados, sustituyendo la opacidad técnica por explicaciones significativas que permitan el control social, judicial y democrático de las decisiones automatizadas.

El segundo mecanismo corresponde al análisis de impacto y las auditorías algorítmicas, consagrado en el artículo 14 de la Directiva 007 de 2025. Este instrumento materializa la prevención ex ante, al exigir que las entidades públicas identifiquen, valoren y mitiguen los riesgos éticos, jurídicos y técnicos antes de la implementación o actualización de un sistema de inteligencia artificial. El análisis de impacto algorítmico se alinea con las mejores prácticas internacionales —como las establecidas por Canadá, la Unión Europea y la OCDE—, que lo reconocen como una garantía estructural de rendición de cuentas y debida diligencia tecnológica.

El tercer mecanismo lo constituyen los canales de revisión y objeción frente a decisiones total o parcialmente automatizadas, regulados en el artículo 12 de la Directiva 007 de 2025 y en la Ley 1755 de 2015. Este instrumento garantiza el ejercicio del derecho fundamental de petición y la intervención humana significativa, permitiendo que las personas afectadas por un sistema automatizado puedan solicitar su revisión, impugnar sus resultados o requerir explicaciones adicionales. Con ello, la transparencia adquiere una dimensión procedimental y correctiva, que fortalece el derecho al debido proceso y la tutela judicial efectiva.

El cuarto mecanismo está constituido por la protección de datos personales y la ciberseguridad por diseño, integrados en la Ley 1581 de 2012, el Decreto 1377 de 2013 y el artículo 5.g de la Directiva 007 de 2025. Este pilar vincula la transparencia con la seguridad digital mediante el principio de seguridad y privacidad desde la concepción, garantizando la trazabilidad, integridad y resiliencia de los sistemas frente a vulnerabilidades, accesos no autorizados y riesgos de manipulación. En armonía con el Decreto 767 de 2022, este enfoque consolida la idea de que la apertura informativa y la protección tecnológica no son categorías excluyentes, sino dimensiones complementarias de la legalidad tecnológica.

Finalmente, el quinto mecanismo de compatibilización está representado por el test de daño, previsto en los artículos 17, 18 y 19 de la Directiva 007 de 2025 y en los artículos 18 y 19 de la Ley 1712 de 2014. Este procedimiento reglado actúa como instrumento de ponderación ex post, que permite determinar si la divulgación de determinada información relativa a sistemas algorítmicos puede generar un daño cierto, probable y específico a bienes constitucionalmente protegidos, como la seguridad digital, la protección de datos personales o los derechos de propiedad intelectual. Su aplicación documentada y motivada asegura que toda restricción al acceso se funde en criterios objetivos y verificables, y no en consideraciones preventivas o genéricas.

En conjunto, estos cinco mecanismos conforman la arquitectura operativa de compatibilización entre transparencia y seguridad digital en el Estado colombiano. Cada uno representa una instancia normativa y procedimental de equilibrio entre los principios de publicidad, rendición de cuentas, legalidad tecnológica y protección de derechos fundamentales, los cuales, articulados en conjunto, sustentan la gobernanza algorítmica responsable y la confianza pública en los sistemas de inteligencia artificial implementados por las autoridades públicas.

1. Explicabilidad y lenguaje claro

El principio de explicabilidad se erige como uno de los pilares de la transparencia algorítmica contemporánea. En el contexto colombiano, su fundamento jurídico se encuentra en los artículos 5.c y 5.d de la Directiva Conjunta 007 de 2025, los cuales establecen que los sistemas algorítmicos y de inteligencia artificial utilizados por los sujetos obligados deberán «ser explicables en su diseño, implementación y uso», y que las entidades deben «comunicar información sobre su funcionamiento en un lenguaje claro, preciso y comprensible para cualquier persona». A su vez, el artículo 17 de la misma Directiva refuerza esta exigencia al disponer que, cuando la publicación del código fuente no sea viable o razonable, las entidades deberán proporcionar «explicaciones significativas en lenguaje claro» sobre el modo en que los sistemas toman decisiones o producen resultados.

Este mandato se alinea con el artículo 22 del Reglamento General de Protección de Datos (RGPD) de la Unión Europea, que reconoce el derecho de toda persona a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, sin que medie una «intervención humana significativa» y sin que el interesado reciba «información pertinente sobre la lógica aplicada, así como la importancia y las consecuencias previstas de dicho tratamiento». La incorporación de este estándar internacional en la Directiva 007/2025 muestra la convergencia normativa entre los sistemas europeo y colombiano hacia un modelo de transparencia significativa y comprensible, en el cual el ciudadano debe poder entender cómo una decisión automatizada afecta sus derechos.

La explicabilidad, en este sentido, no se reduce a una exigencia técnica de documentación del algoritmo; es una obligación jurídica sustantiva que impone a las entidades públicas el deber de garantizar que las decisiones generadas o asistidas por sistemas algorítmicos puedan ser racionalmente comprendidas, justificadas y auditadas. En otras palabras, el principio de explicabilidad traduce el mandato constitucional de publicidad (artículo 209 de la Constitución) al contexto de la inteligencia artificial, asegurando que los procesos automatizados no introduzcan zonas de opacidad o discrecionalidad tecnológica en el ejercicio de la función administrativa.

La exigencia de lenguaje claro, por su parte, complementa la explicabilidad al situar la transparencia en el terreno de la comunicación efectiva. El lenguaje claro constituye un estándar de accesibilidad cognitiva: la información no solo debe estar disponible, sino formulada en términos que permitan a la ciudadanía interpretarla sin mediación técnica especializada. De este modo, la Directiva 007/2025 no se limita a imponer transparencia formal, sino que busca inteligibilidad material, en consonancia con el principio de máxima divulgación de la Ley 1712 de 2014 y con los postulados de la Política de Gobierno Digital adoptada por el Decreto 767 de 2022, que exige que la transformación digital del Estado sea «proactiva, confiable y articulada» con los derechos de los usuarios del ciberespacio.

Desde una perspectiva sistémica, la explicabilidad y el lenguaje claro conforman un mecanismo dual de compatibilización entre transparencia y seguridad digital. Su función es garantizar que las explicaciones sean suficientemente ricas en contenido para permitir el control democrático, pero suficientemente abstractas para evitar la exposición de información sensible o de componentes críticos de ciberseguridad. El artículo 17 de la Directiva materializa este equilibrio al permitir sustituir la entrega del código fuente por explicaciones significativas, siempre que estas sean verificables, actualizadas y comprensibles. En consecuencia, la transparencia se logra no por la apertura indiscriminada del software, sino por la capacidad institucional de traducir el razonamiento algorítmico en información jurídicamente relevante y socialmente inteligible.

Un ejemplo paradigmático de la aplicación de este principio se encuentra en el proceso de focalización y selección de beneficiarios de programas sociales. En estos casos, las entidades públicas utilizan sistemas algorítmicos para priorizar hogares con base en variables socioeconómicas, demográficas o territoriales. La explicabilidad significativa exige que la entidad publique, en lenguaje claro, los criterios generales del modelo —por ejemplo, que pondera ingresos, composición familiar, condición de vulnerabilidad y localización geográfica—, así como el modo en que dichas variables interactúan para producir el resultado final. De este modo, el ciudadano puede comprender por qué su hogar fue o no incluido, identificar posibles errores en los datos y ejercer su derecho de objeción sin necesidad de acceder al código fuente. La transparencia, en este contexto, no revela el algoritmo en sí, pero sí garantiza la rendición de cuentas sobre la lógica de priorización que afecta directamente el acceso a derechos sociales.

