# LA TOMA DE DECISIONES AUTOMATIZADA Y EL CONTROL DE HUMANIDAD

Fernando H. Llano Alonso
Decano de la Facultad de Derecho
Universidad de Sevilla (España)

## INTRODUCCIÓN

La relación entre el ser humano y la Inteligencia Artificial (IA) ante el horizonte de la singularidad ha alimentado tradicionalmente la imaginación de los autores que cultivan el género de la ciencia ficción y la distopía. En la mayoría de estas novelas distópicas son las máquinas inteligentes las que terminan controlando a la humanidad, ante la imposibilidad de ésta de competir con un adversario infinitamente superior en velocidad de procesamiento de datos, capacidad de almacenamiento de información, resistencia y precisión en la ejecución de tareas, y sobre todo en aprendizaje.

> Cuando los humanos crearon un ordenador capaz de almacenar la información y aprender de ella, firmaron la sentencia de muerte de la humanidad.

Con esta cita se inicia *Dune. La yihad butleriana* (2004), primera novela de la trilogía de Brian Herbert y Kevin J. Anderson sobre *Leyendas de Dune*, que a su vez es el preludio de la saga *Dune* de Frank Herbert. Quizás sea esta la primera novela a contracorriente de la mayoría de novelas futuristas y distópicas en la que los últimos humanos libres se revelan en los Planetas Sincronizados, bajo el liderazgo de Serena Butler, contra el dominio de las máquinas pensantes desarrolladas con IA (todas ellas subsumidas en el supercomputador *Omnius*) y consiguen vencerlas rotundamente. Así describen Herbert y Anderson el comienzo del

fin del ciclo de dominio de las máquinas inteligentes y autoconscientes sobre los humanos:

> Desde las profundidades de sus ciudades de energía, Omnius observaba la Tierra. Sus ojos espía grababan cada fase del audaz ataque humano. Vio que las tornas cambiaban (Herbert-Anderson, 2011[5], 619)

A diferencia de otras novelas y películas de ciencia ficción, *Dune* supone una rara excepción a la resignación con la que los seres humanos aceptan la llegada de la era de la posthumanidad. Por ello, apelando precisamente al espíritu indómito de los últimos seres humanos autoconscientes y a su ansia de libertad, Serena Butler prende la llama de la rebelión contra las máquinas pensantes. Hay en esta novela una voluntad de resistencia colectiva de los seres humanos conscientes de su condición de sujetos morales y, por ende, dotados de conciencia moral, sentido de la responsabilidad y sabiduría práctica, de su derecho a la libertad y a la dignidad, a pesar de la opresión que sufren *de facto* a causa de la Singularidad tecnológica.

A propósito de la inmersión de la humanidad en el universo de la Singularidad se pregunta Slavoj Žižek si la entrada de la especie humana en el reino superior de lo posthumano supondrá la desaparición gradual de la humanidad tal y como la conocemos o si, por el contrario, nuestros futuros descendientes posthumanos que vivan en ese "estadio superior" de la perfección serán capaces de recordar de algún modo, incluso con una memoria entrañable, los azares del "estadio inferior" de la humanidad (Žižek, 2023, 142-143). En definitiva, ¿cómo se enfrentará la humanidad a la Singularidad (ese "gran otro" que, según Jacques Lacan, constituye el orden simbólico virtual integrado en la red digital), lo hará de manera consciente o inconsciente? En esa esfera de los múltiples metaversos ¿serán los humanos capaces de diferenciar la realidad virtual de lo real, las experiencias compartidas sincrónicamente

por todos los internautas neuroconectados a través de la IA de la experiencia subjetiva e individual vivida por cada persona?

Para ilustrar esta dicotomía en torno a la conservación del control de la conciencia subjetiva en el espacio infinito de la Singularidad donde se recrea una realidad que no existe pero que es simultáneamente compartida de forma colectiva, y donde no se sabe qué organismo regulador decidirá qué experiencias pueden compartirse ni quién controlará a su vez ese mecanismo, Žižek trae a colación la última escena de un film de ciencia ficción de culto: *Matrix* (1999). En dicha escena, Neo, el protagonista, un programador que actúa como pirata informático y que intuye que algo no va bien en el universo *virtual* en el que cree haber vivido *realmente* todo el tiempo, anuncia la liberación de la humanidad tras destruir Matrix, la realidad inducida por las máquinas a las mentes conectadas de hombres y mujeres alienados y sumidos en un letargo onírico. La paradoja de este final tan ambiguo de la película de las hermanas Wachowski es que, por un lado se anuncia a la humanidad su liberación del dominio de las máquinas en un mundo irreal, pero por otra parte se les arroja al "desierto de lo real", un mundo destruido anteriormente por las máquinas (en el film se muestra concretamente la ciudad de Chicago reducida a escombros) en el que impera el vacío, la nada (Žižek, 2023, 84-85).

