Revista CIDOB d'Afers Internacionals
ISSN: 1133-6595
ISSN: 2013-035X
publicaciones@cidob.org
Barcelona Centre for International Affairs
España

Antonov, Alexander
Gestionar la complejidad: la contribución de la UE a la gobernanza de la inteligencia artificial
Revista CIDOB d'Afers Internacionals, núm. 131, 2022, Mayo-Septiembre, pp. 41-68
Barcelona Centre for International Affairs
España

DOI: https://doi.org/10.24241/rcai.2022.131.2.41

Disponible en: https://www.redalyc.org/articulo.oa?id=695774517006

Revista CIDOB d'Afers Internacionals
n.º 131, p. 41-68
ISSN:1133-6595 | E-ISSN:2013-035X
DOI: doi.org/10.24241/rcai.2022.131.2.41

Fecha de recepción: 25.11.21
Fecha de aceptación: 01.04.22

Gestionar la complejidad: la contribución de la UE a la gobernanza de la inteligencia artificial

Managing complexity: the EU's contribution to artificial intelligence governance

Alexander Antonov
Investigador predoctoral, Departamento de Derecho, Escuela de Negocios y Gobernanza, TalTech-Tallinn University of Technology. *Alexander.Antonov@taltech.ee.*
ORCID: https://orcid.org/0000-0001-6692-647X

Cómo citar este artículo: Antonov, Alexander. «Gestionar la complejidad: la contribución de la UE a la gobernanza de la inteligencia artificial». *Revista CIDOB d'Afers Internacionals*, n.º 131 (septiembre de 2022), p. 41-68. DOI: doi.org/10.24241/rcai.2022.131.2.41

Resumen: En un contexto de ecosistemas digitales mundialmente cuestionados, este artículo examina el papel y la contribución de la UE al concepto emergente de la gobernanza de la inteligencia artificial (IA). Entendida esta por la UE como el ingrediente fundamental para la innovación, la adopción de sistemas de IA ha alterado nuestra comprensión de la gobernanza. Enmarcando la IA como una tecnología digital autónoma integrada en las estructuras sociales, este artículo argumenta que se puede aumentar la confianza de la ciudadanía de la UE hacia la IA si la innovación que esta comporta se fundamenta en un enfoque basado en los derechos fundamentales. Ello se evalúa a partir del trabajo del Grupo de Expertos de Alto Nivel en IA (que ha desarrollado el marco para una IA fiable) y la propuesta recién aprobada de la Comisión Europea para una Ley de inteligencia artificial (con un enfoque basado en el riesgo).

Palabras clave: Unión Europea (UE), inteligencia artificial (IA), gobernanza, derechos fundamentales, Ley de IA, IA fiable, mercado único digital

Abstract: *With digital ecosystems being questioned around the world, this paper examines the EU's role in and contribution to the emerging concept of artificial intelligence (AI) governance. Seen by the EU as the key ingredient for innovation, the adoption of AI systems has altered our understanding of governance. Framing AI as an autonomous digital technology embedded in social structures, this paper argues that EU citizens' trust in AI can be increased if the innovation it entails is grounded in a fundamental rights-based approach. This is assessed based on the work of the High-Level Expert Group on AI (which has developed a framework for trustworthy AI) and the European Commission's recently approved proposal for an Artificial Intelligence Act (taking a risk-based approach).*

Key words: *European Union (EU), artificial intelligence (AI), governance, fundamental rights, AI Act, trustworthy AI, digital single market*

En el contexto de los ecosistemas digitales, cuestionados mundialmente y que evolucionan de forma dinámica, donde el comercio y la producción de bienes, servicios e información se van desplazando progresivamente al ámbito digital, este artículo tiene como objetivo examinar el papel y la contribución de la UE al desarrollo de un marco de gobernanza emergente de sistemas de inteligencia artificial (IA). Puesto que la IA ejerce la función fundamental de amplificar, por no decir automatizar, procesos sociales que tradicionalmente llevaban a cabo seres humanos, es posible que su introducción a gran escala en la sociedad tenga un impacto revolucionario sobre la autonomía personal. Ante las numerosas preguntas que se plantean respecto a la naturaleza exacta y el alcance de la IA, así como a la capacidad para regular su aplicación en diversos contextos sociales, la UE –como uno de los líderes en el desarrollo de IA, junto con Estados Unidos y China– ha contemplado la puesta en marcha de un marco normativo diseñado para aprovechar al máximo el potencial de la IA.

Los estándares únicos a nivel mundial fomentados por la UE al respecto son de especial interés: a) el marco para una IA fiable, desarrollado por el Grupo de Expertos de Alto Nivel en IA (AI HLEG, por sus siglas en inglés, 2019a) y asentado en una concepción basada en los derechos fundamentales, y b) la propuesta posterior de la Comisión Europea para un enfoque de cuatro dimensiones basado en el riesgo contenido en su propuesta de Ley de inteligencia artificial[1] (AIA, por sus siglas en inglés) (Comisión Europea, 2018a, 2018b y 2020f). Las políticas de la UE en materia de IA están motivadas por la intención de crear un «ecosistema de la confianza»[2] y, a la vez, un «ecosistema de la excelencia»[3] (Comisión Europea, 2020). En términos más generales, el objetivo es acelerar el desarrollo del mercado único digital y empoderar a los ciudadanos y a los consumidores con la finalidad transformadora de alcanzar la «Década digital de Europa» (Comisión Europea, 2021a). En este contexto, ¿cómo se alinean estos dos enfoques, el basado en los derechos fundamentales y el asentado en el riesgo de la innovación, para lograr una «IA fiable»?

1. Véase: COM(2021) 206 final. «Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión».
2. Enfoque basado en la ética aplicada que pretende desarrollar la IA bajo una lógica reflexiva, centrándose en los ciudadanos y basándose en los derechos fundamentales.
3. Que gira principalmente en torno a tres pilares: inversión responsable, innovación e implementación de la IA. Véase: COM(2020) 605 final. «Comunicación de la Comisión al Parlamento Europeo, al Consejo Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones sobre la Estrategia de la UE para una Unión de la Seguridad», Bruselas, 24.7.2020.

Sobre eso, este artículo pretende, por un lado, examinar las formas de gobernanza de la UE aplicadas a este ámbito emergente de la política, desde la perspectiva del marco para una IA fiable desarrollado por el IA HLEG; por el otro, argumentar si dichas formas de gobernanza están basadas en los derechos fundamentales, ya que ello supondría la forma más viable para aumentar la confianza de los ciudadanos en una tecnología basada en datos cada vez más autónoma (AI HLEG, 2019a). En consecuencia, el objetivo clave es proporcionar información inicial sobre hasta qué punto los siete requisitos para una IA fiable –esto es, intervención y supervisión humanas; robustez y seguridad; privacidad y gestión de datos; transparencia; diversidad, no discriminación y equidad; bienestar social y medioambiental, así como rendición de cuentas (AI HLEG, 2019a)– se expresan en el enfoque basado en el riesgo estipulado en la AIA. Ello se realizará a través de las siguientes preguntas de investigación:

– ¿Qué es la IA y cómo se puede conceptualizar esta tecnología?
– ¿Cuál es el marco de gobernanza de la IA de la UE y cuál es su contribución al campo emergente de la gobernanza de la IA en general?
– ¿En qué medida el enfoque de cuatro dimensiones basado en el riesgo que se propone en la AIA se alinea con el concepto de la IA fiable?

Este estudio tiene en cuenta que, si bien la IA es de uso dual por naturaleza, el ámbito de aplicación del marco para una IA fiable se limita al desarrollo y utilización de la IA en el sector público (Comisión Europea, 2021b); asimismo, se centra principalmente en los factores de gobernanza externa, con lo que no proporciona un análisis de los factores de gobernanza interna variables, como el impacto de los asuntos presupuestarios en el marco de gobernanza de la IA (por ejemplo, en materia de contratación pública de IA).

Problematización de la IA

La IA como reto para la gobernanza

Calificada como una tecnología de uso general (Brynjolfsson y McAfee, 2017) y concebida como una «caja negra», el carácter de arma de doble filo de la IA proporciona razones convincentes para establecer un régimen regulador global, holístico y de naturaleza multidimensional. Promocionada como una de las tecnologías más estratégicas del siglo xxi, la adopción a gran escala de la IA en el marco de la UE condensa la promesa de aumentar la calidad de

los productos y los servicios, mejorar la eficiencia y propiciar un crecimiento económico anual por valor de 176.600 millones de dólares –si no billones de dólares– siempre y cuando su adopción en el comercio electrónico como parte de la estrategia para el mercado único digital europeo resulte exitosa en los próximos años (Scott *et al.*, 2019). Desde el punto de vista sociopolítico, promete contribuir a afrontar retos sociales en sectores como el *sanitario*, al detectar células cancerígenas; el *agrícola*, al disminuir el agotamiento del suelo, o el del *transporte*, al aumentar la seguridad y reducir potencialmente la huella de carbono (Taeihagh, 2021).

Sin embargo, entre las dimensiones social y económica de la IA aparecen tensiones, al debilitarse la confianza social hacia el uso de la tecnología tanto por parte de actores privados como del sector público. Esto se produce, principalmente, por el mal uso e incluso un abuso deliberado de la IA, como ya se ha visto, por ejemplo, en áreas como la de la aplicación de la ley para predecir la probabilidad de reincidencia, en el contexto estadounidense, o en el del uso de sistemas de identificación biométrica a distancia y de crédito social, en el contexto chino, o, más pertinente para este artículo, en el de la asignación de prestaciones sociales en la UE (Chiusi *et al.*, 2020). Estos y otros casos plantean retos para la adopción social a gran escala de la IA y exigen medidas reguladoras para restringir lo que en ocasiones se ha concebido en los informes académicos como el «leviatán digital» emergente (Langford, 2020).