Este tipo de explicaciones fortalece la confianza pública y cumple simultáneamente con las exigencias de seguridad digital establecidas en el Decreto 767 de 2022, pues permite divulgar la lógica general sin exponer información que pueda ser manipulada o utilizada indebidamente para alterar los resultados. Así, la entidad preserva la integridad del sistema y la confidencialidad de los datos personales, mientras facilita la comprensión ciudadana del proceso decisorio. Estos mecanismos de explicabilidad y lenguaje claro constituyen, en consecuencia, una garantía esencial de debido proceso algorítmico. La persona afectada por una decisión automatizada debe poder entender las razones que la motivaron y conocer las vías para solicitar revisión o corrección. La explicación cumple entonces una triple función: jurídica, al permitir el ejercicio del derecho de defensa y de control judicial; ética, al reforzar la rendición de cuentas; y técnica, al fomentar la documentación y validación responsable de los modelos algorítmicos.

La explicabilidad y el lenguaje claro representan la concreción normativa del principio de transparencia significativa. Constituyen la capacidad institucional de comunicar el «cómo» y el «por qué» de las decisiones automatizadas en términos accesibles, verificables y seguros. Su aplicación adecuada no solo incrementa la confianza y legitimidad de las políticas públicas basadas en inteligencia artificial, sino que consolida la sujeción de la tecnología al derecho y el

fortalecimiento del control democrático sobre el Estado digital. La explicabilidad se exige en general, pero el estándar de publicación adicional es reforzado solo para los supuestos del art. 13.

2. Análisis de impacto y auditorías

El análisis de impacto algorítmico previsto en el artículo 14 de la Directiva Conjunta 007 de 2025 constituye una obligación específica aplicable únicamente a los sistemas algorítmicos comprendidos en el artículo 13, es decir, aquellos que el legislador reglamentario ha calificado como «sujetos a mayores exigencias de divulgación». Esta restricción delimita con precisión el ámbito de aplicación del análisis, que no se extiende a todos los sistemas utilizados por el Estado, sino exclusivamente a los que generan efectos jurídicos o materiales de especial intensidad sobre los derechos de las personas.

El artículo 13 establece siete supuestos que determinan la procedencia del deber reforzado de transparencia y, por consiguiente, del análisis de impacto. Dichos supuestos incluyen los sistemas utilizados para evaluar, priorizar o determinar la asignación o denegación de beneficios, subsidios, apoyos, empleos o servicios; los que predicen riesgos asociados a conductas como criminalidad, reincidencia, fraude o deserción; los que generan perfiles individuales o colectivos de personas; aquellos que realizan inferencia de emociones; los que intervienen en servicios dirigidos a niñas, niños y adolescentes, minorías étnicas, personas con discapacidad, migrantes u otras poblaciones vulnerables; los que predicen o diagnostican enfermedades; y los que acompañan personas con enfermedades mentales.

En cualquiera de estos casos, los sujetos obligados del artículo 5 de la Ley 1712 de 2014 deben realizar, antes de la adquisición, desarrollo, implementación o uso del sistema, un análisis de impacto algorítmico conforme al artículo 14 de la Directiva. Este análisis tiene por objeto identificar, evaluar y mitigar los riesgos asociados al funcionamiento, uso y efectos del sistema, y documentar las medidas adoptadas para garantizar la transparencia, la seguridad digital y la protección de los derechos de las personas potencialmente afectadas. El contenido mínimo del análisis debe incluir una descripción del objetivo del sistema y del proceso que automatiza o apoya; la identificación de las fuentes y calidad de los datos; la detección de posibles sesgos o efectos discriminatorios; las medidas de mitigación aplicadas; los procedimientos de revisión y actualización del modelo; y la designación de las áreas responsables de su ejecución y supervisión. Esta información debe hacerse pública de forma accesible y comprensible, conforme al mandato del artículo 13, de manera que permita un mayor grado de comprensión y escrutinio ciudadano sobre el funcionamiento y los efectos del sistema.

El análisis de impacto tiene así una función preventiva y de garantía constitucional: asegura que los sistemas algorítmicos de alto impacto solo se implementen después de haber evaluado su conformidad con los principios de igual-

dad, proporcionalidad, seguridad digital y rendición de cuentas. Su realización constituye evidencia de diligencia administrativa reforzada, en cumplimiento del principio de legalidad tecnológica previsto en el Decreto 767 de 2022, según el cual las tecnologías utilizadas en el sector público deben operar dentro del marco constitucional y reglamentario vigente. El artículo 20 de la Directiva complementa esta obligación con el deber de realizar auditorías periódicas y reportes de actualización. Las entidades deben mantener un registro de desempeño, documentar los incidentes detectados y reportar las medidas correctivas adoptadas. Estas auditorías ex post garantizan la trazabilidad del sistema y la actualización permanente de los controles, evitando que la transparencia se agote en la fase inicial del ciclo de vida algorítmico.

Por ejemplo, un sistema de priorización de subsidios de vivienda, comprendido en el numeral (i) del artículo 13, debe someterse a un análisis de impacto previo a su despliegue para identificar sesgos en los datos de ingresos, composición familiar o localización territorial. El informe deberá describir las medidas correctivas adoptadas —como ponderaciones compensatorias o revisión humana de casos límite— y prever auditorías periódicas para verificar su eficacia. De igual modo, un sistema predictivo de reincidencia penal, incluido en el numeral (ii) del mismo artículo, debe evaluar la proporcionalidad y confiabilidad de los datos históricos utilizados, garantizar la intervención humana significativa en la decisión final y establecer criterios de revisión técnica periódica para detectar desviaciones o sesgos emergentes.

El análisis de impacto algorítmico y las auditorías conforman, por tanto, un binomio de control que traduce los principios de proporcionalidad, legalidad tecnológica y rendición de cuentas. La transparencia no se limita a la divulgación posterior de información, sino que implica la capacidad institucional de anticipar riesgos, documentar las medidas de mitigación y verificar su efectividad en el tiempo.

Sin embargo, este enfoque plantea dos problemas jurídicos y administrativos relevantes que deben ser reconocidos.

El primer problema es la capacidad institucional limitada de muchas entidades públicas para realizar análisis de impacto con la profundidad técnica y metodológica exigida por la Directiva. La evaluación de sesgos, la verificación de calidad de datos o la identificación de riesgos algorítmicos requieren competencias interdisciplinarias (jurídicas, estadísticas, informáticas y éticas) que no siempre están disponibles en los niveles nacional o territorial. Esta carencia puede conducir a evaluaciones incompletas o meramente formales, reduciendo el análisis a un requisito documental y debilitando su valor preventivo.

El segundo problema radica en la ausencia de un marco metodológico uniforme que estandarice los procedimientos, formatos y criterios de evaluación. Aunque la Directiva impone la obligación del análisis de impacto, no establece una guía técnica vinculante que defina indicadores, métricas de riesgo o niveles de severidad. Esta ambigüedad normativa genera heterogeneidad en los resultados y dificulta la comparación o supervisión por parte de los órganos de control. En consecuencia, la eficacia del sistema depende de la pronta elabora-

ción —por parte del Ministerio Público y la Agencia Nacional Digital— de lineamientos metodológicos complementarios que aseguren la coherencia y verificabilidad de los análisis realizados.

En suma, el análisis de impacto algorítmico previsto para los sistemas descritos en el artículo 13 de la Directiva 007/2025, junto con las auditorías periódicas del artículo 20, constituyen un mecanismo normativo robusto para equilibrar transparencia y seguridad digital. No obstante, su efectividad depende de la capacidad institucional para ejecutarlos con rigor técnico y de la adopción de metodologías estandarizadas que eviten la dispersión interpretativa. Superar estos dos retos es condición indispensable para que el Estado colombiano avance hacia una transparencia algorítmica verificable, uniforme y jurídicamente exigible.

3. Canales de revisión y objeción

El fundamento normativo se encuentra en el Artículo 12 de la Directiva 007/2025: «Canal de contacto para objeciones y solicitudes de revisión». El precepto ordena que los sujetos obligados «mantengan habilitados canales de contacto para revisiones, solicitudes u objeciones frente a decisiones total o parcialmente automatizadas», precisando el área o funcionario responsable, los plazos para presentar y responder y asegurando «el respeto al debido proceso, la transparencia y la posibilidad de corregir o explicar las decisiones adoptadas por los sistemas algorítmicos». Esta obligación no está limitada a los sistemas del art. 13; aplica a todo sistema algorítmico implementado por los sujetos del art. 5 de la Ley 1712/2014. La transparencia pasiva operativa que habilita estas solicitudes se articula con el Artículo 16 («Solicitud de información sobre los sistemas algorítmicos utilizados por el Estado»), que impone responder en términos claros, veraces, completos y comprensibles, dentro de los plazos de la Ley 1755/2015.