El recurso a la novela y el cine de ciencia ficción en la presentación de este último capítulo se debe en realidad a la necesidad de situar las conjeturas y elucubraciones del evolucionismo tecnológico y posthumanista fuera de los márgenes estrictamente científicos, es decir, el que se corresponde con los hechos probados. Recientemente Erik J. Larson se ha referido a la supuesta inevitabilidad de la Singularidad tecnológica en un libro que pretende desmontar el mito de la inexorabilidad e infalibilidad de la Inteligencia Artificial, que cuenta entre sus principales valedores a futurólogos transhumanistas como Ray Kurzweil y Nick Bostrom.

Según Larson, el peligro de tomarse en serio el futuro del mito de la IA es que puede deparar consecuencias muy negativas, ya que subvierte la ciencia. En este sentido, sostiene este científico experto en computación y emprendedor tecnológico

> La ciencia de datos (la aplicación de la IA a los macrodatos) es, en el mejor de los casos, una prótesis del ingenio humano; en caso de usarla de manera correcta, nos ayudará a lidiar con el "diluvio de datos" contemporáneo. Cuando se la usa para reemplazar la inteligencia individual, tiende a estropear la inversión sin ofrecer ningún resultado (...) Estamos pagando un precio demasiado elevado por este mito. Como no poseemos ninguna buena razón científica para creer que el mito pueda hacerse realidad, puesto que contamos con todos los motivos para rechazarlo a fin de alcanzar la prosperidad en el futuro, tenemos que repensar de manera radical la conversación sobre la IA (Larson, 2022, 11).

Dejemos pues la ficción distópica aparte y vayamos a la realidad de la interacción humana con la automatización y las simulaciones interactivas con robots y sistemas de IA. Un claro ejemplo de la interacción remota entre el ser humano y la máquina de circuito cerrado lo proporciona el Internet táctil con el ser humano "en el bucle", es decir, dentro del sistema de aprendizaje automático para ayudar a la computadora a tomar las decisiones correctas en la construcción de un modelo o en una simulación (*Tactile Internet with Human-in-the-Loop*).

En el siguiente epígrafe se presentarán sucintamente las cuatro variantes posibles que los expertos en robótica e ingeniería de telecomunicaciones distinguen a propósito de la interacción entre los humanos y las máquinas inteligentes, desde la que asegura el control humano de la máquina (*Human-in-the-Loop*), pasando por dos tipos intermedios de transferencia de habilidades y retroalimentación entre humanos y máquinas (*Human-in-the-Loop*

*for Exceptions* y *Human-on-the-Loop*), hasta la cuarta modalidad, radicalmente opuesta a la primera, en la que se excluye al ser humano de la toma de decisiones automatizada de la máquina (*Human-out-of-the-Loop*).

## CUATRO MODELOS DE INTERACCIÓN EN LA RELACIÓN ENTRE HUMANOS Y MÁQUINAS INTELIGENTES DENTRO DEL PROCESO DE TOMA DE DECISIONES

El proceso de digitalización y automatización vinculado al fenómeno de la revolución 4.0 ha metamorfoseado el modo tradicional de gestionar la toma de decisiones de las administraciones, de las empresas, de los servicios profesionales en todos los órdenes de la vida pública (sanitario, jurídico, educativo, científico-tecnológico...)

Imaginar un sistema autónomo y mecanizado de toma de decisiones resulta tan sencillo que no son pocos los psicólogos y especialistas en el estudio del comportamiento humano que coinciden en señalar cómo el uso de nuevos y sofisticados generadores de textos e imágenes impulsados por IA -por ejemplo, mediante herramientas como el Chat GPT- está afectando a la creatividad humana, el juicio y la toma de decisiones.

Ahora bien, desde el momento en que en el proceso de toma de decisión de una máquina inteligente interviene el ser humano, en la medida en que las consecuencias de la micro-decisión automatizada pueden afectarle (piénsese, por ejemplo, en un vehículo autónomo que ofrece a su conductor alternar las opciones de conducción manual o asistida), resulta imprescindible el diseño de un sistema de decisión que permita al ser humano tener una mínima interacción con la máquina (Ross-Taylor, 2021).

Los cuatro principales modelos de gestión que se han desarrollado para la interacción entre los humanos y las máquinas inteligentes varían en función del grado de implicación y de la naturaleza de la intervención humana: *Human-in-the-Loop* (HITL), *Human-in-the-Loop for Exceptions* (HITLFE), *Human-on-the-Loop* (HOTL), y *Human-out-of-the-Loop* (HOOTL).