Para sacar partido de la IA, el discurso social y jurídico del estudio –todavía fragmentado– de la gobernanza de la IA se centra fundamentalmente en cómo lograr una «IA responsable», una IA «ética por concepción» y una «IA fiable» (Van den Hoven, 2017; Theodorou y Dignum, 2020; Hamulák, 2018). Puesto que los ingenieros son uno de los grupos clave de las partes interesadas (*stakeholders*) en el desarrollo de la IA, se han establecido iniciativas de investigación como la *inspección-Z* con el objetivo de implicar a los desarrolladores de IA en marcos de codiseño iterativo y de hacerlos partícipes de los debates con un grupo heterogéneo de especialistas en ese campo concreto (Zicari *et al.*, 2021). Esta metodología de codiseño holística e interdisciplinaria aporta un elemento importante para concretar los requisitos del AI HLEG para una IA fiable, conciencia acerca de la sociotecnicidad de la IA y, con ello, reafirma la importancia de un enfoque basado en los derechos fundamentales para la gobernanza de la IA en el seno de la UE. Además, existen debates centrados en la creación de marcos de responsabilidad y rendición de cuentas en casos de IA potencialmente discriminatoria o con puntos débiles (Ebers, 2021). En una sociedad cada vez más interconectada y dataficada, es-

tructurada en torno a las tecnologías de la información y la comunicación[4], la adopción de la IA pone a prueba fundamentalmente la intervención humana y, con ella, los pilares y valores del proyecto de integración europea, como la dignidad humana, la democracia y el Estado de derecho (AI HLEG, 2019a; Murray, 2020).

El doble desafío, que resuena con el discurso anterior sobre las tecnologías emergentes (Larsson, 2021), es idear un mecanismo regulador proporcionado y personalizado que, por un lado, aporte un cierto grado de libertad para la innovación en IA y, por el otro, aborde las trampas que esta pueda acometer y que ya se han detectado. Dicho de otro modo, la UE se enfrenta al reto de establecer un marco normativo para el diseño, el desarrollo y la aplicación de la IA que no «subordine, coaccione, engañe, manipule, condicione o guíe en rebaño a los seres humanos de manera injustificada», sino que «aumente, complemente y empodere las habilidades cognitivas, sociales y culturales de estos» (AI HLEG, 2019a). En este sentido, si adopta un enfoque basado en los derechos fundamentales, ello puede contribuir a situar y contextualizar la aplicación a gran escala de sistemas sociotécnicos basados en la IA, alineados con los principios de proporcionalidad y necesidad. Además, daría acceso a mecanismos de reparación y rendición de cuentas en casos adversos que involucran a la IA como, en particular, hacia los grupos vulnerables de la sociedad. Asimismo, desde el punto de vista del desarrollador de la IA, el marco de derechos fundamentales ayuda a anticipar y, en consecuencia, abordar los riesgos potenciales que pueden surgir al utilizar la IA (Smuha, 2021) en un estadio inicial. La clave reside en cómo podemos lograr una IA fiable en el contexto de factores endógenos y exógenos variables, teniendo en cuenta la competencia mundial por el desarrollo de la IA, materializada en conceptos como «carrera por la IA» entre Estados Unidos, China y la UE, así como las reivindicaciones cada vez mayores por la «soberanía digital» (Pohle y Thiel, 2020).

La UE se enfrenta al reto de establecer un marco normativo para el diseño, el desarrollo y la aplicación de la IA que no «subordine, coaccione, engañe, manipule, condicione o guíe en rebaño a los seres humanos de manera injustificada», sino que «aumente, complemente y empodere las habilidades cognitivas, sociales y culturales de estos».

4. Definidas como un «campo amplio y no consolidado (...) de (i) productos, (ii) infrastructura y (iii) procesos (…) que incluye las telecomunicaciones y las tecnologías de la información, de (a) la radio y (b) las líneas telefónicas a (c) los satélites, (d) los ordenadores y (e) Internet» (ITU, 2015).

Marco conceptual de los sistemas de IA

La IA como una tecnología digital autónoma integrada en las estructuras sociales

La IA puede dividirse en diferentes metodologías y subdisciplinas (Gasser y Almeida, 2017), la más prometedora de las cuales es la formada por aplicaciones basadas en el aprendizaje automático (AA), por ejemplo, en las áreas del procesamiento del lenguaje natural, el reconocimiento de imágenes o la robótica. Tras el surgimiento de la IA como disciplina de investigación en la Conferencia de Dartmouth en el verano de 1956[5], el interés político y económico por la IA ha ido en aumento, no sin haber pasado por varios altibajos, los llamados «inviernos de la IA» y «veranos de la IA» (Russel y Norvig, 2010). Catalizada por un incremento exponencial en la cantidad de datos legibles por el ordenador, junto con la aceleración de la potencia computacional proporcionada por métodos estadísticos de AA mejorados, parece existir un nuevo impulso en la adopción a gran escala de la IA, tanto en la Administración pública como en el sector privado.

La naturaleza intangible del autoaprendizaje continuo de la IA, que funciona con software y códigos, o la información digitalizada codificada en cadenas de bits diseñadas para traducir los impulsos electrónicos con un determinado fin para amplificar o incluso substituir la capacidad física o mental del ser humano, complica nuestra interpretación de cómo diseñar un marco normativo centrado en las personas. Según la AIA, la IA es: «el software que se desarrolla con (…) (a) enfoques de aprendizaje automático (…), (b) enfoques basados en el conocimiento y en la lógica (…), (c) enfoques estadísticos (…) y que puede, para un determinado conjunto de objetivos definidos por los seres humanos, generar resultados tales como contenidos, predicciones, recomendaciones o decisiones que influyen sobre los entornos con los que interactúan» (Comisión Europea, 2021c). En este sentido, la IA se puede conceptualizar mejor como «un medio que se materializa en dispositivos basados en un código concreto» (Lawson, 2017), lo que implica un cambio en el estado mental de un ser humano y/o en el estado físico de objetos.

5. La Conferencia de Dartmouth tuvo lugar en el verano de 1956 en la universidad privada Dartmouth College (Hanover, estado estadounidense de Nuevo Hampshire), duró unas ocho semanas y fue organizada por John McCarthy (informático), Marvin Minsky (científico), Nathaniel Rochester (arquitecto jefe de IBM) y Claude Shannon (matemático, ingeniero eléctrico y criptógrafo). Este evento se considera como el germen de ver la inteligencia artificial como una esfera o campo de actividad.

De ello se deduce que el concepto y la interpretación de la tecnología digital de la IA dependen del contexto de su aplicación, de su impacto en el mundo material, así como de los métodos o medios subyacentes a través de los cuales se espera alcanzar determinados objetivos preprogramados. Su materialización siempre se traduce en un efecto sobre el mundo real y, por lo tanto, está ligada a la actualización de los procesos preprogramados de su diseñador. Al respecto, elementos de la inteligencia y el conocimiento humanos se replican y representan en la capacidad tecnológica para *percibir* el entorno digital y/o físico, *interpretar y procesar* datos estructurados o no estructurados, *decidir* realizar la acción más racional en el contexto de alcanzar un objetivo predefinido, *aprender* de este proceso y establecer *inductivamente* nuevas reglas para alcanzar los mismos objetivos de forma más eficiente (AI HLEG, 2019a). En consecuencia, la IA podría conceptualizarse como un sistema tecnológico y sociotécnico tan pronto como se alcance el umbral de sus efectos, materializándose en la actualización a través de artefactos o dispositivos materiales o dispositivos informáticos y/o robóticos. En este sentido, vale la pena recordar la relevancia de los datos, ya que el proceso de aprovechar los datos se traduce en una reflexión acerca de los valores, las normas y las estructuras sociales sobre los que se despliega la IA o que ya estarían integrados en ella (Rahwan, 2018; Larsson, 2019 y 2021).

La novedad de la IA radica en su característica inherente de autonomía, proporcionada por los algoritmos basados en aprendizaje automático, así como por su inteligencia, la cual, aunque limitada, se va desarrollando progresivamente.

Si bien este marco podría cuestionarse debido a ser de aplicación demasiado amplia, lo que convierte las tecnologías de la información y comunicación (TIC) anteriores basadas en software en «sistemas de IA», es decir, la novedad de la IA radica más bien en su característica inherente de autonomía, proporcionada por los algoritmos basados en AA, así como por su inteligencia, la cual, aunque limitada, se va desarrollando progresivamente. De forma más general, la IA es única en comparación con los sistemas sociotécnicos tradicionales debido a que tiene sus propias características «autónomas, adaptativas e interactivas» (Van de Poel, 2021), con un diseño que cambia continuamente a través de la interacción y el compromiso con el entorno a lo largo del tiempo y del espacio. Así, el nexo entre la naturaleza de la IA y su despliegue en contextos del mundo real requiere una reflexión adicional sobre el impacto de este nuevo tipo de sistema sociotécnico sobre el entorno humano y, particularmente, sus interacciones con los seres humanos. Plantear la IA como un sistema sociotécnico autónomo integrado en un entorno económico y sociolegal ayuda, de esta forma, a establecer el vínculo entre el diseño de la IA y el impacto de las opciones de diseño para el

desarrollo de la IA en la interacción humano-ordenador. Asimismo, aumenta el rol de los ciudadanos empoderándolos en los enfoques de diseño, que se pueden extender a las necesidades humanas desde una perspectiva centrada en el ser humano en el contexto de la innovación en IA. Esto resulta todavía más relevante si se tiene en cuenta que la innovación en IA ha tenido lugar fundamentalmente en clústeres de investigación privados motivados por intereses comerciales, cuya lógica se inclinaba más hacia los consumidores que al empoderamiento de los ciudadanos (Umbrello, 2021).