El ámbito reforzado aparece cuando el sistema encaja en los supuestos del Artículo 13 («sistemas sujetos a mayores exigencias de divulgación»: asignación/denegación de beneficios o servicios; predicción de riesgos como criminalidad, reincidencia, fraude o deserción; perfilamiento; inferencia de emociones; intervención en servicios a NNA, minorías étnicas, personas con discapacidad, migrantes u otras poblaciones vulnerables; predicción/diagnóstico de enfermedades; acompañamiento en salud mental). En estos casos, además de los canales del art. 12 y el derecho de solicitud del art. 16, rigen la publicación adicional del propio art. 13 y la obligación de análisis de impacto ex ante del Artículo 14. Los canales del art. 12 permiten activar revisión/objeción y exigir respuesta motivada; pero no habilitan, por sí solos, entrega de código fuente. El acceso al código o su sustitución por «explicación significativa en lenguaje claro» se rige por los Artículos 17–19 (condiciones de entrega, test de daño, excepciones). En consecuencia, ante una objeción, la entidad debe ofrecer respuesta sustantiva y, si corresponde, explicación significativa; la entrega del código se decide con la metodología y límites de los arts. 17–19.

Operación mínima exigible. Para cumplir el art. 12, la entidad debe: (i) publicar y mantener canal visible y accesible (digital y presencial) para objeciones y revisiones; (ii) identificar el responsable institucional y plazos; (iii) registrar cada objeción (fecha, fundamento, decisión, corrección si procede) asegurando trazabilidad; (iv) emitir respuesta de fondo en lenguaje claro, dejando constancia de la intervención humana significativa cuando la decisión se modifique, confirme o anule; y (v) retroalimentar el sistema (ajustes de datos/modelo), enlazando con auditorías del Artículo 20 cuando se detecten errores sistemáticos. Si un hogar impugna una denegación automatizada de un subsidio (supuesto del Art. 13.i), la entidad debe recibir la objeción por el canal del Art. 12, verificar datos y reglas aplicadas, y responder con explicación significativa (Art. 17), indicando si corrige el resultado o lo confirma. El caso queda registrado para fines de auditoría (Art. 20). Si la objeción incluyera solicitud de código, la decisión se adopta conforme a Art. 17–19 (posible entrega condicionada o sustitución por explicación, previo test de daño del Art. 19).

4. Protección de datos y ciberseguridad por diseño

La protección de datos personales y la ciberseguridad por diseño constituyen los pilares sobre los que se sustenta la legitimidad jurídica de la transparencia algorítmica en la administración pública. Ambas obligaciones actúan como condiciones habilitantes del derecho de acceso a la información en entornos automatizados, al garantizar que los derechos fundamentales a la intimidad y a la autodeterminación informativa, así como la integridad de la infraestructura digital del Estado, no resulten comprometidos por la apertura informativa asociada al funcionamiento de los sistemas algorítmicos y de inteligencia artificial.

Desde el plano legal, la Ley 1581 de 2012 fija los principios del tratamiento en su artículo 4, dentro de los cuales el principio de seguridad impone que la información sujeta a tratamiento sea gestionada mediante medidas técnicas, humanas y administrativas adecuadas para preservar su confidencialidad, integridad y disponibilidad, previniendo accesos o usos no autorizados. Esta exigencia se concreta en los deberes del responsable y del encargado del tratamiento previstos en los artículos 17 y 18, que obligan a adoptar medidas eficaces de protección y a garantizar el cumplimiento de la normativa de datos personales en todas las fases del ciclo de tratamiento. La ley establece además un estándar reforzado de protección cuando se trata de datos sensibles o de sujetos de especial protección constitucional. El artículo 19 designa a la Superintendencia de Industria y Comercio (SIC) como autoridad de vigilancia encargada de garantizar el cumplimiento de los principios, derechos y garantías previstos en la norma. El Decreto 1377 de 2013, reglamentario de la Ley 1581, desarrolla el alcance del principio de seguridad y detalla las obligaciones de los responsables del tratamiento, incluyendo la adopción de políticas internas, la definición de controles de acceso, la gestión de incidentes y la documentación de las medidas de seguridad implementadas. Estas disposiciones convierten la

seguridad en un componente estructural del régimen de protección de datos, con carácter obligatorio y verificable por la autoridad competente.

La Directiva Conjunta 007 de 2025, en su artículo 5 literal g), refuerza este marco al establecer que los sistemas algorítmicos utilizados por los sujetos obligados deberán garantizar el cumplimiento de la normativa vigente en materia de protección de datos personales y seguridad digital. Este mandato unifica la transparencia, la protección de datos y la ciberseguridad en un mismo deber jurídico, vinculando la apertura informativa con la existencia de condiciones materiales de seguridad, trazabilidad, integridad y confidencialidad en el tratamiento de la información y en la operación de los sistemas.

Por su parte, el Decreto 767 de 2022, que subroga el Capítulo 1 del Título 9 de la Parte 2 del Libro 2 del Decreto 1078 de 2015, reconoce la seguridad y privacidad de la información como un habilitador esencial de la Política de Gobierno Digital. En virtud del artículo 2.2.9.1.2.1 (numeral 3.2), las entidades públicas están obligadas a integrar la seguridad y la privacidad por diseño y por defecto en todos los procesos, trámites, servicios, sistemas e infraestructuras tecnológicas, preservando de manera continua la confidencialidad, integridad, disponibilidad y trazabilidad de los activos de información. Este enfoque transforma la seguridad en una manifestación concreta del principio de legalidad tecnológica, de modo que el cumplimiento de las obligaciones de protección no constituye una práctica discrecional, sino una condición de validez del sistema digital en la función administrativa.

La protección de datos por diseño impone que todo tratamiento de información se estructure conforme a criterios de licitud, finalidad determinada, minimización de datos, limitación de acceso y control integral del ciclo de vida de la información. Los sistemas algorítmicos deben incorporar desde su arquitectura mecanismos que posibiliten la seudonimización o anonimización de los datos personales cuando su identificación no sea indispensable para el cumplimiento de la finalidad pública. La gestión del ciclo de vida de los datos debe documentarse exhaustivamente, de modo que se pueda demostrar la proporcionalidad y pertinencia del tratamiento en cada fase.

En correspondencia, la ciberseguridad por diseño exige que la protección de los sistemas, los datos y las comunicaciones se integre desde la planeación y desarrollo de los sistemas algorítmicos. De conformidad con el Decreto 767 de 2022, la seguridad debe garantizar la confidencialidad mediante controles de autenticación y cifrado; la integridad mediante verificaciones criptográficas, firmas digitales y gestión de versiones; la disponibilidad mediante planes de continuidad y recuperación ante incidentes; y la trazabilidad mediante registros inalterables de auditoría. La Directiva 007 de 2025 extiende estas exigencias a todos los sistemas algorítmicos del Estado, imponiendo la obligación de documentar y conservar evidencia técnica de cada modificación o actualización, de modo que la toma de decisiones automatizadas sea plenamente auditada.

El cumplimiento de estas obligaciones se complementa con la adopción de estándares internacionales reconocidos. El Manual de Gobierno Digital recomienda la aplicación de las normas ISO/IEC 27001 (sistemas de gestión de seguridad

de la información), ISO/IEC 27005 (gestión del riesgo de seguridad de la información) y ahora la norma ISO/IEC 42001 (sistema de gestión para inteligencia artificial), que establecen procedimientos de control de acceso, gestión de incidentes, desarrollo seguro, continuidad operativa y evaluación de vulnerabilidades.