A continuación se explicarán sucintamente cada una de estas variantes:

## HUMAN-IN-THE-LOOP (HITL)

Este primer modelo de interacción con la máquina permite al usuario llevar el control en el proceso de aprendizaje automático (*machine learning*) y hacer una transferencia de competencias a la máquina o robot inteligente. En el ámbito de las simulaciones virtuales HITL, por citar algunos supuestos: simuladores de vuelo (por ejemplo, el X-62A Vista, utilizado por la Fuerza Aérea de los Estadios Unidos), de conducción urbana (como el sistema español CARLA, acrónimo del inglés Car Learning to Act), o incluso de medicina "in silico" para probar medicinas y tratamientos experimentales en pacientes virtuales (uno de ellos, el simulador HeartFlow Analysis ayuda a localizar la enfermedad de las arterias coronarias a partir de las imágenes de tomografía computerizada del corazón de un paciente), los usuarios interactúan en tiempo real con la simulación a través de una interfaz gráfica, y se pueden recopilar directamente datos de los usuarios en un entorno experimental controlado (Rao, Chernyakhovsky, Rao, 2011, 157).

Se pueden elegir distintos caminos para intercambiar habilidades entre humanos y máquinas. Una vía para este tipo de transferencia de habilidades consiste en equipar a un experto humano con cualquier tipo de interfaz hombre-máquina que funciona como un simple mando a distancia (por ejemplo, el controlador de una videoconsola). Sin embargo, hay que tener en cuenta que no todos los expertos humanos son capaces de manejar un mando de este

tipo, ni entienden la relación entre los movimientos y la precisión del robot. Sin embargo, la industria de la robótica reconoce este problema de especialización humana desde hace tiempo, y por eso se establecen procesos estándar para transmitir conocimientos y experiencia a los robots industriales.

Cada máquina necesita ser controlada por un solo humano. Por eso, incluso en escenarios con millones de consumidores que demandan una habilidad específica del robot, el control humano debe ser escalonado a la hora de transferir las habilidades de un solo operador a millones de robots para satisfacer adecuadamente la demanda de los consumidores (Fitzek-Li-Speidel-Strufe, 2021, 4-5).

Pero la robótica actual también permite un *feedback* o retroalimentación multimodal que permite la transferencia a los humanos de competencias aprendidas automáticamente por la máquina a través del proceso de *machine learning*. En esta transferencia de habilidades en sentido inverso es la máquina quien enseña al humano; pensemos, por ejemplo, en los *wearables* o dispositivos inteligentes que portamos a diario, como pulseras, gafas o relojes, y que al estar interconectados a Internet permiten conocer con precisión datos como nuestra localización exacta en tiempo real, nuestro ritmo cardiaco, los kilómetros que hemos recorrido o las calorías consumidas.

Suponiendo que los *wearables* no sólo estén equipados con sensores sino también con *actuadores* (dispositivos mecánicos esenciales que sirven para mover o actuar sobre otros dispositivos mecánicos), las señales de aprendizaje pueden generarse en directo o con antelación y transmitirlas al usuario humano. En la rehabilitación física, por ejemplo, los movimientos de un fisioterapeuta a distancia podrían generarse en línea (es decir, a distancia) y transmitirse a los *wearables* equipados con actuadores que lleve una persona mayor con movilidad reducida, o a un dispositivo ciberfísico que ayude a realizar ejercicios de

fisioterapia a los pacientes en casa. Los posibles ámbitos de aplicación de esta transferencia de habilidades no se limitan tan solo a los cuidados sanitarios y de enfermería, sino que también comprenden la enseñanza de nuevas habilidades en la escuela, el trabajo o los intereses personales, es decir, que se encuentran dentro del amplio espectro del Internet de las Habilidades.

## HUMAN-IN-THE-LOOP-FOR-EXCEPTIONS (HITLFE)

La mayoría de las decisiones están automatizadas en este segundo modelo, y el humano sólo se ocupa de las excepciones. Para las excepciones, el sistema requiere algún juicio o aportación del humano antes de tomar la decisión. Además, los humanos también controlan la lógica del sistema para determinar cuáles serían las excepciones susceptibles de revisarse en el caso de que su intervención fuera necesaria, por ejemplo, cuando se pone en cuestión el nivel de confianza en las predicciones de un sistema predictivo de IA basado en algoritmos de aprendizaje automático.