Por consiguiente, a fin de concretar el marco para una IA fiable del AI HLEG y, de esta forma, aprovechar el potencial de las características de naturaleza autónoma de la IA, se pide a los ingenieros, en particular, que se familiaricen y apliquen enfoques de pensamiento sistémico al diseño de la IA, los cuales se fundamentan en lógicas basadas en los derechos fundamentales. Para ello, por ejemplo, Umbrello (2021) sugiere utilizar el enfoque del diseño de concepción ética (*Value Sensitive Design* [VSD]), un marco que proporciona una valiosa caja de herramientas sobre el estudio de la interacción humano-ordenador, y cuándo/cómo adaptarla a las especificidades de la IA, particularmente respecto a la gobernanza de la IA (Umbrello y van de Poel, 2021). Como metodología específica, reflexiva, interdisciplinaria y transcultural, este marco puede generar y fomentar la concienciación sobre la importancia de las opciones de diseño en la innovación de la IA y el impacto a largo plazo de incorporar valores específicos del contexto de la IA en los ciudadanos, los usuarios finales y otras partes interesadas en los ecosistemas digitales emergentes de IA. Esto es relevante, ya que el diseño ético puede complementar el trabajo del AI HLEG, traduciendo los criterios para una IA fiable en normas concretas a través de su diseño y, por lo tanto, puede ayudar a fomentar un sistema de pensamiento basado en los derechos fundamentales para la gobernanza de la IA.

Respecto a la segunda problemática, la mejor forma de describirla es con el llamado «efecto IA», una paradoja que subyace a la utilización de todas las tecnologías basadas en IA: tan pronto como el público general adopte la IA, esta perderá la naturaleza de IA y se convertirá en una tecnología convencional (Troitiño, 2021). Pero el hecho de definir la IA en función de la gravedad y el alcance de los efectos de su utilización en los seres humanos y en el entorno en general –como se estipula en el enfoque de cuatro dimensiones basado en el riesgo de la AIA– ayuda a abordar este reto no solo desde el punto de vista semántico, sino posiblemente desde el punto de vista del Estado de derecho y de los derechos fundamentales (Comisión Europea, 2021c). Por ejemplo, un sistema de IA «que utiliza técnicas subliminales más allá de la conciencia de la persona para distorsionar materialmente su comportamiento» (Comisión Europea, 2021c) nunca podría ser tratado como una tecnología conven-

cional por y en el interior de una sociedad democrática. Podrían surgir tensiones durante la aplicación de «sistemas de identificación biométrica remota "a tiempo real" en espacios públicos con el fin de hacer aplicar la ley» (ibídem), con excepciones a su aplicación, enfrentándose a importantes críticas por parte del Parlamento Europeo y varias ONG europeas.

Sin embargo, el hecho de que las mismas aplicaciones estén restringidas por los principios legales de derechos fundamentales de necesidad y proporcionalidad no implica que estas puedan ser aprobadas sin una reflexión crítica previa por parte de una sociedad democrática bien informada por una prensa independiente y, por lo tanto, consciente de los peligros potenciales que plantea la IA a su propia dignidad. Al respecto, podrían surgir problemas en los países de la UE donde el Estado de derecho y la independencia de la prensa y del poder judicial se están socavando progresivamente. Esto podría complicar los procedimientos de revisión judicial para impugnar solicitudes individuales de IA en tribunales administrativos, por ejemplo, por la aplicación de la ley, lo que a su vez podría tener un efecto negativo sobre la confianza de la sociedad en la tecnología digital.

Igualmente, sería necesaria una mayor concienciación respecto al uso de la biometría sobre ciudadanos de «terceros países», por ejemplo, en el contexto de la protección de las fronteras. En este sentido, el campo de la biometría seguirá siendo objeto de controversia y proporcionará un terreno fértil para los debates sobre si las solicitudes procedentes de esos países deben transferirse total o parcialmente a los criterios de «prácticas prohibidas de IA» (Comisión Europea, 2021c). Aunque podríamos mencionar otras áreas grises referentes a las tensiones entre los principios de los derechos humanos y la salvaguardia de la seguridad pública, toda esta controversia se reduce a que, mientras la sociedad civil o las entidades jurídicas estén informadas sobre las trampas de la IA y estén empoderadas para desafiarlas a través de procedimientos de revisión judicial, los impactos adversos de la IA continuarán siendo cuestionados por la sociedad, lo que convertiría el «efecto de la IA» en un criterio pertinente sobre el cual podría evaluarse el despliegue de la IA. En consecuencia, al permanecer bajo el escrutinio público durante todo el ciclo de vida de la IA, la paradoja presenta un criterio sobre el umbral a partir del cual se puede evaluar el carácter democrático de una sociedad.

En esencia, ya que la IA puede construirse como un artefacto tecnológico digital autónomo integrado en un entorno económico y sociolegal, la regulación de la IA se resuelve en torno a las preguntas fundamentales de «para *quién* y *para qué propósito* se diseñarán [los sistemas de IA] y, relacionado con ello, *de quién* son propiedad, *quiénes* los usarán y *en qué contextos* se aplicarán» (Antonov y Kerikmäe, 2020).

Cómo gestionar la IA: de la gobernanza a la gobernanza de la IA

El concepto de «gobernanza» se aplica a varios campos y áreas para referirse, a grandes rasgos, a forma(s) compleja(s) de gestión y cogestión por parte de actores públicos y privados de procesos sociales, políticos y/o económicos, en el ámbito internacional, nacional o subnacional. Por lo general, por lo tanto, el marco de la gobernanza puede entenderse como un ejercicio de ordenación social y, en su definición, se incluyen los siguientes cinco elementos: a) multitud de actores –instituciones, estados, organizaciones internacionales y no gubernamentales–, b) variedad de mecanismos y c) de estructuras, d) grados de institucionalización y e) distribución de la autoridad (Katzenbach y Ulbricht, 2019).

Aunque algunos politólogos han cuestionado la profundidad analítica de este marco, debido a su amplio alcance y aplicabilidad (Kohler-Koch y Rittberger, 2006), en la adopción de tecnologías disruptivas y ante la necesidad de una concepción holística de las ramificaciones sociales de la IA (Murray, 2020), el marco de la gobernanza brinda, en primer lugar, orientación –como si fuera una caja de herramientas– para explorar y potencialmente abordar la complejidad en torno a la IA, la diversidad de actores, así como los procesos que se producen alrededor de la infraestructura tecnológica y, a su vez, sobre las modalidades y las configuraciones existentes en la interrelación entre nuestra comprensión actual del marco de gobernanza y la IA. Eso incluye, por tanto, las medidas necesarias, ya sean políticas, legislativas, reguladoras, incluyendo modalidades alternativas de regulación (Lessig, 1999) –como directrices éticas (Larsson, 2021), estándares, códigos de conducta–, o de adjudicación, a través de las cuales las fricciones sociales, políticas y económicas causadas por la adopción de la IA pueden suavizarse y, por consiguiente, equilibrar proporcionalmente los intereses públicos y privados.

En esencia, el concepto de gobernanza surgió a finales de la década de 1970, cuando actores privados entraron en el campo de la gobernanza pública impulsados por motivos de reducción de costes y eficiencia; áreas y campos sobre los cuales el Estado tradicionalmente había asumido la autoridad política, lo que con el tiempo condujo a lo que Rhodes (1996: 661) describió como una «fragmentación de la autoridad política» (véase también, Calcara *et al.*, 2020: 8). Por lo que respecta a la dimensión internacional y externa del marco de la gobernanza, en el contexto del final de la Guerra Fría e impulsada por el proceso de la globalización, la concepción Estado-céntrica había sido reemplazada progresivamente por nuevos tipos de gobernanza que culminaron en la siguiente definición consensuada de la Comisión sobre Gobernanza Global (Commission

on Global Governance, 1995: 2): «La gobernanza es la suma de las numerosas formas en que los individuos y las instituciones, públicas y privadas, gestionan sus asuntos comunes».

Esta definición plasma acertadamente el carácter cambiante del Estado, de manera que, tras ser moldeado gradualmente por factores endógenos y exógenos, «gobernar» se acabó transformando en «gobernanza». El aspecto más crucial que cabe mencionar para ambas dimensiones es la adopción de las TIC y el posterior proceso de digitalización[6]. Por lo tanto, el fenómeno de la fragmentación se aplica igualmente a –y está representado en y por– el contexto actual de disrupción tecnológica, al que han contribuido las grandes plataformas digitales, en particular los gigantes tecnológicos del GAFAM (Google, Amazon, Facebook, Apple y Microsoft). Como tal, la digitalización, que tiene un impacto en todas las dimensiones de la sociedad, ha distorsionado más la concepción tradicional del funcionamiento del Estado. El continuo proceso de fragmentación caracterizado por una complejidad tecnológica cada vez mayor tiene diversos impactos en la formulación de políticas: en su diseño, en la selección y la participación de nuevos actores, en los objetivos de las políticas y, por lo tanto, en los métodos por los que dichos objetivos se establecen para ser alcanzados.