En el plano operativo, la integración del sistema de gestión de seguridad con el régimen de protección de datos personales constituye una obligación indeclinable. En un sistema de focalización de transferencias monetarias, la entidad responsable debe justificar la base jurídica de cada dato, aplicar técnicas de anonimización, cifrar las bases de datos, restringir los accesos por niveles de autorización y registrar todas las consultas o modificaciones. Ante un incidente de seguridad o filtración de información, la entidad está obligada a activar el protocolo de respuesta correspondiente y notificarlo a la autoridad competente. En sistemas de apoyo diagnóstico en salud, la arquitectura debe separar los módulos de procesamiento y almacenamiento, aplicar cifrado a las bases de datos médicas y realizar auditorías periódicas de integridad. En plataformas de asignación de causas judiciales, la trazabilidad de las operaciones y el control de versiones de los parámetros del algoritmo garantizan el debido proceso y la posibilidad de auditoría.

La implementación del modelo enfrenta dos desafíos institucionales relevantes. El primero es la falta de coordinación entre las entidades responsables del diseño de políticas digitales, la definición de lineamientos técnicos y la vigilancia en materia de protección de datos, lo que puede producir solapamientos o vacíos regulatorios. El segundo desafío radica en la limitada interoperabilidad de los sistemas de gestión de seguridad, que en muchos casos operan de forma independiente y dificultan la verificación transversal de incidentes, controles y cumplimiento normativo. La protección de datos personales y la ciberseguridad por diseño concretan los principios de confianza, legalidad tecnológica y resiliencia digital establecidos en el Decreto 767 de 2022, y materializan para el ámbito algorítmico el principio de seguridad y los deberes consagrados en la Ley 1581 de 2012, bajo la vigilancia de la Superintendencia de Industria y Comercio. Estas obligaciones constituyen la base material y normativa sobre la cual puede ejercerse una transparencia algorítmica válida, equilibrada y segura, que haga posible el control democrático sin poner en riesgo la integridad técnica del Estado ni los derechos fundamentales de las personas.

5. Test de daño y ponderación de riesgos en la transparencia algorítmica

El test de daño constituye el instrumento jurídico de ponderación que delimita, caso a caso, el alcance del derecho de acceso a la información pública frente a bienes constitucionalmente protegidos comprometidos por la divulgación de información técnica relativa a sistemas algorítmicos y de inteligencia artificial. Su fundamento normativo directo se encuentra en los artículos 17, 18 y 19 de la Directiva Conjunta 007 de 2025, que regulan, respectivamente, la

entrega de explicaciones significativas como alternativa a la entrega de código fuente, el régimen de entrega condicionada cuando excepcionalmente proceda, y la aplicación obligatoria del test de daño previo a cualquier restricción; y en los artículos 18 y 19 de la Ley 1712 de 2014, que establecen el régimen de excepciones al acceso y la exigencia de una prueba de daño motivada, específica y proporcional. Este marco se integra con la Sentencia T-067 de 2025, que reafirma la transparencia significativa y exige motivaciones concretas y verificables cuando se imponga una reserva o se sustituya la información solicitada.

El test de daño es un procedimiento reglado de ponderación y no un margen discrecional. Vincula a toda autoridad sujeta a la Ley 1712 de 2014 que deba decidir sobre solicitudes de información referentes a sistemas algorítmicos (documentación funcional, datos de entrenamiento y evaluación, parámetros, arquitecturas, artefactos de despliegue, bitácoras y, en su caso, código fuente). La decisión debe atender al principio de máxima divulgación y al principio de divisibilidad de la información, de manera que la reserva sea excepcional, necesaria, idónea y proporcional y, cuando sea posible, limitada a los fragmentos cuya divulgación ocasionaría el daño, privilegiando la entrega parcial, la disociación o la anonimización.

El test se articula, metodológicamente, en fases sucesivas y documentables. En primer lugar, la autoridad debe identificar con precisión el objeto informativo: naturaleza, formato, granularidad y contexto de uso del insumo solicitado (por ejemplo, manuales de operación, matrices de variables, pesos de modelos, registros de auditoría, reglas de negocio, configuraciones de inferencia). En segundo lugar, debe verificar la concurrencia de una causal de excepción prevista por la Constitución o la ley, conforme al artículo 18 y 19 de la Ley 1712 de 2014 (por ejemplo, seguridad y defensa, secretos comerciales o industriales de terceros, protección de datos personales, seguridad de la infraestructura crítica o de la información). En tercer lugar, debe determinar el bien jurídico a proteger afectado por la divulgación: seguridad digital e integridad de la infraestructura algorítmica; protección de datos personales, incluidos datos sensibles o de poblaciones especialmente protegidas; propiedad intelectual y secretos comerciales de proveedores cuando su revelación sea jurídicamente relevante; o debido proceso y confiabilidad de procedimientos administrativos y judiciales automatizados.

A partir de esa determinación, procede la caracterización del daño: debe ser concreto, probable y específico, no meramente hipotético. La autoridad debe describir el escenario causal por el cual la divulgación del insumo solicitado permitiría, razonablemente, la ingeniería inversa de controles, la explotación de vulnerabilidades, la elusión de reglas, la exposición de datos personales, la afectación de la continuidad del servicio o la lesión del interés público asociado al uso legítimo del sistema. Finalmente, debe valorar la gravedad del daño y su reversibilidad, ponderando la entidad e inminencia del riesgo frente al beneficio público de la divulgación para fines de control democrático, participación informada y defensa de derechos.

La valoración exige criterios objetivos y verificables. En términos de probabilidad, la autoridad debe acreditar un nexo plausible entre el contenido a divulgar y el vector de riesgo (por ejemplo, que publicar pesos y arquitectura de un modelo de detección de fraude habilita su evasión sistemática; que exponer reglas operativas de priorización permite manipular criterios de elegibilidad; o que revelar registros de trazabilidad con identificadores indirectos reidentifica titulares). En términos de severidad, debe estimarse la magnitud del impacto potencial: indisponibilidad del servicio esencial, pérdida de integridad de la base de datos, afectación masiva de titulares de datos, desarticulación de controles antifraude o compromisos de seguridad pública. En términos de especificidad, debe demostrarse que el daño se sigue de esa divulgación concreta y no de supuestos genéricos.

Para asegurar un estándar uniforme, resulta exigible la corroboración técnica mediante insumos de los responsables de seguridad de la información y de protección de datos de la entidad, incluyendo referencias a evaluaciones de vulnerabilidades, perfiles de amenazas, análisis de superficie de ataque del sistema, y, cuando corresponda, evidencia de pruebas de penetración o de modelado de amenazas que sustenten la probabilidad y la severidad aducidas. En solicitudes que involucren potencial exposición de datos personales, debe hacerse constar si la información puede anonimizarse efectivamente, con referencia a técnicas reconocidas (p. ej., supresión, generalización, k-anonimato, l-diversidad o t-closeness) que reduzcan el riesgo de reidentificación a niveles aceptables.

Antes de disponer una reserva total, la autoridad debe agotar alternativas de divulgación menos restrictivas, en coherencia con los artículos 17 y 28. Esto incluye, según el caso, la entrega parcial con testado de segmentos críticos; la agregación o reducción de granularidad; la anonimización o pseudonimización de datos; la diferición temporal de la entrega cuando la inmediatez eleve el riesgo (p. ej., despliegues en curso); la sustitución por explicaciones significativas en lenguaje claro que describan la lógica, variables y salvaguardas del sistema; o la entrega bajo condiciones del artículo 18 (acta de entrega, restricciones de uso, no divulgación, prohibición de explotación comercial y deberes de custodia), cuando ello reduzca el riesgo a niveles aceptables y la entrega sea necesaria para fines de control académico, judicial o social.

La elección de la medida mitigadora debe motivar por qué esa opción es idónea, necesaria y proporcional frente al riesgo acreditado, y por qué resulta preferible a la reserva absoluta.