En relación con la falibilidad de los sistemas de algoritmos predictivos, un caso paradigmático es el de Glossier, una marca de belleza online que desarrolló un algoritmo de aprendizaje automático para predecir el comportamiento de sus clientes y su compromiso futuro en el aumento de ventas de distintas promociones de sus productos. Sin embargo, la predicción *machine learning* asumió erróneamente que las visitas a su sitio web "Into the Gloss" de los potenciales clientes se traducirían necesariamente en compras online teniendo en cuenta factores como la oferta, el apoyo de marketing y la estacionalidad para crear una previsión automatizada. Y en efecto, para muchas promociones la predicción *machine learning* funcionó bien, pero los directivos perdieron rápidamente la confianza después de que, tras su éxito inicial, se produjera una serie de estrepitosos fracasos predictivos que se tradujeron en importantes pérdidas de ventas.

Cuando los científicos de datos revisaron las predicciones algorítmicas basadas en los datos extraídos de los millones de datos y patrones de comportamiento de sus clientes en las plataformas digitales y redes sociales, descubrieron que el algoritmo de aprendizaje automático tenía dificultades para predecir ciertos tipos de promociones. Pero, en lugar de abandonar el proyecto, los científicos decidieron desarrollar un enfoque HITLFE. La clave consistía en codificar el nivel de confianza de la máquina en sus predicciones y hacer que los humanos revisaran dichas predicciones de forma excepcional cuando la máquina no les ofreciera la fiabilidad esperada (Ross-Taylor, 2021).

## HUMAN-ON-THE-LOOP (HOTL)

En el tercer modelo de interacción entre humanos y máquinas inteligentes o robots, son los humanos quienes asisten a las máquinas en su toma de microdecisiones mecánicas. En los sistemas de HOLP la toma de decisiones debe automatizarse al 100% porque la respuesta debe producirse casi de manera inmediata, en un periodo de tiempo demasiado corto como para incluir en el sistema al ser humano. Sin embargo, aunque el operador humano quede excluido en primera instancia de ese proceso de toma de decisión automatizada, sí puede supervisar posteriormente los resultados de estas decisiones automatizadas e incluso ajustar reglas y parámetros para futuras decisiones. En una configuración algorítmica más avanzada, la máquina también recomienda parámetros o cambios en las reglas que luego son supervisados por un humano.

Un ejemplo de aplicación del sistema autónomo de toma de decisiones HOTL lo encontramos en una aseguradora global de vida y salud que necesita aumentar el valor de sus ventas. En este supuesto, los agentes de ventas de esta empresa utilizan un dispositivo móvil para trabajar con posibles clientes, recopilando datos sobre sus necesidades financieras y su situación, les y ayudan

a solicitar electrónicamente productos de protección e inversión. El aumento de los ingresos resultará de gran utilidad a los agentes para identificar las oportunidades adecuadas de venta cruzada y venta adicional mientras tratan con los clientes potenciales.

Posteriormente, el departamento de marketing de esta compañía aseguradora utilizará análisis avanzados e IA para identificar ofertas para cada cliente potencial en función de su situación financiera y de los productos que compre. Este sistema de toma de decisiones a nivel de cliente se integra perfectamente en la aplicación móvil de la aseguradora y se ejecuta en tiempo real, de forma totalmente autónoma. Luego el equipo de marketing gestionará el bucle de retroalimentación y obtendrá datos detallados sobre cómo se seleccionan las ofertas, podrá hacer un seguimiento de las estrategias de los clientes y realizar cambios en el comportamiento del sistema.

Por cierto, según algunos estudios recientes sobre el impacto de la IA en el futuro del sector de los seguros en la próxima década (2030), las aseguradoras que prosperarán más serán precisamente aquellas que mejor utilicen las nuevas tecnologías para crear productos innovadores, aprovechar los conocimientos del aprendizaje cognitivo a partir de nuevas fuentes de datos, agilizar los procesos y reducir los costes, y superar las expectativas de los clientes en cuanto a individualización y adaptación dinámica (Balasubramanian-Libarikian-McElhaney, 2021).

## HUMAN-OUT-OF-THE-LOOP (HOOTL)

En el cuarto modelo de interacción, radicalmente opuesto al primero (HITL), la máquina es quien toma todas las micro-decisiones, aunque ésta es supervisada desde fuera del sistema por el hombre que interviene tan solo para establecer nuevos límites y objetivos. En esta modalidad el *feedback* del sistema se encuentra dentro de un bucle cerrado, por lo que los ajustes de la

retroalimentación de datos entre los humanos y las máquinas se hallan automatizados.