En este sentido, han surgido nuevos marcos de la gobernanza, como la *e-gobernanza* (Garson, 2006), la *gobernanza de Internet* (Kettemann, 2020), la *gobernanza de la ciberseguridad* (Von Solms y Von Solms, 2018) y la *gobernanza algorítmica* (Katzenbach y Ulbricht, 2019). La profética observación que Lessing (1999) realizó en los inicios de la adopción de las TIC al final de la década de 1990, de que «el código [era] la ley» y de que la arquitectura del ciberespacio presentaba características únicas que requerían una revisión de los marcos de gobernanza tradicionales, es un precursor bien fundamentado de la comprensión de los fenómenos actuales socialmente disruptivos, como la elaboración de perfiles sociales con fines comerciales y el *nudging* («empujoncito»), o la proliferación de discursos de odio y la difusión de desinformación, todo ello impulsado por la innovación en las TIC a través del ciberespacio. Fundamentalmente, el gran poder económico que ejercen las plataformas digitales sobre la sociedad (Zuboff, 2019; Nemitz y Pfeffer, 2020) y la datificación que se deriva de ello (Murray, 2020) han exigido la revisión de los marcos de gobernanza tradicionales. En particular, la creciente asimetría de

6. Aquí consideramos la digitalización como un proceso tanto técnico como social de traducción de información y datos análogicos, tradicionalmente almacenados en forma de textos, a un formato legible por un aparato a través del código binario (Altwicker, 2019).

conocimiento al respecto entre los programadores y los responsables políticos aporta razones para repensar la gobernanza (Lessig, 1999; Van den Hoven, 2017; Buiten, 2018). El concepto de «*multistakeholderism*» (múltiples partes interesadas), derivado del discurso de la gobernanza de Internet (Kettemann, 2020: 30), ha ido ganando terreno, lo que requiere una participación amplia de expertos con conocimientos técnicos, jurídicos, sociales y económicos en los procesos de elaboración de políticas y leyes.

A la luz de todo esto, el desarrollo de la inteligencia artificial (IA) puede considerarse una continuación de la innovación en las TIC o las tecnologías digitales en general. Simplificando, podemos decir que la digitalización, apoyándose en las capacidades de la infraestructura actual –las capas técnica, lógica y social del ciberespacio (Schmitt, 2017)–, culmina con la llegada de la inteligencia digital en la cognificación de los procesos sociales (Kelly, 2016) como presentación de una inteligencia humana aún restringida en su definición. El problema del *control* de estos procesos, sumado a la velocidad que alcanza la luz de las señales electrónicas que recorren los cables de fibra óptica, se ve agravado por el aspecto más característico y propio de la IA, su autonomía (que parte de mecanismos de aprendizaje inductivos basados en AA) –y, por tanto, su condición de «caja negra»–, que a su vez trastoca la idea tradicional de transparencia, justicia, legalidad y rendición de cuentas. Otro factor que complica la gobernanza de la IA es la multiplicidad de actores que intervienen en su concepción, implantación, mantenimiento y fiscalización. Por último, la velocidad de la innovación en este ámbito llega a inducir a algunos investigadores a sostener que una IA no alineada con nuestro sistema de valores podría constituir una amenaza existencial para la propia humanidad (Bostrom, 2014; Dafoe, 2018).

Se puede distinguir un nuevo campo de gobernanza a partir del discurso actual presente en la literatura interdisciplinar sobre gobernanza tecnológica: la gobernanza de la IA. Tanto académicos como responsables políticos han propuesto diversos planteamientos para gestionar la IA.

Consecuentemente, además de la lista no exhaustiva de marcos de gobernanza presentada anteriormente, se puede distinguir un nuevo campo de gobernanza a partir del discurso actual presente en la literatura interdisciplinar sobre gobernanza tecnológica: la *gobernanza de la IA*. De forma dinámica, aunque fragmentada y no consolidada (Butcher y Beridze, 2019; Taeihagh, 2021), tanto académicos como responsables políticos han propuesto diversos planteamientos para gestionar la IA. El discurso sobre la gobernanza de la IA —holístico en muchos sentidos— se articula principalmente en torno a la cuestión de anticipar y reducir los riesgos futuros a corto y largo plazo mediante unas directrices de ética, procesos de construcción institucional y la inclusión de las directrices de ética en las nor-

mas de carácter vinculante (Larsson, 2021b). Como primeros investigadores que establecieron un marco de gobernanza de la IA, Gasser y Almeida (2017) y Dafoe (2018) buscaron concretar las reflexiones relativas a las oportunidades y la ramificaciones sociales de la IA en torno a las dimensiones de a) *nivel social y jurídico*, b) *nivel ético* y c) *nivel técnico* y, respectivamente, a) el *panorama técnico*, b) la *política de la IA* y c) la *gobernanza ideal* en políticas, nuevas instituciones y normas a través de consultas reiterativas con múltiples partes interesadas en el plano internacional. A partir de la motivación por crear una IA ética y explicable, han surgido subcomunidades como la iniciativa AI4Good de Naciones Unidas, vinculada a los Objetivos de Desarrollo Sostenible (ODS) de esta organización (ITU, 2021; Cowls *et al.*, 2021), y el Movimiento Éticamente Alineado de la comunidad del Instituto de Ingeniería Eléctrica y Electrónica (IEEE, 2019), lo que en sentido más amplio atiende al llamamiento de «abrir la caja negra de la IA» (Buiten, 2019) a la sociedad.

Para ayudar a los ingenieros en IA, en especial, a traducir las directrices éticas y los valores específicos de la IA a través del diseño, Umbrello y van de Poel (2020) sugirieron enfoques híbridos y de abajo arriba (*bottom-up*) para la gobernanza de la IA, partiendo del marco autorreflexivo de concepción ética VSD (Umbrello, 2021b). Demostrando cómo los principios del AI HLEG se pueden concretar como valores de orden superior a través del diseño de la IA mientras se tienen en cuenta las tensiones éticas y los posibles dilemas que este proceso pueda contener, el trabajo de Umbrello y van de Poel no solo otorga credibilidad a la operatividad del marco del AI HLEG, sino que también reafirma cuán instrumental es el pensamiento del sistema basado en los derechos fundamentales para la gobernanza y la innovación de la IA. Así, para lograr una IA fiable, piden a los ingenieros de IA que hagan su trabajo *teniendo en cuenta* principalmente los valores humanos para contrarrestar la lógica actual de la innovación mercantilizada de la IA y su influencia en el diseño de la IA; asimismo, ofrecen pautas complementarias sobre cómo hacerlo de manera más efectiva a largo plazo (Umbrello y van de Poel, 2021).

Desde la misma perspectiva, los discursos sociojurídicos sobre IA (Rahwan, 2018; Floridi *et al.*, 2018; Dignum, 2019; Larsson, 2021b) reflejan los argumentos para considerar la IA no solo como una tecnología basada en la ciencia computacional, sino también como un elemento que está integrado y situado en las estructuras sociales, por lo que plantea por igual cuestiones éticas, normativas y de valores. En particular, la propuesta de Rahwan (2018) de un contrato social algorítmico arroja luz sobre el reto social fundamental de retener la agencia humana en la adopción de la IA, como se ejemplifica en el modelo de «sociedad en el bucle» del mismo autor. Examinar y explicar la interrelación entre el desarrollo de la IA y los valores humanos en el ámbito social proporciona vías para una concepción revisada de cómo la sociedad como conjunto puede prosperar

adoptando la IA. Esta reflexión se materializa en la petición de Rahwan de un contrato social renovado en un entorno cada vez más cuantificable donde «los humanos y los algoritmos de gobernanza» interactúen entre sí (ibídem). A su vez, todo el ciclo de vida de la IA, desde su diseño hasta su auditoría, debe entenderse en el contexto social en el que se despliega la tecnología digital.

Contribución de la UE a la gobernanza de la IA

Los elementos elaborados para avanzar en propuestas de marcos de gobernanza en materia de IA se reflejan en documentos de políticas, como las múltiples *estrategias nacionales de IA* (OECD, 2021; Van Roy, 2020) y, en especial, los documentos de la UE *Directrices éticas para una IA fiable* y *Recomendaciones de políticas e inversión para una IA fiable* (AI HLEG, 2019a y 2019b), el *Libro Blanco sobre la inteligencia artificial* de la Comisión Europea (2020) y la propuesta legislativa de la UE sobre IA, la AIA (Comisión Europea, 2021c).

Por ejemplo, el marco autorreflexivo de concepción ética VSD está de acuerdo con los siete requisitos clave del AI HLEG y se refleja en el marco de la UE para una IA fiable, que trata «no solo de la fiabilidad del propio sistema de IA, sino también (…) de la fiabilidad de todos los *procesos* y *actores* que forman parte del ciclo de vida del sistema», siendo este reiterativo por naturaleza (AI HLEG, 2019a). En el nivel meta, la propuesta holística de la UE para la AIA presenta no solo la primera concreción de los marcos analizados en los discursos académicos sobre gobernanza de la IA, en particular sobre IA responsable y la iniciativa AI4Good (Rahwan, 2018; Floridi *et al.*, 2018; Dignum, 2019; Theodorou y Dignum, 2020; Cowls *et al.*, 2021; Larsson, 2021b), sino también el primer instrumento legislativo general para abordar la creciente recomendación, desde la literatura sobre gobernanza de la IA, de llevar a cabo el desarrollo institucional y la construcción de estructura de gobernanza en materia de IA, lo que consolida los debates internacionales sobre esta cuestión. Esto también se debe, en parte, a la participación de algunos de los académicos más destacados del AI HLEG, como Floridi o Dignum, cuyas investigaciones abarcan temas de los campos interdisciplinarios de la ciencia computacional, la filosofía y el derecho. Y viceversa, los informes elaborados por el AI HLEG han configurado y cambiado el discurso sobre gobernanza de la IA de los debates anteriores, poniendo el foco en la gobernanza de la IA *centrada en el riesgo* existencial (Dafoe, 2018) hacia la gobernanza de la IA *fiable*, reflejada en la adopción de los principios de la OCDE y los principios estadounidenses sobre IA (Thiebes *et al.*, 2021).

¿Confianza, excelencia o ambas?