La decisión que resuelve la solicitud debe ser motivada en derecho y en hecho, describiendo el objeto informativo, la causal legal invocada, el bien jurídico protegido, el escenario del daño, la probabilidad y severidad estimadas, las medidas de mitigación analizadas y la razón por la cual se adopta la entrega, la entrega parcial, la sustitución por explicación significativa o la negativa. Debe, además, respetar los plazos y formas de la Ley 1755 de 2015 y dejar constancia de la intervención de los roles técnicos pertinentes (seguridad de la información, protección de datos, responsables del sistema). Cuando se opte por

explicación significativa, esta debe cumplir los estándares de claridad, suficiencia y verificabilidad previstos por la Directiva, permitiendo al solicitante comprender la lógica decisional, las variables relevantes, los controles de sesgo y la intervención humana significativa.

Conforme al principio de divisibilidad, si parte de la información puede divulgarse sin riesgo, la autoridad debe proceder a su entrega y explicitar los motivos de la reserva sobre los segmentos testados. La decisión debe informar las vías de impugnación y control disponibles (recursos administrativos, control contencioso y acción de tutela, según el caso).

El test de daño debe quedar documentado en un expediente de acceso que contenga: la solicitud; la identificación del objeto informativo; los informes técnicos que sustentan probabilidad y severidad; el análisis de alternativas menos lesivas; la decisión y su motivación; y, en su caso, el acta de entrega condicionada del artículo 18. Este expediente garantiza la trazabilidad para auditoría interna, control externo y eventual revisión judicial. Dado el carácter excepcional y temporal de las reservas, la autoridad debe revisarlas periódicamente y levantar o modificar la restricción cuando desaparezcan las circunstancias que justificaban el daño, dejando constancia de ello.

Cuando la solicitud se refiera a sistemas sujetos a mayores exigencias de divulgación (artículo 13), la ponderación adquiere un estándar reforzado: debe presumirse un mayor interés público en la transparencia, y la autoridad debe demostrar con mayor rigor la existencia del daño y la insuficiencia de medidas mitigadoras menos lesivas. La decisión debe armonizarse con las obligaciones ex ante de análisis de impacto (artículo 14) y con las salvaguardas de protección de datos y ciberseguridad por diseño desarrolladas en el apartado 3.4, de suerte que la eventual reserva no remedie fallas de diseño, sino proteja bienes cuya afectación sería grave, específica y probable.

En una solicitud de acceso al código fuente de un modelo de detección de fraude fiscal, la autoridad identifica que el objeto incluye pesos, arquitectura y umbrales de disparo. Verifica la causal legal relativa a seguridad y secretos industriales; determina como bienes protegidos la seguridad digital del sistema y la eficacia del control antifraude. Acredita, con informe del equipo de seguridad y de ciencia de datos, que la divulgación permitiría ingeniería inversa de umbrales y reglas, elevando sustancialmente la probabilidad de elusión sistemática y afectando la recaudación. Analiza alternativas: anonimización de datos (irrelevante al tratarse de artefactos de modelo), entrega parcial de documentación funcional, explicación significativa de la lógica y variables, y entrega condicionada a un órgano de control, con acta del artículo 18. Concluye que la sustitución por explicación significativa, acompañada de documentación técnica verificable sobre desempeño, controles de sesgo, trazabilidad y gobernanza, es idónea y suficiente para fines de control público, y menos lesiva que la entrega del código. Motiva la decisión, ofrece recursos y archiva el expediente con soporte técnico del test.

En una solicitud sobre reglas operativas de priorización de subsidios (sistema comprendido en el artículo 13), la autoridad acredita que publicar pondera-

ciones exactas en tiempo real habilita manipulación de postulaciones. Opta por la divulgación parcial de los criterios y su peso en rangos, describe la lógica de decisión y publica indicadores de desempeño y sesgo, reservando temporalmente los umbrales operativos mientras el riesgo se mantenga. La motivación explica por qué la medida es proporcional y cómo se revisará periódicamente la reserva.

El test de daño es el último eslabón del régimen de compatibilización: asegura, ex post, que la transparencia algorítmica se ejerza dentro de límites constitucionales y legales, complementando los mecanismos ex ante de explicabilidad, análisis de impacto y seguridad por diseño. Su correcta aplicación impide tanto la exposición indebida de información crítica como la invocación abusiva de la reserva, y consolida un modelo de responsabilidad demostrable en el que la Administración acredita, con evidencia técnica y argumentación jurídica, la proporcionalidad de cada decisión de acceso

El test de daño, en su configuración final, es el instrumento ex post que equilibra el derecho de acceso con la obligación de protección, completando el cuadro de compatibilización junto con los mecanismos ex ante (explicabilidad, análisis de impacto, seguridad y protección de datos por diseño). Su correcta aplicación impide la exposición indebida de información crítica y la utilización arbitraria de la reserva, consolidando un modelo de responsabilidad demostrable en el que la Administración acredita, con evidencia técnica y motivación jurídica, la proporcionalidad de su decisión.

No obstante, la implementación de este mecanismo enfrenta dos problemas estructurales que amenazan su eficacia práctica. El primero es la ausencia de metodologías estandarizadas para la aplicación del test. Aunque la Directiva 007 de 2025 exige su realización, no define criterios uniformes de valoración ni instrumentos de medición del daño. En la práctica, cada entidad aplica metodologías propias, lo que genera heterogeneidad en la interpretación y dificulta la supervisión por parte de los órganos de control. Resulta indispensable la expedición de una guía técnica interinstitucional —liderada por la Procuraduría, la Defensoría y la Agencia Nacional Digital— que unifique procedimientos, variables de análisis y formatos de documentación del test.

El segundo problema es la limitada capacidad técnica de las entidades para identificar, cuantificar y documentar los riesgos asociados a la divulgación de información algorítmica. La evaluación de amenazas de seguridad digital, la estimación de probabilidad y severidad, y la ponderación de daños requieren conocimientos especializados en ciberseguridad, ciencia de datos, derecho de acceso y protección de datos personales. Muchas entidades carecen de equipos interdisciplinarios o de recursos suficientes para realizar el test con el rigor técnico que demanda la Directiva. Esta limitación puede derivar en decisiones deficientemente motivadas, que o bien restringen indebidamente el acceso, o bien exponen información que compromete la seguridad institucional.

Superar estos dos retos exige fortalecer las capacidades institucionales en evaluación de riesgos algorítmicos y establecer un marco metodológico homogéneo que permita aplicar el test de daño de forma coherente, verificable y

jurídicamente controlable. Solo bajo esas condiciones el test cumplirá su función como herramienta de equilibrio efectivo entre transparencia y protección, asegurando que la reserva informativa deje de ser un espacio de discrecionalidad y se consolide como una práctica de ponderación racional dentro del régimen de transparencia algorítmica del Estado colombiano.

IV. CRITERIOS E INDICADORES DEL CUADRO OPERATIVO DE COMPATIBILIZACIÓN

La implementación del cuadro operativo de compatibilización exige definir criterios claros e indicadores verificables que permitan evaluar el grado de cumplimiento de las obligaciones de transparencia y seguridad digital en todas las fases del ciclo de vida de los sistemas algorítmicos. Dichos criterios, de naturaleza normativa y técnica, constituyen el puente entre la regulación y la práctica administrativa: expresan cómo se materializan, en la gestión institucional, los principios de legalidad tecnológica, trazabilidad, rendición de cuentas y protección de derechos fundamentales.

En la dimensión ex ante, los indicadores deben demostrar que la transparencia y la seguridad han sido incorporadas desde la concepción del sistema. El primer criterio es el de justificación y necesidad pública, que obliga a documentar la finalidad legítima del sistema, los objetivos que persigue y las alternativas evaluadas. Este cumplimiento se evidencia mediante la existencia de un documento técnico o acto administrativo que acredite el origen del proyecto, su relación con un problema público definido y la evaluación de opciones no algorítmicas o menos intrusivas. La ausencia de esta documentación impide verificar la proporcionalidad de la decisión de automatizar un proceso.