A propósito del sistema de toma automatizada de decisiones HOOTL, el *Mayflower Autonomous Ship*, un buque completamente autónomo (no tripulado) de investigación marina cuyo software ha sido co-diseñado por la compañía IBM y la organización sin fines de lucro Promoting Marine Research and Exploration (ProMare), recopila a lo largo de su travesía cantidades masivas de datos, incluso en entornos meteorológicos hostiles, y tiene capacidad de tomar decisiones cruciales en fracciones de segundo sin intervención humana de ningún tipo. La capitana de la nave es la IA que navega autónomamente, surca los mares evitando los peligros oceánicos y observa rigurosamente las reglas del Derecho Marítimo, con el fin de alcanzar los objetivos preestablecidos por los responsables del proyecto científico (que son seres humanos).

Respecto a los cuatro modelos de interacción que se han explicado de forma sucinta solo cabría añadir que, con independencia de cuál sea la opción de gestión con la que deseemos interactuar con la IA (ya sea desde la mayor implicación humana en la supervisión algorítmica del HITL, a la completa autonomía de la máquina en la toma de decisiones propia del HOTL), todo sistema de micro-decisión debe ser revisado para garantizar que la toma de decisión automatizada es la adecuada, para ello es fundamental que se eviten las decisiones herméticas que se amparan en la opacidad de los algoritmos de cajas negras, y que ninguna decisión algorítmica que se aplique al mundo real pueda optimizarse en tanto que esté basada en una sola métrica, sin que exista una mínima supervisión y una compensación entre los parámetros que sirven para medir la toma de decisiones.

En el próximo epígrafe me referiré precisamente a la transparencia algorítmica y a la necesaria garantía de la reserva y el control de humanidad en materia de libertades y derechos fundamentales.

# LA RESERVA DE HUMANIDAD Y EL CONTROL DE LOS ALGORITMOS EN MATERIA DE LIBERTADES Y DERECHOS FUNDAMENTALES

En la doctrina iuspublicista española más reciente ha alcanzado especial fortuna la expresión "reserva de humanidad", una idea comparable a otros conceptos normativos afines como el de "reserva de ley" (por ejemplo, la reserva de ley orgánica que establece el art. 81.1 CE para el desarrollo de los derechos fundamentales y de las libertades públicas), o el término "reserva de ejercicio de potestades para los funcionarios", según se prevé en el art. 9.2 del Real Decreto Legislativo 5/2015, de 30 de octubre, que considera que las potestades administrativas sólo deben ser ejercidas por los empleados públicos, debido a su relación estatutaria con la Administración, actúan al servicio los intereses generales de acuerdo con los principios de imparcialidad y objetividad.

En el ámbito reglamentario de la UE también se regula la reserva de humanidad por el Reglamento 2016/679 del Parlamento Europeo y del Consejo 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 (UE RGPD), en cuyo art. 22.1 se garantiza a quienes estén interesados el derecho a no ser objetos de decisiones individuales automatizadas, incluida la elaboración de perfiles, que produzca en ellos efectos jurídicos o les afecten significativamente. Por otra parte, en el art. 22.3 UE RGPD se introduce el control de humanidad (la supervisión humana) sobre la decisión automática para salvaguardar los derechos, las libertades y los intereses legítimos de los interesados.

Sin embargo, como ha señalado Juli Ponce, la inminente entrada en vigor del Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de Inteligencia Artificial (en adelante RUEIA) plantea problemas de coherencia normativa en la interpretación del art. 22 UE RGPD,

que establece una reserva de humanidad *generalizada*, salvo las excepciones previstas por la norma jurídica y su sustitución por la supervisión humana, y los arts. 5 y 14 RUEIA, que establecen, respectivamente, la prohibición de la IA *solo* en ciertos casos y la supervisión humana *únicamente* en supuestos de sistemas de IA de alto riesgo; por lo demás, en el resto de usos de la IA que no implican alto riesgo, el RUEIA no prevé ni reserva de humanidad ni supervisión humana (Ponce Solé, 2022, 66).

En el décimo punto de la Exposición de Motivos del RUEIA se indica que los Estados miembros deben estar facultados para especificar la aplicación de las disposiciones del RUEIA a través de su propia normativa nacional, para lo cual el Reglamento concede a los Estados un margen de maniobra en la concreción de dicha normativa interna, incluso en el tratamiento de datos personales ("datos sensibles").

También en el párrafo 16 de la Resolución del Parlamento Europeo, de 3 de mayo de 2022, sobre Inteligencia Artificial en la era digital se encomienda a los legisladores el deber de "abordar los riesgos que plantea actualmente la toma de decisiones basada en la IA", porque, como corresponde a un Estado de Derecho y a los principios proclamados en el art. 9.3 CE que le son consustanciales, el uso de la IA por las administraciones públicas no escapa al cumplimiento de los principios de legalidad, publicidad de las normas, seguridad jurídica, responsabilidad e interdicción de la arbitrariedad de los poderes públicos (Presno Linera, 2022, 51).