Tanto las medidas legislativas ad hoc como los esfuerzos de estandarización son pilares para la competitividad mundial de la UE en la adopción de la IA (Data Ethics Commission, 2019). La presidenta de la Comisión Europea, Ursula von der Leyen, se comprometió con el objetivo clave recogido en las *Directrices políticas de la Comisión para 2024* de diseñar un marco normativo para la IA basado en los valores y las normas europeos (Comisión Europea, 2021d).

Desde 2018, con el establecimiento del AI HLEG –formado por 52 miembros de múltiples partes interesadas– hasta la AIA, presentada el 21 de abril de 2022 por la Comisión, la UE ha recurrido a varios instrumentos políticos para crear un marco de gobernanza para la IA, de entre los cuales destacan los tres siguientes: a) las *directrices éticas y de políticas sobre IA*, que sienta las bases para el concepto de IA fiable centrada en el ser humano y establece un estándar general; b) el *Libro Blanco sobre la inteligencia artificial*, que expone la visión de un ecosistema de IA basado en la confianza y la excelencia, y c) las *propuestas legislativas*, en especial, la Ley de Servicios Digitales, la Ley de Mercados Digitales y la última propuesta de Ley de inteligencia artificial-AIA (AI HLEG, 2019a, 2019b; Comisión Europea, 2021a).

Tanto las medidas legislativas ad hoc como los esfuerzos de estandarización son pilares para la competitividad mundial de la UE en la adopción de la IA. Ursula von der Leyen, presidenta de la Comisión, se comprometió con el objetivo clave de diseñar un marco normativo para la IA basado en los valores y las normas europeos.

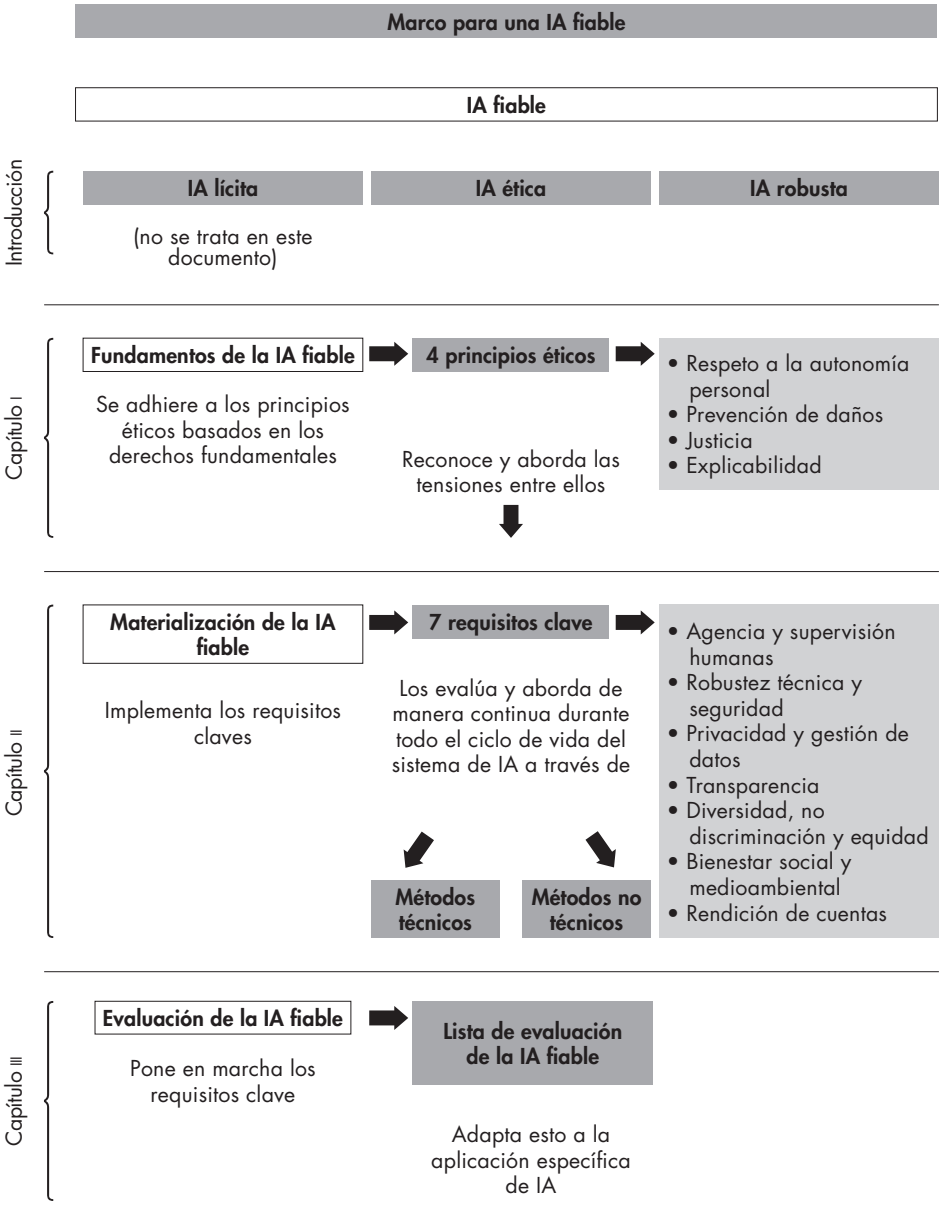
A su vez, estas iniciativas conjuntas se han ido incorporando gradualmente en un marco normativo general, aunque no exento de controversia respecto a la IA. Junto con la introducción anterior del Reglamento General de Protección de Datos (RGPD), las últimas políticas de Bruselas sobre gobernanza de la IA podrían también concebirse como una reafirmación del papel que tradicionalmente ha asumido la UE de «superpotencia reguladora» basada en principios (Bradford, 2020a; Bakardjieva Engelbrekt *et al.*, 2021). Es decir, que lo que Bakardjieva Engelbrekt *et al.* (2021) analizan como «cambio tecnológico» se ha abordado en el contexto del «efecto Bruselas» desde la perspectiva de la capacidad innata de la UE –como el mercado único más grande del mundo– para promulgar estándares legales globales que posteriormente sirven de inspiración y son adoptados más allá de su jurisdicción (Bradford, 2020b). Debemos suponer que esta concepción de Bradford (ibídem), probada empíricamente, se ve reflejada en los esfuerzos de la UE en cuanto a

regulación global de la tecnología, un ejemplo paradigmático de lo cual es el RGPD. El mismo patrón se sigue aplicando a sus esfuerzos reguladores para contrarrestar el poder monopolístico del GAFAM y otros actores tecnológicos mundiales. Si bien este artículo no pretende adoptar la metodología de Bradford, sí que partimos de su leitmotiv, a saber, la resolución inherente de la UE a actuar como un modelo para la democracia, el Estado de derecho y la promoción de los derechos fundamentales, todos principios y conceptos holísticos que son más relevantes que nunca en la coyuntura del «cambio tecnológico». Estos mismos principios, basados en los derechos fundamentales globalmente desafiados, se materializan en el marco para una IA fiable, un concepto que la UE pretende exportar y con el que, a través de la IA, quiere ostentar el liderazgo.

En este sentido, la Comisión Europea se propone gestionar el desarrollo, la comercialización y la aplicación de la IA basándose principalmente en un enfoque de cuatro dimensiones fundamentado en el riesgo (Comisión Europea, 2021c). Si bien es más completo y, por lo tanto, tiene un alcance más matizado en comparación con la propuesta binaria prevista anteriormente en el *Libro Blanco sobre inteligencia artificial*, que solo distinguía entre aplicaciones de IA de alto riesgo y de bajo riesgo (Comisión Europea, 2020), el método de clasificación basado en el riesgo de la AIA todavía parece estar sujeto a críticas. Para que la «IA hecha en la UE» sea considerada fiable (véase la figura 1), los desarrolladores y los diseñadores de estos sistemas deben cumplir tres requisitos clave: a) la IA debe cumplir todas las normas legales; b) acatar los valores y estándares éticos de las sociedades democráticas, y c) presentar un alto grado de robustez, tanto técnica –por ejemplo, en el área de la ciberseguridad– como social, por lo que respecta a la seguridad de la IA y a los principios como la explicabilidad, la equidad y la prevención de los daños y el respeto hacia la autonomía humana, en particular (AI HLEG, 2019a).

Mientras que el AI HLEG expuso en profundidad los requisitos segundo y tercero, la Comisión Europea, como órgano ejecutivo de la UE, concibió una propuesta legislativa para realizar una «IA lícita», complementando el marco para una IA fiable con el enfoque de cuatro dimensiones basado en el riesgo (véase la figura 2; Comisión Europea, 2021c). Esto, a su vez, plantea la pregunta sobre cómo está representado el marco para una IA fiable basada en derechos fundamentales en la AIA, concretamente en el contexto de la intención de la Comisión Europea de crear «una estructura de gobernanza ligera» sobre IA (Comisión Europea, 2021c). En otras palabras, ¿puede la UE alcanzar una IA fiable con una estructura de gobernanza ligera, incluida en el enfoque basado en el riesgo?

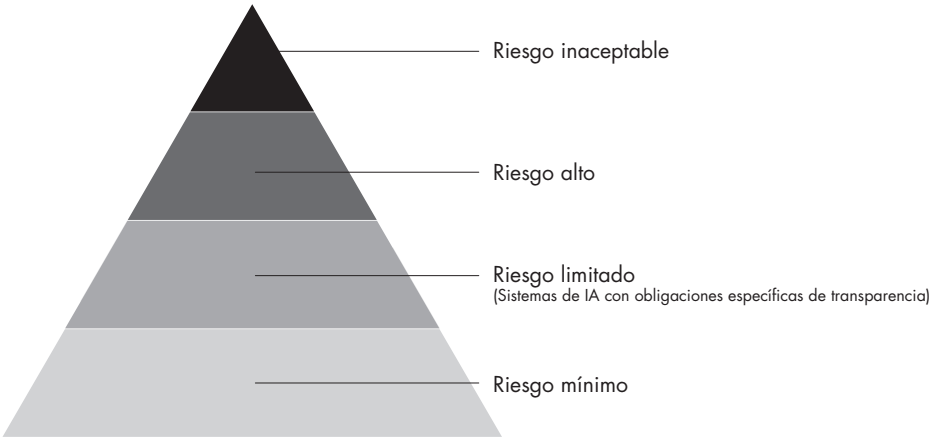
Figura 1. Marco para una IA fiable del Grupo de Expertos de Alto Nivel en IA (AI HLEG)



Fuente: AI HLEG (2019a).