El segundo criterio es el análisis de impacto algorítmico, establecido en el artículo 14 de la Directiva Conjunta 007 de 2025, que constituye el principal instrumento de prevención. Su aplicación requiere la elaboración de un informe formal de evaluación de riesgos éticos, jurídicos y técnicos, acompañado de una matriz de riesgos que identifique las medidas de mitigación, los responsables de su ejecución y las fechas de seguimiento. La aprobación del análisis de impacto antes del despliegue del sistema es el principal indicador de diligencia debida institucional.

El tercer criterio se refiere al cumplimiento normativo y a la licitud del tratamiento de datos personales, conforme a la Ley 1581 de 2012. El indicador de cumplimiento consiste en la existencia de una base jurídica documentada para el tratamiento, un protocolo de manejo y protección de datos personales, y un dictamen de conformidad emitido por el delegado o área responsable de protección de datos, que certifique la adecuación del sistema a los principios de legalidad, finalidad y seguridad.

El cuarto criterio es el de seguridad y privacidad por diseño, derivado del artículo 5.g de la Directiva 007 de 2025 y del Decreto 767 de 2022. La verificación de este criterio depende de la incorporación de medidas de seguridad

técnicas y organizativas desde la fase de diseño, tales como cifrado de la información, autenticación multifactor, control de accesos, registro automatizado de auditoría y existencia de un plan de continuidad y recuperación ante incidentes aprobado por la autoridad competente. Estos elementos constituyen indicadores directos de cumplimiento y de madurez en la gestión de riesgos.

El quinto criterio corresponde a la explicabilidad y documentación técnica, prevista en los artículos 5.c y 5.d de la Directiva, que obliga a las entidades a garantizar que los sistemas sean capaces de ofrecer explicaciones significativas en lenguaje claro. Los indicadores asociados son la existencia de un documento de arquitectura algorítmica que describa variables, reglas y lógica de decisión; la integración del sistema en el registro institucional de sistemas algorítmicos, con designación de responsable técnico; y la adopción de un formato validado de explicación significativa, aprobado por las áreas jurídica y de comunicaciones.

El sexto criterio de esta fase es el de control y aprobación institucional previa, que exige la validación del sistema por los órganos de supervisión internos antes de su despliegue. Los indicadores son la acta o certificación de aprobación emitida por el comité competente y el registro formal del sistema en el inventario nacional o sectorial de sistemas algorítmicos. La existencia de este control previo demuestra la sujeción del proyecto al principio de legalidad y a la función pública de control preventivo.

En la dimensión ex post, los criterios e indicadores se orientan a verificar la capacidad del sistema para ser auditado, revisado y corregido una vez en funcionamiento. El primero es el criterio de revisión y objeción, previsto en el artículo 12 de la Directiva y en la Ley 1755 de 2015, que exige garantizar a las personas el derecho a cuestionar decisiones automatizadas y solicitar revisión humana. Este criterio se acredita con la existencia de canales digitales y presenciales accesibles, registros de solicitudes recibidas y estadísticas de tiempos y resultados de respuesta. El cumplimiento implica, además, la trazabilidad documental de cada objeción, evidenciando la intervención humana significativa en las decisiones revisadas.

El segundo criterio es el de auditorías periódicas, regulado en el artículo 20 de la Directiva. Los indicadores son la existencia de un plan anual de auditoría aprobado, los informes de resultados que incluyan medidas correctivas implementadas y la publicación de un resumen ejecutivo en la sede electrónica de la entidad. La falta de evidencia documental de auditoría constituye incumplimiento de la obligación de rendición de cuentas tecnológica.

El tercer criterio se refiere a la trazabilidad y registro de operaciones, que garantiza la posibilidad de reconstruir decisiones y validar la integridad de los sistemas. Sus indicadores son la existencia de logs de ejecución y modificación, registros automatizados de intervenciones humanas y un sistema de versionamiento documentado que permita conservar, durante un tiempo determinado, la evolución de modelos, parámetros y configuraciones.

El cuarto criterio corresponde a la aplicación del test de daño, establecido en los artículos 17 a 19 de la Directiva y en los artículos 18 y 19 de la Ley 1712 de 2014. Los indicadores de cumplimiento son la existencia de expedientes de

acceso con motivación jurídica y técnica, la documentación de la ponderación de riesgo-beneficio, el registro de test realizados y la revisión periódica de las reservas adoptadas. Su observancia demuestra que la entidad aplica la transparencia bajo parámetros de proporcionalidad y no de restricción arbitraria.

El quinto y último criterio de la fase ex post es el de retroalimentación institucional y mejora continua, que cierra el ciclo de gobernanza. Sus indicadores incluyen la existencia de planes de acción derivados de auditorías, con responsables y plazos definidos; la actualización de modelos y protocolos conforme a los hallazgos; y la elaboración de informes de seguimiento que acrediten la ejecución de las medidas correctivas. Este criterio materializa la obligación de aprendizaje institucional prevista en el Decreto 767 de 2022 y garantiza la resiliencia del sistema.

El conjunto de estos criterios e indicadores conforma un marco verificable de gobernanza algorítmica, en el que la transparencia, la legalidad tecnológica, la protección de datos y la seguridad digital pueden ser evaluadas de manera homogénea entre entidades públicas. Su integración al Modelo Integrado de Planeación y Gestión (MIPG), a los Planes Estratégicos de Tecnología e Información (PETI) y al Manual de Gobierno Digital permite articular la rendición de cuentas tecnológica con la gestión pública ordinaria, fortaleciendo la trazabilidad institucional y la capacidad de supervisión de los órganos de control.

La incorporación de estos indicadores convierte la compatibilización entre transparencia y seguridad digital en una práctica administrativa mensurable, sustentada en evidencia documental y susceptible de auditoría. El Estado puede así demostrar, de manera objetiva y verificable, la responsabilidad demostrable en el uso ético, transparente y seguro de la inteligencia artificial en la gestión pública, consolidando un modelo de gobernanza algorítmica basado en el equilibrio efectivo entre apertura, protección y legitimidad.

V. DISCUSIÓN CRÍTICA

El modelo de compatibilización entre transparencia algorítmica y seguridad digital, tal como ha sido sistematizado en este capítulo, ofrece una arquitectura normativa y operativa robusta. No obstante, su eficacia depende de supuestos técnicos y organizativos que, en el contexto institucional colombiano, plantean desafíos de primera magnitud. La presente discusión expone, con enfoque crítico y técnico, los principales problemas a afrontar para que el régimen transite de la prescripción normativa a la efectividad material, sin erosionar los bienes constitucionales en tensión.

La determinación de qué constituye un «sistema algorítmico» y, en particular, una «decisión total o parcialmente automatizada» genera fricciones interpretativas cuando el sistema funciona como apoyo no vinculante, cuando la decisión se compone de cadenas de modelos (preprocesamiento, scoring y reglas de negocio), o cuando se emplean modelos generativos de propósito general. La frontera entre recomendación técnica e incidencia decisoria jurídicamente rele-

vante no siempre es nítida. Esta ambigüedad afecta el alcance de los deberes reforzados del artículo 13, el umbral de exigibilidad del análisis de impacto y la intensidad del deber de explicabilidad. Sin una taxonomía funcional precisa —por tipología de sistema, rol en el proceso decisorio y efectos jurídicos o materiales—, se corre el riesgo de subinclusión (sistemas relevantes que quedan fuera) o sobreinclusión (cargas desproporcionadas sobre herramientas instrumentales de bajo riesgo).

El análisis de impacto algorítmico, la explicabilidad significativa y el test de daño requieren metodologías homogéneas para ser comparables y auditables. La ausencia de umbrales y métricas mínimas —por ejemplo, definiciones operativas de riesgo, criterios de severidad y probabilidad, niveles de intervención humana significativa, o formatos estandarizados de «explicación significativa»— produce heterogeneidad en la práctica y dificulta la supervisión. La consecuencia es un cumplimiento formal que satisface el trámite documental, pero no garantiza la reducción efectiva de riesgos ni la verificabilidad.