En este sentido, en relación con la legislación española sobre IA y la toma de decisiones administrativas, en el marco de la Estrategia Nacional de la Inteligencia Artificial, de la Carta de Derechos Digitales, y de las iniciativas europeas en torno a la IA, el art. 23 de la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación, aparte de ser considerada por algunos juristas como la primera regulación de la IA en Derecho español, establece los principios de buena administración y diligencia debida como

límites de las decisiones administrativas automatizadas (Ponce Solé, 2019a; 2019b).

De conformidad con este estándar jurídico, las administraciones públicas tendrán en cuenta criterios de minimización de sesgos, transparencia y rendición de cuentas, y, siguiendo las recomendaciones de la UE, promoverán "el uso de una IA ética, confiable y respetuosa con los derechos fundamentales". A hilo de este razonamiento sobre la pertinencia de la justificación o explicación de las decisiones públicas, y considerando los riesgos derivados del funcionamiento técnico de los algoritmos y de la IA empleados por la administraciones públicas, comenta Andrés Boix Palop que un mínimo de explicabilidad de los algoritmos es exigible, aún en el caso de que entendamos funcionalmente a los algoritmos como actuaciones administrativas y no como normas reglamentarias, de conformidad con lo establecido en los arts. 9.3 CE y 35 Ley 39/2015, de Procedimiento Administrativo Común (Boix Palop, 2022, 99).

La mención a la necesidad de avanzar en la rendición de cuentas de los programas informáticos con capacidad para tomar decisiones algorítmicas resulta oportuna, no solo en aras de la protección de los derechos y libertades de los ciudadanos, tanto en el ámbito de lo público como en el de lo privado, sino también porque, como ha advertido Manuel Medina Guerrero, a pesar de las innegables ventajas del empleo de sistemas de decisiones automatizadas basadas en algoritmos en términos de eficacia y economía, no pueden soslayarse los riesgos que conllevan estas decisiones para los derechos e intereses de la ciudadanía,

> riesgos que se condensan en el temor genérico a que el ser humano se convierta en un mero objeto de los programas informáticos (Medina Guerrero, 2022, 141).

Algunos autores plantean ya la conveniencia de amparar por vía constitucional el derecho de los ciudadanos al conocimiento de los

algoritmos utilizados por las administraciones en la toma de decisiones automatizadas en el sector público. La justificación de la ampliación del reconocimiento constitucional a este pretendido derecho se motivaría por vía del art. 18.4 CE, en el que se consagran como derechos fundamentales la libertad informática y la protección de datos personales. Por cierto, esta interpretación abierta del art. 18.4 para integrar el derecho a conocer los algoritmos usados en la toma de decisiones parece haber encontrado receptividad en el legislador a través del derecho de acceso a la información pública regulado en la Ley 19/2013, de 9 de diciembre, y en la jurisprudencia constitucional reciente (STC 292/2020) a propósito de las facultades y poderes jurídicos que confiere al titular del derecho a la protección de datos (Cerrillo i Martínez, 2021, 41-78; Medina Guerrero, 2022, 142).

## INTELIGENCIA ARTIFICIAL Y EL COTO VEDADO DE LA HUMANIDAD

En la toma de decisiones automatizadas hay una dimensión fundamental que debe tenerse en cuenta en la valoración del modelo de interacción entre humanos y máquinas, y es la perspectiva ético-jurídica desde la que se estudia el impacto de la cibernética, la informática, la IA y la robótica en el ámbito de los principios y los valores, un terreno de común interés para la Ética de la Inteligencia Artificial y la Filosofía del Derecho.

El pionero de los estudios de informática jurídica e iuscibernética ha sido, precisamente, un iusfilósofo: Mario G. Losano, quien en uno de sus primeros libros, *Giuscibernetica* (1969), analizaba la estructura interdisciplinar de la iuscibernética para ofrecer una síntesis adecuada de los principales problemas teóricos y prácticos que se encontraban los programadores y los juristas en el uso de los programas informáticos (problemas de índole sociológico, filosófico, lógico, lingüístico y técnico).

Por cuanto respecta a la idea de reserva de humanidad fue apuntada por primera vez, en sentido iusfilosófico, por Ernesto Garzón Valdés. En efecto, en uno de los primeros números de la revista *Doxa. Cuadernos de Filosofía del derecho* el iusfilósofo argentino publicó dos artículos; en el primero de ellos, titulado "Representación y democracia", el iusfilósofo argentino se refiere por primera vez al "coto vedado" de los bienes básicos, que son aquellos que son condición necesaria para la realización de cualquier plan de vida (Garzón Valdés, 1989b, 209), y justifica el ejercicio del paternalismo jurídico por parte del Estado de derecho en el caso de que los miembros de la comunidad no comprendan la importancia de estos bienes básicos.