Partiendo de los cinco elementos del marco de gobernanza presentado anteriormente, combinándolos con los siete requisitos para una IA fiable, este artículo examina el *feedback* recibido sobre la propuesta de AIA, en particular, el proporcionado por académicos del ámbito jurídico de los derechos humanos (véanse, por ejemplo, Smuha *et al.*, 2021), y sostiene que la AIA en su forma actual carece de elementos fundamentales para poder ser considerada fiable en la dimensión de «IA lícita».

Figura 2. Enfoque de cuatro dimensiones basado en el riesgo de la Ley de inteligencia artificial (AIA)



Fuente: Comisión Europea (2021d).

Multiplicidad de actores

Si bien la UE ha permitido la participación pública en el desarrollo de la AIA –por ejemplo, mediante encuestas a las pequeñas y medianas empresas sobre el marco para una IA fiable, el establecimiento de un foro europeo de debate sobre políticas de IA o la «Alianza IA» o el AI HLEG–, todavía quedan dudas sobre la inclusividad del proceso. Esta crítica se ve reflejada en la AIA, que no estipula *derechos procedimentales* ni *substantivos* para los ciudadanos de la UE afectados por una IA adversa (Smuha *et al.*, 2021), ya sea en áreas de aplicaciones de bajo o alto riesgo, y que está en desacuerdo con los principios para una IA fiable de *agencia humana*, *equidad*, *bienestar social* y *rendición de cuentas*. Como tal, la AIA en su forma actual no empodera a los ciudadanos hasta el punto de permitir mecanis-

mos de reparación informados y transparentes (Smuha *et al.*, 2021), en casos de usos potencialmente defectuosos o ilegales de la IA, como anteriormente había previsto el AI HLEG.

Por consiguiente, el aspecto de la participación de la ciudadanía de la UE en esta materia debe ser mejorado. Basándose en los principios del Estado de derecho de legalidad, equidad y rendición de cuentas, el ingrediente esencial de la inclusividad debería ganar peso en el debate sobre la gobernanza de la IA en el ámbito de la UE. Los marcos de gobernanza inclusiva conllevan elementos que permiten la participación ciudadana, legitimándolos, y están orientados a aumentar la transparencia del propio proceso de elaboración de políticas. Por último, pueden ajustarse con el tiempo, por lo que su alcance es reiterativo y dependiente del contexto. Aunque no es exhaustiva, la combinación de estos factores clave permite que la confianza de los ciudadanos hacia el proceso de elaboración de políticas y hacia las instituciones del Estado crezca y se desarrolle (Pierre y Peters, 2021), independientemente del momento y del contexto en que las tecnologías, de naturaleza digital, autónoma o compleja, se introduzcan.

De esta forma, las voces de los ciudadanos de la UE de a pie deberían reflejarse en la AIA, especialmente para empoderarlos, pero no solo a través de mecanismos de formulación de quejas y rendición de cuentas. En este sentido, el Parlamento Europeo, como representante de 450 millones de ciudadanos, debe defender una mayor participación ciudadana en la futura formulación y definición de la AIA, por ejemplo, introduciendo el derecho a la participación directa en las revisiones potenciales de la aplicación de la IA de riesgo. Los ciudadanos de la Unión podrían presentar casos y quejas directamente ante un consejo europeo de IA o ante sus respectivas autoridades nacionales de supervisión.

Mecanismos, estructuras, institucionalización y autoridad

En este sentido, aunque la AIA prevé establecer un *Consejo Europeo de inteligencia artificial*, con representación de expertos de los 27 estados miembros y la Comisión Europea, para permitir una aplicación adecuada de la ley, quedan dudas en cuanto a las buenas prácticas administrativas derivadas, ya que se deben desarrollar los niveles requeridos de conocimientos respecto a la infraestructura de la IA. Ello exige una exhaustiva formación de los expertos y el nivel de inversión adecuado para desplegar las capacidades de la infraestructura de las autoridades nacionales de supervisión, ya que la experiencia en torno al contexto del RGPD, por ejemplo, señala que estas instituciones normalmente han estado infradotadas en varios países de la UE, lo que ha obstaculizado y

paralizado la administración efectiva del cumplimiento de este reglamento y la aplicación de las leyes de protección de datos.

Para la perspectiva a largo plazo sobre la implementación y el seguimiento de la AIA, es especialmente importante el hecho de capacitar a los funcionarios públicos en alfabetización digital avanzada, en particular sobre cómo identificar amenazas específicas de la IA para los derechos humanos, considerando que los diseñadores de sistemas de IA tienen un gran margen de maniobra en su decisión de realizar evaluaciones de conformidad *ex ante*, lo que, dudosamente, se refiere solo a las aplicaciones de la IA de alto riesgo[7] enumeradas en el artículo 6(2) y en el Anexo III (Comisión Europea, 2021c), pero no a las de bajo riesgo. Al respecto, debido a la naturaleza compleja y a la dependencia en el contexto de la IA, si esto no se aborda, ello podría tener un impacto negativo en los derechos de privacidad y la agencia de la ciudadanía de la UE. Si bien la creación de una base de datos sobre aplicaciones de la IA de alto riesgo de ámbito europeo respeta el objetivo de supervisión y seguimiento efectivos de los sistemas críticos de IA, desde la perspectiva de la ley de derechos fundamentales, cabe preguntarse si la base de datos no debería extenderse a las cuatro dimensiones del riesgo.

Además, al transponer un espíritu de protección de los derechos fundamentales en la AIA, valdría la pena establecer un órgano de auditoría externa complementario, que revise de forma continua los criterios de *necesidad* y *proporcionalidad*, cuestionando si las vulneraciones temporales de los derechos fundamentales son *necesarias en una sociedad democrática*, legítimas y, por lo tanto, *proporcionadas* (Smuha *et al.*, 2021). Esto tiene una relevancia especial en la utilización de la IA por parte del sector público, teniendo en cuenta que la AIA permite el uso de sistemas de identificación biométrica en el ámbito del cumplimiento de la ley. Las aplicaciones adversas o mal interpretadas de los requisitos de necesidad y proporcionalidad podrían tener repercusiones graves en los requisitos para una IA fiable de: *agencia y supervisión humanas; privacidad y gestión de datos*; *diversidad*, *no discriminación y equidad; bienestar social* y *rendición de cuentas*. Aunque son necesarias para monitorear los requisitos de salud y seguridad de los productos en el área de la IA, en

7. Ejemplos de aplicaciones de la IA de alto riesgo: «(i) el componente de seguridad de los productos sujeto a la evaluación de conformidad *ex ante* por parte de terceros, y (ii) los sistemas de IA autónomos con implicaciones principalmente en materia de derechos fundamentales en las áreas de identificación biométrica y categorización de personas físicas; gestión y funcionamiento de infraestructuras esenciales; educación y formación profesional; empleo, gestión de trabajadores y acceso al autoempleo; acceso y disfrute de servicios públicos y privados esenciales y sus beneficios: aplicación de la ley; gestión de la migración, el asilo y el control fronterizo; así como administración de justicia y procesos democráticos».

general, las autoridades de vigilancia del mercado no pueden reemplazar ni asumir los roles de las instituciones, cuya competencia principal radica en salvaguardar los derechos fundamentales de los ciudadanos de la UE, extendiéndolos a áreas sensibles como la protección de datos y la privacidad.

En consecuencia, remitiéndonos a la última pregunta de investigación: ¿en qué medida el enfoque de cuatro dimensiones basado en el riesgo que se propone en la AIA se alinea con el concepto de la IA fiable?, podemos afirmar que, si bien se muestran y respetan los requisitos para una IA fiable, en especial en términos de *riesgos inaceptables* de usos de la IA que van en contra de los valores democráticos, el Estado de derecho y los derechos fundamentales como tales, la primera propuesta legislativa horizontal global en materia de IA se inclina más bien a favor de la creación de un ecosistema de la excelencia. Por lo tanto, si se compara con el RGPD, la AIA debe entenderse como una propuesta legislativa inspirada en la innovación, lo cual cambia el debate sobre la gobernanza de la IA desde un lenguaje de *IA fiable* –un enfoque basado en los derechos y respaldado por el AI HLEG– hasta un lenguaje de una *IA basada en el riesgo*, que se considera más bien favorable a la innovación, manteniéndose los valores *económicos* y *humanos* en tensión entre sí.

A modo de conclusión

El fundamento de este artículo ha sido examinar si el marco de la gobernanza de la UE en su forma actual está en consonancia con el objetivo de crear un ecosistema de IA inclusiva, basada en los derechos fundamentales y, por lo tanto, fiable. Tras tener en cuenta la naturaleza de arma de doble filo de los sistemas de IA y los retos que presentan para el desarrollo de un marco de gobernanza de la IA centrado en los seres humanos, este artículo ha analizado el discurso más reciente de la literatura sobre la gobernanza de la IA desde la perspectiva de la UE. A continuación, ha evaluado las medidas políticas de la UE en materia de IA y ha discutido la estructura de gobernanza de la IA de la UE desde una perspectiva basada en derechos fundamentales, estableciendo un contraste entre la propuesta de AIA inspirada por la innovación de la Comisión Europea, que se interpreta que representa un *ecosistema de la excelencia*, y el marco para una IA fiable del Grupo de Expertos de Alto Nivel en IA (AI HLEG), concebido principalmente para crear un *ecosistema de la confianza*. Al situar las medidas políticas de la UE en los discursos académicos sobre la gobernanza de la IA y viceversa, este artículo proporciona información inicial sobre el papel y la contribución de la UE para un marco de gobernanza emergente de la IA.