La aplicación rigurosa del régimen exige equipos interdisciplinarios con competencias en derecho administrativo y de datos, ciencia de datos, ciberseguridad y ética aplicada, así como herramientas de gestión (versionamiento de modelos, trazabilidad, registros de auditoría, inventarios de sistemas). Muchas entidades —en especial del nivel territorial— carecen de estas capacidades o de presupuestos para adquirirlas. El riesgo práctico es la «conformidad performativa»: documentos de evaluación de impacto, protocolos de transparencia o explicaciones estandarizadas que no reflejan el funcionamiento real del sistema ni habilitan un control sustantivo.

La contratación de soluciones algorítmicas en modalidad de software como servicio o mediante APIs propietarias introduce asimetrías informativas que dificultan la auditabilidad. Cláusulas de confidencialidad y restricciones de acceso a artefactos técnicos (datos de entrenamiento, pesos de modelos, pipelines de inferencia) pueden tornar ilusorio el cumplimiento de obligaciones de trazabilidad, auditoría y explicabilidad. La compatibilización entre reserva legítima de secretos comerciales y transparencia significativa exige rediseñar los pliegos de contratación para incluir derechos de auditoría técnica, exigencias de documentación y mecanismos de «entrega condicionada» conforme a los parámetros del test de daño.

Los sesgos de datos —por subrepresentación, medición deficiente o proxies socioeconómicos— comprometen igualdad material y razonabilidad técnica. La trazabilidad del linaje de datos (origen, transformaciones, descartes y justificaciones) es condición de explicabilidad y de debida diligencia. Sin catálogos de datos, metadatos estandarizados y políticas de minimización y retención, la protección de datos por diseño se vuelve meramente declarativa. La interoperabilidad entre fuentes heterogéneas añade riesgos de calidad y consistencia que impactan tanto la transparencia como la seguridad.

La seguridad digital por diseño debe evolucionar para enfrentar amenazas propias de la IA: envenenamiento de datos de entrenamiento, inversión y extracción de modelos, inferencia de pertenencia a conjuntos de entrenamiento,

ejemplos adversariales, manipulación de prompts en sistemas generativos y fuga de información a través de explicaciones. Los registros de auditoría y artefactos de trazabilidad —indispensables para la transparencia— se convierten, a su vez, en activos sensibles que requieren protección criptográfica, control de acceso granular y resguardo contra alteraciones. La seguridad de la cadena de suministro (dependencias, librerías, contenedores, artefactos de modelos) demanda inventarios y declaraciones de materiales de software y de modelos (SBOM/MBOM) para gestionar vulnerabilidades.

Modelos estocásticos, aprendizaje continuo y actualizaciones frecuentes complican la reproducibilidad estricta de resultados, elemento clave para auditoría y control jurisdiccional. La fijación de semillas, el sellado de versiones y los registros de inferencia ayudan, pero no eliminan la no determinidad inherente a ciertos modelos. Debe definirse un estándar de «reproducibilidad suficiente» para fines probatorios y de control de legalidad, compatible con la naturaleza técnica del sistema y con el debido proceso.

Los canales de revisión y objeción garantizan intervención humana significativa, pero su efectividad depende de plazos, competencias y recursos. La complejidad técnica puede trasladar cargas excesivas al ciudadano, quien, sin asistencia, difícilmente logrará impugnar decisiones. La accesibilidad, el lenguaje claro y la obligación de motivación sustantiva deben traducirse en protocolos operativos, con indicadores de tiempos de respuesta, tasas de corrección y aprendizaje institucional derivado de objeciones reiteradas.

La concurrencia de competencias entre autoridades de control, supervisores sectoriales y responsables de política digital genera solapamientos y vacíos. Sin un esquema de coordinación interinstitucional que unifique criterios, formatos y calendarios de auditoría, la supervisión se vuelve errática y desigual. La publicación de criterios interpretativos comunes y guías técnicas vinculantes reduciría incertidumbre y litigiosidad, a la par que elevaría el piso mínimo de cumplimiento.

La adopción de medidas ex ante y ex post tiene costos directos en talento, herramientas y auditorías. Un enfoque proporcional —basado en niveles de riesgo y en escalabilidad por tamaño y función de la entidad— es imprescindible para evitar que el cumplimiento devenga prohibitivo. La provisión de servicios compartidos (plantillas, repositorios de modelos, herramientas de auditoría, soporte centralizado) y la estandarización de contratos y cláusulas técnicas pueden reducir costos sin sacrificar garantías.

La presentación de información técnicamente correcta pero irrelevante o inaccesible para el control social produce una ilusión de transparencia sin capacidad real de escrutinio. Las explicaciones significativas deben responder a preguntas jurídicamente relevantes: por qué la decisión es razonable, cómo se controlan sesgos, cuál es la intervención humana, qué medidas existen para corregir errores y cómo se asegura la integridad del sistema. La métrica de éxito no es el volumen de documentos publicados, sino su capacidad para habilitar revisión y corrección efectivas.

El alineamiento con marcos internacionales —particularmente en protección de datos y en gobernanza de IA— es determinante cuando existen flujos trans-

fronterizos de datos o servicios en nube. Divergencias regulatorias pueden limitar intercambios, auditorías y exigibilidad de derechos. Un enfoque de compatibilidad normativa y cláusulas contractuales estándar que contemplen auditoría y portabilidad minimiza fricciones y preserva la capacidad soberana de control.

La velocidad de evolución de la IA tensiona la vigencia de estándares técnicos y guías operativas. Los instrumentos deben concebirse como «normas vivas», con ciclos de actualización definidos, repositorios versionados de criterios y mecanismos de consulta pública que permitan incorporar nueva evidencia científica y mejores prácticas sin demoras incompatibles con el riesgo.

Para enfrentar estos retos se imponen acciones de fortalecimiento con efectos inmediatos: adopción de guías técnicas unificadas para evaluación de impacto, explicabilidad y test de daño; incorporación obligatoria de derechos de auditoría y de documentación técnica en la contratación pública de soluciones algorítmicas; creación de servicios compartidos de trazabilidad, versionamiento y gestión de incidentes; establecimiento de indicadores transversales de desempeño (tiempos de respuesta a objeciones, tasas de corrección, métricas de sesgo y estabilidad de modelos); y desarrollo de programas de capacitación interinstitucional que integren derecho, ciencia de datos y ciberseguridad. Estas medidas convierten la compatibilización en práctica verificable, reducen la variabilidad entre entidades y elevan el estándar de responsabilidad demostrable en el uso público de la inteligencia artificial.

En conjunto, los problemas aquí identificados no desvirtúan el modelo, pero sí delinean sus condiciones de posibilidad. La transparencia algorítmica, para ser jurídicamente exigible y socialmente confiable, requiere metodologías estables, capacidades técnicas suficientes, coordinación regulatoria y una cultura institucional orientada a la trazabilidad y a la corrección temprana. Solo así el equilibrio entre apertura y protección dejará de ser una aspiración normativa para convertirse en una realidad administrativa sustentable.

VI. CONCLUSIONES

La transparencia algorítmica, tal como fue delineada por la Sentencia T-067 de 2025 y el artículo 2.h de la Directiva Conjunta 007 de 2025, constituye un principio jurídico autónomo dentro del bloque de constitucionalidad digital. No se limita a un deber de divulgación técnica, sino que se configura como una herramienta de control democrático sobre los sistemas de decisión automatizada e inteligencia artificial utilizados por el Estado. Su objeto es garantizar que el funcionamiento interno de estos sistemas —la lógica, los datos y los criterios de decisión— sea comprensible, verificable y evaluable por la ciudadanía. En consecuencia, la transparencia algorítmica se proyecta como un nuevo estándar de publicidad estatal aplicable al entorno digital y, por tanto, como un elemento constitutivo del derecho fundamental de acceso a la información pública en contextos de automatización y gobernanza algorítmica.