Por lo tanto, según Garzón Valdés, el coto vedado de los bienes básicos no dependería, en última instancia, de la voluntad o deseos de la comunidad, porque, en su opinión:

> Quien no comprende la relevancia de los bienes básicos puede ser incluido en la categoría de incompetente básico (Garzón Valdés, 1989a, 157).

Para este autor, el rechazo de la garantía de los propios bienes básicos supone una muestra evidente de irracionalidad o de ignorancia de relaciones causales elementales como las que existen entre la disponibilidad de estos bienes y la realización de cualquier plan de vida.

Por otra parte, señala Garzón Valdés, los derechos y necesidades básicas incluidos en el coto vedado deben ser universalizables e iguales para todos los ciudadanos, pues sólo se puede concluir que una sociedad es homogénea cuando todos sus miembros gozan de los derechos incluidos en el coto vedado de los bienes básicos. Desde un punto de vista jurídico-positivo, el coto vedado de los intereses universalizables o derechos humanos no pueden ser objeto de recortes resultantes de negociaciones parlamentarias, en la medida que éstos constituyen

> el núcleo no negociable de una constitución democrático-liberal que propician el Estado social (Garzón Valdés, 1989a, 162).

Existe una vinculación conceptual entre el coto vedado de los derechos, las necesidades y bienes básicos, y la democracia representativa, en la que está vigente el principio de la mayoría, dentro de un marco de homogeneidad social compatible con la realización de la esperanza de la autodeterminación individual. En definitiva, como sostiene Carlos Santiago Nino, en sintonía con Ernesto Garzón Valdés, el concepto de bienes y necesidades básicas

> no sólo sería central en una concepción liberal de la sociedad, sino que haría de puente -al permitir su satisfacción simultánea- entre las dos ideas básicas del liberalismo: la de que los fines de los individuos deben ser respetados y la de que todo individuo es un fin en sí mismo (Nino, 1989, 34).

Garzón Valdés se refiere al coto vedado de los derechos, bienes y necesidades en clave liberal y neokantiana. En este sentido, conviene aclarar que su aproximación al concepto de bien básico desde un punto de vista moral recuerda a la idea del bien primario en la posición original rawlsiana e implica la ejecución exitosa de un plan de vida racional. Esta concepción de los bienes básicos (*primary goods*) permite a los individuos presumir la consecución de sus fines: libertad y oportunidad, renta y patrimonio, y sobre todo, el respeto de la persona por sí misma (s*elf-respect*) (Rawls, 1971, 433).

En línea de continuidad con esta lectura kantiana y liberal de los derechos fundamentales (en forma de lista de capacidades) Martha C. Nussbaum también es partidaria de un "blindaje" de los mismos, por ejemplo, ante cualquier intento de recorte o frente a posibles cambios que pudieran ser decididos por mayoría simple

en un nuevo proceso constituyente (Nussbaum, 2017, 69; 2012, 97-98).

Proyectando la tesis iusfilosófica del coto vedado defendida por Garzón Valdés al contorno ético-jurídico demarcado por la reserva de humanidad, en la que se protegen derechos, bienes y necesidades básicas de los ciudadanos frente la indeseada intromisión de las decisiones automatizadas basadas en algoritmos de alto riesgo, cabría preguntarse qué áreas prioritarias de los derechos y libertades serían aquellas en las que operaría este ámbito de protección de lo que es esencialmente humano, absolutamente irreductible e irrenunciable para la dignidad de la persona y la conciencia humana, ante la irrupción de la IA, la robótica y las tecnologías conexas.

En el art. 14 de la 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[1], se regula la vigilancia humana de los sistemas de IA de alto riesgo, que -según la definición del texto normativo- son aquellos que comprometen la salud, la seguridad o los derechos fundamentales, y que pueden surgir

> cuando un sistema de IA de alto riesgo se utiliza conforme a su finalidad prevista o cuando se le da un uso indebido razonablemente previsible (art. 14.2).