Al enmarcar la IA como una tecnología digital autónoma integrada en contextos y estructuras sociales, mediada a través de dispositivos digitales, el artículo sostiene que surgen tensiones potenciales entre ambos enfoques que se materializan en el enfoque de cuatro dimensiones basado en el riesgo contenido en la AIA, lo que, en parte, compromete los principios de una IA fiable, especialmente respecto a la *agencia humana*, *la equidad* y *la rendición de cuentas*. Sin embargo, dada la naturaleza incipiente del campo de la gobernanza de la IA, la primera propuesta legislativa de la historia en materia de IA allana el camino a debates más amplios sobre cómo las sociedades democráticas, basadas en valores inspirados en los derechos fundamentales y el Estado de derecho, pretenden vivir en un entorno cada vez más condicionado por la tecnología y la IA.

Este artículo, por lo tanto, reclama mecanismos adicionales para empoderar y permitir a los ciudadanos de la UE participar directamente en la futura configuración y gestión de la implementación de la AIA, con el fin de promover los requisitos para una IA fiable de *agencia humana*, *y supervisión* y *rendición de cuentas*. Cuestionando reiteradamente el enfoque basado en el riesgo, la investigación futura necesitaría estudios de caso empíricos basados en riesgos dependientes del contexto sobre la efectividad de las evaluaciones de autoconformidad de los diseñadores de IA partiendo del concepto de IA fiable. Deberían realizarse encuestas de opinión pública de amplio alcance, para recopilar información sobre la percepción de los ciudadanos de la UE acerca de los conceptos basados tanto en el riesgo como en los derechos fundamentales. Estos hallazgos deben respaldar y sostener las discusiones sobre ambos enfoques para alcanzar una IA fiable inclusiva. No son menos importantes las medidas para recopilar más información sobre el impacto ambiental de la adopción de las tecnologías de la IA, en estrecha colaboración con los estudios acerca del papel de la IA en la potencial reducción de la huella de carbono.

Son necesarios mecanismos adicionales para empoderar y permitir a los ciudadanos de la UE participar directamente en la configuración y gestión de la implementación de la Ley de IA, con el fin de promover los requisitos para una IA fiable de agencia humana, y supervisión y rendición de cuentas.

En esencia, en este contexto es fundamental situar los derechos de los ciudadanos en el centro y empoderarlos a través de la transformación digital a fin de desarrollar y adoptar una IA fiable. La propuesta de la Comisión para una AIA proporciona un punto de partida único de ámbito global para estos debates, en los niveles internacional, nacional y subnacional. Sin embargo, falta voluntad política no solo para respaldar parcialmente sino también para integrar por completo las ideas de los enfoques de gobernanza de la IA híbridos, por no decir de abajo arriba (*bottom-up*), en particular los que están basados en los

métodos de pensamiento del sistema basado en los derechos fundamentales del diseño de concepción ética VSD. Solo a través de la comprensión, basándose en un proceso a largo plazo y evaluando reiteradamente los riesgos de la adopción social a gran escala de la IA, los representantes públicos elegidos democráticamente pueden ayudar a empoderar a los ciudadanos hasta un punto que les permita potenciar la tecnología digital para el bien social común en los ecosistemas digitales mundialmente cuestionados.

Referencias bibliográficas

AI HLEG-High-level Expert on *Artificial Intelligence*. «Ethics guidelines for trustworthy AI». *European Commission*, (8 de abril de 2019a) (en línea) [Fecha de consulta: 15.03.2022] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

AI HLEG-High-level Expert on *Artificial Intelligence*. «Policy and investment recommendations for trustworthy Artificial Intelligence». *European Commission* (8 de abril de 2019b) (en línea) [Fecha de consulta: 15.03.2022] https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence

Altwicker, Tilmann. «International Legal Scholarship and the Challenge of Digitalization». *Chinese Journal of International Law*, vol. 18, n.º 2 (2019), p. 217-246. ¡Error! Referencia de hipervínculo no válida.

Antonov, Alexander y Kerikmäe, Tanel. «Trustworthy AI as a Future Driver for Competitiveness and Social Change». En: Troitiño, David; Kerikmäe, Tanel; de la Guardia, Ricardo Martín y Pérez Sánchez, Guillermo Á. (eds.). *The EU in the 21st Century Challenges and Opportunities for the European Integration*. Cham: Springer, p. 135-154.

Bakardjieva Engelbrekt, Antonina; Leijon, Karin; Michalski, Anna y Oxelheim, Lars. «What Does the Technological Shift Have in Store for the EU? Opportunities and Pitfalls for European Societies». En: Bakardjieva Engelbrekt, Antonina; Leijon, Karin; Michalski, Anna y Oxelheim, Lars (eds.). *The European Union and the Technology Shift*. Cham: Springer Nature, 2021, p. 1-25.

Bostrom, Nick. *Superintelligence: Paths, dangers, strategies.* Oxford: Oxford University Press, 2014.

Bradford, Anu. *The Brussels effect: How the European Union rules the world*. Oxford: Oxford University Press, 2020a.

Bradford, Anu. «The Brussels Effect Comes for Big Tech». *Project Syndicate*, (17 de diciembre de 2020b) (en línea) [Fecha de consulta: 15.03.2022] https://www.project-syndicate.org/commentary/eu-digital-services-and-markets-regulations-

on-big-tech-by-anu-bradford-2020-12
Brynjolfsson, Erik y McAfee, Andrew. «What's Driving the Machine Learning Explosion?». *Harvard Business Review*, (18 de julio de 2017) (en línea) [Fecha de consulta: 15.03.2022] https://hbr.org/2017/07/whats-driving-the-machine-learning-explosion
Buiten, Miriam. «Towards Intelligent Regulation of Artificial Intelligence». *European Journal of Risk Regulation*, vol. 10, n.º 1 (2019), p. 41-59. ¡Error! Referencia de hipervínculo no válida.
Butcher, James and Beridze, Irakli. «What Is the State of Artificial Intelligence Governance Globally?». *The RUSI Journal*, vol. 164, n.º 5–6 (2019), p. 88-96. ¡Error! Referencia de hipervínculo no válida.
Calcara, Antonio; Csernatoni, Raluca and Lavallée, Chantal (eds.). *Emerging Security Technologies and EU Governance: Actors, Practices and Processes*. Londres: Routledge, 2020.
Chiusi, Fabio; Fischer, Sarah; Kayser-Bril, Nicolas y Spielkamp, Matthias. «Automating Society Report 2020». *Algorithm Watch,* (30 de septiembre de 2020) (en línea) [Fecha de consulta: 15.03.2022] https://automatingsociety.algorithmwatch.org/wp-content/uploads/2020/12/Automating-Society-Report-2020.pdf
Comisión Europea. «Artificial Intelligence for Europe». *CE*, COM/2018/237 final, (25 de abril de 2018a) (en línea) [Fecha de consulta: 15.03.2022] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN
Comisión Europea. «Coordinated Plan on Artificial Intelligence». *CE*, COM/2018/795 final, (7 de diciembre de 2018b) (en línea) [Fecha de consulta: 15.03.2022] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0795
Comisión Europea. «European White Paper on Artificial Intelligence: a European approach to excellence and trust». *CE*, COM(2020) 65 final, (19 de febrero de 2020) (en línea) [Fecha de consulta: 15.03.2022] https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en
Comisión Europea. «A Europe fit for the digital age». *CE*, (9 de marzo de 2021a) (en línea) [Fecha de consulta: 15.03.2022] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age_en
Comisión Europea. «Fostering a European approach to Artificial Intelligence». *CE*, COM/2021/205 final, (21 de abril de 2021b) (en línea) [Fecha de consulta: 15.03.2022] ¡Error! Referencia de hipervínculo no válida.
Comisión Europea. «Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts». *CE*, COM (2021) 206, (21 de abril de 2021c) (en línea) [Fecha de consulta: 15.03.2022] https://

eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
Comisión Europea. «Regulatory framework proposal on Artificial Intelligence». *CE*, (31 de agosto de 2021d) (en línea) [Fecha de consulta: 15.03.2022] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
Commission on Global Governance. *Our Global Neighbourhood: The Report of the Commission on Global Governance*. Oxford: Oxford University Press, 1995
Cowls, Josh; Tsamados, Andreas; Taddeo, Mariarosaria y Floridi, Luciano. «A definition, benchmark and database of AI for social good initiatives». *Nature Machine Intelligence*, vol. 3, n.º 2 (2021), p. 111-115. ¡Error! Referencia de hipervínculo no válida.
Dafoe, Allan. «AI governance: a research agenda». *Governance of AI Program, Future of Humanity Institute*, (27 de agosto de 2018) (en línea) [Fecha de consulta: 15.03.2022] https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf
Data Ethics Commission of the [German] Federal Government. «Opinion of the Data Ethics Commission». *DEK*, (diciembre de 2019) (en línea) [Fecha de consulta: 15.03.2022] https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3
Dignum, Virginia. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Cham: Springer Nature, 2019.
Ebers, Martin. «Liability for Artificial Intelligence and EU Consumer Law». *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, vol. 12, n.º 2 (2021), p. 204-220.
Floridi, Luciano; Cowls, Josh; Beltrametti, Monica; Chatila, Raja; Chazerand, Patrice; Dignum, Virginia; Luetge, Christoph; Madelin, Robert; Pagallo, Ugo; Rossi, Francesca; Schafer, Burkhard; Valcke, Peggy y Vayena, Effy. «AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations». *Minds and Machines*, vol. 28, n.º 4 (2018), p. 689-707. ¡Error! Referencia de hipervínculo no válida.
Garson, George D. *Public information technology and e-governance: Managing the virtual state*. Burlington: Jones & Bartlett Learning, 2006.
Gasser, Urs y Almeida, Virgilio. «A Layered Model for AI Governance». *IEEE Internet Computing*, vol. 21, n.º 6 (2017), p. 58-62. https://doi.org/10.1109/MIC.2017.4180835
Hamulák, Ondrej. "La carta de los derechos fundamentales de la union europea y los derechos sociales." *Estudios constitucionales* 16.1 (2018): 167-186.
IEEE-Institute of Electrical and Electronics Engineers. «Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems». *IEEE*, (31 de marzo de 2019) (en línea) [Fecha de consulta: 15.03.2022] ¡Error! Referencia de hipervínculo no válida.
ITU-International Telecommunication Union. «Measuring the Information Socie-

ty Report 2015». *UN*, (30 de noviembre de 2015) (en línea) [Fecha de consulta: 15.03.2022] https://www.itu.int/en/ITU-D/Statistics/Pages/publications/mis2015.aspx