Las facetas activa y pasiva de la transparencia representan dos dimensiones complementarias e inseparables del derecho de acceso. La transparencia activa, sustentada en la publicación proactiva, clara y accesible de información sobre los sistemas de IA y SDA, refuerza la rendición de cuentas institucional; la transparencia pasiva, sustentada en el derecho ciudadano a solicitar y recibir información significativa, garantiza la participación y el control social. Este doble mecanismo, previsto en los artículos 7, 8, 9 y 16 de la Directiva 007 de 2025, traduce el principio de publicidad del artículo 209 de la Constitución a la era digital. La adecuada articulación entre ambas facetas evita tanto la opacidad estructural derivada del secretismo tecnológico como la dispersión informativa que impide ejercer un control efectivo sobre las decisiones automatizadas.

La seguridad digital, conforme al Decreto 767 de 2022, es un pilar estructural del modelo de gobierno digital colombiano y no un límite restrictivo a la transparencia. El decreto eleva la seguridad digital a la categoría de principio jurídico transversal, integrando la confianza, legalidad tecnológica y resiliencia tecnológica como fundamentos del ejercicio legítimo de la transparencia y la protección de datos. En este sentido, la seguridad digital se convierte en condición de posibilidad de la transparencia algorítmica: garantiza que el acceso a la información y la auditabilidad de los sistemas no comprometan la confidencialidad, integridad ni disponibilidad de los datos. Esta relación dialéctica entre apertura y protección es la que permite consolidar una confianza pública informada, en la que la ciudadanía puede comprender los algoritmos estatales sin poner en riesgo los derechos fundamentales ni la infraestructura digital del Estado.

El dilema del código fuente, analizado a partir de los artículos 17, 18 y 19 de la Directiva Conjunta 007 de 2025 y de los parámetros constitucionales de la Sentencia T-067 de 2025, sintetiza el punto de equilibrio entre transparencia y seguridad digital. El acceso al código no puede entenderse como un derecho absoluto ni como una prohibición categórica: debe resolverse mediante la aplicación documentada del test de daño, que permita determinar cuándo el interés público en la transparencia prevalece sobre los riesgos técnicos, patrimoniales o de ciberseguridad. Este modelo de ponderación convierte la transparencia algorítmica en una práctica jurídicamente estructurada y evita tanto la exposición indiscriminada de información sensible como el uso abusivo de la reserva. En consecuencia, la verdadera transparencia significativa no reside en revelar líneas de código, sino en garantizar explicaciones comprensibles, verificables y sometidas a control público, compatibles con los estándares de seguridad establecidos por el derecho administrativo digital.

La explicabilidad constituye el primer nivel de garantía de la transparencia algorítmica. Su propósito es asegurar que las decisiones derivadas de sistemas automatizados e inteligentes sean comprensibles y auditables, sustituyendo la complejidad técnica por explicaciones significativas en lenguaje claro. Este mecanismo convierte la transparencia en una obligación de inteligibilidad y permite ejercer control público sobre la racionalidad de los sistemas sin requerir acceso directo al código fuente.

El análisis de impacto y las auditorías algorítmicas operan como herramientas ex ante de prevención y control. Su aplicación permite identificar, mitigar y documentar los riesgos éticos, jurídicos y de seguridad antes de la implementación o modificación de un sistema. Estos instrumentos constituyen la base probatoria de la diligencia debida institucional y garantizan la trazabilidad del cumplimiento, alineando el funcionamiento de los sistemas con los principios de legalidad, proporcionalidad y rendición de cuentas.

Los canales de revisión y objeción trasladan la transparencia al terreno del ejercicio efectivo de derechos. Permiten que las personas afectadas por decisiones automatizadas obtengan revisión humana, corrección de resultados y explicaciones razonadas. Este mecanismo fortalece la rendición de cuentas y consolida una dimensión procedimental del control ciudadano, al asegurar que toda decisión algorítmica sea revisable y trazable conforme al principio de intervención humana significativa.

La protección de datos personales y la ciberseguridad por diseño constituyen los pilares técnicos que legitiman la transparencia. Integran la seguridad y la privacidad en el ciclo de vida de los sistemas, garantizando la confidencialidad, integridad, disponibilidad y trazabilidad de la información. Este enfoque convierte la seguridad digital en un requisito de validez del tratamiento algorítmico y asegura que la apertura informativa no comprometa derechos fundamentales ni la integridad tecnológica del Estado.

El test de daño es el mecanismo ex post que equilibra el derecho de acceso a la información con la protección de bienes constitucionalmente relevantes. Su aplicación documentada y motivada impide tanto la exposición indebida de información sensible como la invocación arbitraria de la reserva. A través de criterios técnicos y jurídicos de probabilidad, severidad y proporcionalidad, el test de daño convierte la reserva en un acto administrativo controlable y consolida la compatibilización final entre transparencia y seguridad digital.

El modelo colombiano de transparencia algorítmica se encuentra normativamente consolidado, pero enfrenta una brecha estructural entre la regulación y la capacidad institucional para su cumplimiento efectivo. La falta de estandarización metodológica, la dispersión de competencias, la dependencia tecnológica de terceros y la insuficiencia de capacidades técnicas en muchos niveles de la administración pública amenazan con transformar la transparencia en un ejercicio formal más que sustantivo. Para superar esta distancia, el régimen requiere institucionalizar metodologías comunes de evaluación de impacto, explicabilidad y test de daño; asegurar recursos y personal especializado en protección de datos y ciberseguridad; y fortalecer la coordinación entre autoridades de control. Solo mediante esta consolidación operativa será posible transitar de una transparencia declarativa hacia una transparencia efectiva, verificable y jurídicamente exigible.

La sostenibilidad del modelo depende de su capacidad de adaptarse a la evolución tecnológica sin erosionar los derechos fundamentales ni debilitar la seguridad digital del Estado. Ello exige entender la compatibilización no como un equilibrio estático, sino como un proceso dinámico de mejora continua,

apoyado en evidencia técnica, trazabilidad y control público. La transparencia algorítmica debe concebirse como una práctica viva de gobernanza digital, en la que los mecanismos de rendición de cuentas evolucionen al mismo ritmo que los riesgos emergentes de la inteligencia artificial. Solo un enfoque adaptable, interdisciplinario y basado en la responsabilidad demostrable permitirá que el principio de publicidad constitucional se mantenga vigente en la era de la automatización y que la confianza ciudadana se fundamente en hechos verificables y no en declaraciones normativas.

CONSTITUCIONAL

En la era de la inteligencia artificial, las administraciones públicas deben integrar las tecnologías algorítmicas sin poner en riesgo los derechos fundamentales. *Garantías ante las decisiones y perfiles automatizados en el sector público* aborda esta cuestión desde una perspectiva multidisciplinar y europea, ofreciendo un análisis riguroso de los marcos normativos, jurisprudenciales y éticos que deben guiar el uso responsable de la IA en la gestión pública.

La obra, dirigida por **Jorge Castellanos Claramunt** y **Adrián Palma Ortigosa**, reúne a destacados especialistas que exploran cuestiones clave como la transparencia algorítmica, la supervisión humana, el derecho a explicación, la protección de datos y los sesgos en la toma de decisiones automatizadas. Desde el impacto de la pandemia en la gestión digital hasta la evaluación automatizada de solicitudes de asilo o la transformación post-burocrática del sector público, cada capítulo realiza un estudio del nuevo ecosistema tecnológico-administrativo.

Esta obra colectiva invita a repensar el papel del Derecho en la era algorítmica, proponiendo un equilibrio entre innovación tecnológica y garantía de los derechos fundamentales. Con un enfoque crítico, propositivo y europeísta, *Garantías ante las decisiones y perfiles automatizados en el sector público* ofrece una reflexión imprescindible para juristas, académicos y responsables públicos que buscan comprender y orientar la transformación digital del Estado desde los valores de la transparencia, la igualdad y la dignidad humana.

MINISTERIO DE CIENCIA E INNOVACIÓN
UNIÓN EUROPEA
FONDO EUROPEO DE DESARROLLO REGIONAL
"Una manera de hacer Europa"
AGENCIA ESTATAL DE INVESTIGACIÓN

ISBN 979-13-87867-91-1