ITU-International Telecommunication Union. «AI for Good». *UN,* (2021) (en línea) [Fecha de acceso: 15.03.2022] https://aiforgood.itu.int/about/

Katzenbach, Christian y Ulbricht, Lena. «Algorithmic governance». *Internet Policy Review,* vol. 8, n.º 4 (2019) (en línea) ¡Error! Referencia de hipervínculo no válida.

Kelly, Kevin. *The inevitable: Understanding the 12 technological forces that will shape our future*. Nueva York: Viking Press, 2016.

Kettemann, Matthias C. *The normative order of the internet: A theory of rule and regulation online*. Oxford: Oxford University Press, 2020.

Kohler-Koch, Beate y Rittberger, Berthold. «Review Article: The Governance Turn in EU Studies». *Journal of Common Market Studies*, vol. 44, (2006), p. 27-50. ¡Error! Referencia de hipervínculo no válida.

Langford, Malcolm. «Taming the digital leviathan: Automated decision-making and international human rights». *American Journal of International Law*, vol. 114 (2020), p. 141-146.

Larsson, Stefan. «The Socio-Legal Relevance of Artificial Intelligence». *Droit et société*, vol. 103, n.º 3 (2019), p. 573-593.

Larsson, Stefan. «AI in the EU: Ethical Guidelines as a Governance Tool. Why Ethics Guidelines?». En: Bakardjieva Engelbrekt, Antonina; Leijon, Karin; Michalski, Anna y Oxelheim, Lars (eds.). *The European Union and the Technology Shift*. Cham: Springer, 2021, p. 85-111.

Lawson, Clive. *Technology and isolati*on. Cambridge: Cambridge University Press, 2017.

Lessig, Lawrence. *Code and Other Laws of Cyberspace*. Nueva York: Basic Books, 1999.

Murray, Andrew. «Talk at Sixth Annual T.M.C. Asser Lecture on Law and Human Agency in the Time of Artificial Intelligence». *Annual T.M.C. Asser Lecture 2020*, (26 de noviembre de 2020) (en línea) [Fecha de consulta: 15.03.2022] https://www.asser.nl/annual-lecture/annual-tmc-asser-lecture-2020/

Nemitz, Paul y Pfeffer, Matthias. *Prinzip Mensch: Macht, Freiheit und Demokratie im Zeitalter der Künstlichen Intelligenz*. Bonn: Dietz, 2020.

OECD-Organización para la Cooperación y el Desarrollo Económicos. «National AI policies and strategies». *OECD.AI*, (septiembre de 2021) (en línea) [Fecha de consulta: 15.03.2022] ¡Error! Referencia de hipervínculo no válida.

Pierre, Jon and Peters, Guy. *Advanced Introduction to Governance*. Cheltenham: Edward Elgar Publishing, 2021.

Pohle, Julia y Thiel, Thorsten. «Digital sovereignty». *Internet Policy Review*, vol. 9, n.º 4 (2020), p. 1-19. ¡Error! Referencia de hipervínculo no válida.

Rahwan, Ilyad. «Society-in-the-loop: programming the algorithmic social contract». *Ethics and Information Technology*, vol. 20 (2018), p. 5-14. ¡Error! Referencia de hipervínculo no válida.

Rhodes, Roderick. «The New Governance: Governing Without Government». *Political Studies*, vol. 44, n.º 4 (1996), p. 652-667. ¡Error! Referencia de hipervínculo no válida.

Russell, Stuart y Norvig, Peter (eds.). *Artificial Intelligence: A Modern Approach*. Londres: Pearson, 2010.

Schmitt, Michael N. *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. Cambridge: Cambridge University Press, 2017.

Scott, Marcus; Petropoulos, Georgios y Yeung, Timothy. «Contribution to growth: The European Digital Single Market. Delivering economic benefits to citizens and businesses». *European Parliament*, PE 631.044, (enero de 2019) (en línea) [Fecha de consulta: 15.03.2022] https://www.bruegel.org/wp-content/uploads/2019/02/IPOL_STU2019631044_EN.pdf

Smuha, Nathalie. «Beyond a human rights-based approach to AI governance: Promise, pitfalls, plea». *Philosophy and Technology*, vol. 34, (2021), p. 91-104. *https://doi.org/10.1007/s13347-020-00403-w*

Smuha, Nathalie; Ahmed-Rengers, Emma; Harkens, Adam; Li, Wenlong; MacLaren, James; Piseli, Ricardo y Yeung, Karen. «How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act». *SSRN Electronic Journal*, (2021), p. 1-59. ¡Error! Referencia de hipervínculo no válida.

Taeihagh, Araz. «Governance of artificial intelligence». *Policy and Society*, vol. 40, n.º 2 (2021), p. 137-157. ¡Error! Referencia de hipervínculo no válida.

Theodorou, Andreas y Dignum, Virginia. «Towards ethical and socio-legal governance in AI». *Nature Machine Intelligence*, vol. 2, n.º 1 (2020), p. 10-12. ¡Error! Referencia de hipervínculo no válida.

Troitiño, David Ramiro y Kerikmäe, Tanel. «Europe facing the digital challenge: obstacles and solutions.» *IDP: revista d'Internet, dret i política*, n.º 34 (2021).

Umbrello, Steven. «The Role of Engineers in Harmonising Human Values for AI Systems Design». *Journal of Responsible Technology*, vol. 10, (2021a), p.1-19. ¡Error! Referencia de hipervínculo no válida.

Umbrello, Steven.«Conceptualizing Policy in Value Sensitive Design: A Machine Ethics Approach». En: Thompson, Steven John (ed.). *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence*. Hershey: IGI Global, 2021b, p. 108-125.

Umbrello, Steven y van de Poel, Ibo. «Mapping value sensitive design onto AI for social good principles». *AI and Ethics*, vol. 1, n.º 3 (2021), p. 283-296.

Van de Poel, Ibo. «Embedding values in artificial intelligence (AI) systems». *Minds and Machines*, vol. 30, n.º 3 (2020), p. 385-409. ¡Error! Referencia de hipervínculo no válida.

Van den Hoven, Jerome. «Ethics for the Digital Age: Where Are the Moral Specs? – Value Sensitive Design and Responsible Innovation». En: Werthner, Hannes y van Harmelen, Frank (eds.). *Informatics in the Future*. Cham: Springer, 2017, p. 65-76.

Van Roy, Vincent. «AI Watch - National strategies on Artificial Intelligence: A European perspective». *European Union*, JRC122684 (junio 2020) (en línea) [Fecha de consulta: 15.03.2022] https://publications.jrc.ec.europa.eu/repository/handle/JRC119974

Von Solms, Basie y Von Solms, Rossouw. «Cybersecurity and information security–what goes where?». *Information and Computer Security*, vol. 26, n.º 1 (2018), p. 2-9.

Zicari, Roberto V.; Ahmed, Sheraz; Amann, Julia; Braun, Stephan; Brodersen, John; Bruneault, Frédérick; Brusseau, James; Campano, Erik; Coffee, Megan; Dengel, Andreas; Düdder, Boris; Gallucci, Alessio; Krendl Gilbert, Thomas; Gottfrois, Philippe; Goffi, Emmanuel; Bjerre Haase, Christoffer; Hagendorff, Thilo; Hickman, Eleanore; Hildt, Elisabeth; Holm, Sune; Kringen, Pedro; Kühne, Ulrich; Lucieri, Adriano; Madai, Vince I.; Moreno-Sánchez, Pedro; Medlicott, Oriana; Ozols, Matiss; Schnebel, Eberhard; Spezzatti, Andy; Jahan Tithi, Jesmin; Umbrello, Steven; Vetter, Dennis; Volland, Holger; Westerlund, Magnus y Wurth, Renne. «Co-design of a trustworthy AI system in healthcare: Deep learning based skin lesion classifier». *Frontiers in Human Dynamics*, vol. 3, (2021), p. 1-20. https://doi.org/10.3389/fhumd.2021.688152

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Londres: Profile Books, 2019.

Traducción del original en inglés: Maria Gené Gil y redacción CIDOB.

Este artículo forma parte del proyecto de investigación de la Cátedra Jean Monnet "La Europa Digital y su influencia en la integración futura". N.º identificación: 101082988. ERASMUS-JMO-2022-HEI-TCH-RSCH. Programa: ERASMUS2027.