En los últimos años ha habido casos paradigmáticos y problemáticos derivados del uso de algoritmos de alto riesgo que han afectado precisamente a las tres grandes áreas a las que se refiere el mencionado precepto de la propuesta de Reglamento UE

[1] Texto de la Comisión Europea COM (2021) 206 final, publicado el 21 de abril de 2021, 2021/0106 (COD).

de IA. Veamos un ejemplo representativo por cada una de estas áreas:

En relación con la seguridad: el miércoles 19 de diciembre de 2018, a las 21:03 hrs, el aeropuerto de Gatwick se vio obligado a cerrar sus vuelos de entrada y salida durante 32 hrs debido al avistamiento cerca del aeródromo de varios drones. Más de 120.000 personas se vieron afectadas por este cierre en el segundo aeropuerto londinense (Shackle, 2020). Tan solo unas semanas más tarde, el 8 de enero de 2019, el aeropuerto de Heathrow, el primero en volumen de tráfico aéreo del Reino Unido, también suspendió durante una hora sus vuelos tras detectar la torre de control, a las 17.12 hrs, un dron sobrevolando su espacio aéreo. Desde un punto de vista legal y judicial estos incidentes generaron cambios (el 31 de diciembre de 2020 entró en vigor una nueva normativa para propietarios de drones, clasificando tipos de drones en diversas categorías, introduciendo requisitos de registro y seguro para cada clase). Por otra parte, se suscitó la cuestión a propósito de las reclamaciones de indemnización a las compañías de seguros por parte de los viajeros, y determinar sobre quién recaería la responsabilidad civil en ambos incidentes, se presentaría también la ocasión de plantearse la oportunidad de reconocer la personalidad electrónica de los robots o máquinas inteligentes entrenadas con IA para responder por los perjuicios causados a terceros por su mal funcionamiento o por sus erróneas tomas de decisión automatizadas (Benon-Decker, 2021).

Respecto a la reserva de humanidad y al control de las decisiones administrativas basadas en algoritmos predictivos cuando limitan el disfrute de los derechos fundamentales, en el ámbito de la educación (art. 27 CE) resulta ilustrativo el caso de los exámenes de acceso a las universidades británicas en el verano de 2020, el primero de la pandemia del Covid-19, cuyos resultados no se obtuvieron a partir de una prueba presencial, sino excepcionalmente mediante un algoritmo que combinó la calificación obtenida por los alumnos en sus colegios o institutos en los exámenes realizados en los tres últimos cursos y una

calificación estimada de cada alumno en comparación con sus compañeros de su centro.

El resultado de las predicciones del algoritmo utilizado para determinar las notas GCSE y A-Levels de los alumnos de Inglaterra, Gales y Escocia fue, anunciado el 13 de agosto, fue decepcionante: las calificaciones asignadas por el algoritmo fueron, en general, inferiores a las previsiones realizadas por los profesores de los colegios e institutos; por otra parte, se apreció también en la decisión automatizada un sesgo discriminatorio hacia los alumnos de las escuelas públicas situadas en zonas humildes y que, por el contrario, favorecía a los estudiantes de colegios privados o centros públicos de alto rendimiento. A raíz de la polémica causada por la prueba de acceso al sistema universitario británico mediante el empleo de un algoritmo predictivo, el gobierno anunció el 17 de agosto que los resultados del nivel A se modificarían para introducir en ellos como criterio corrector las estimaciones originalmente realizadas por los profesores de los centros de procedencia de cada alumno (por cierto, como ha podido apreciarse, este caso encajaría en el segundo modelo de control de humanidad: HITLFE).

A propósito de los algoritmos predictivos, se pregunta Alejandro Huergo Lora si en nuestra experiencia jurídica se pueden utilizar para determinar el contenido de resoluciones administrativas, al igual que sucediera con los A-Levels en 2020. En este sentido, el catedrático de Derecho Administrativo de la Universidad de Oviedo establece la siguiente diferenciación:

> En el caso de potestades discrecionales, en las que la norma no vincula el contenido de la decisión administrativa a la constatación de uno o varios hechos, sino que establece un marco dentro del que es válida cualquier decisión que tome la Administración siempre que esté adecuadamente motivada y se hayan respetado las normas procedimentales, es perfectamente posible y válido que la

> Administración utilice, entre otros factores, predicciones algorítmicas (Huergo Lora, 2020, 78-79).

Ahora bien, matiza Huergo Lora, no se puede soslayar que el hecho de que una decisión administrativa se constante mediante un modelo algorítmico predictivo plantea problemas de principio (en la medida en que, a su parecer, resultaría incoherente que se pueda decidir en función de predicciones y no de hechos probados) y problemas de orden práctico (¿cómo se pueden tomar decisiones a partir de algoritmos dada su opacidad?).

Teniendo en cuenta que el riesgo de error en la predicción algorítmica es aún muy alto, y que, dadas las circunstancias de error y opacidad de los algoritmos, el empleo de dichas predicciones en las resoluciones administrativas sería incompatible con las ideas de dignidad de la persona y del libre desarrollo de la personalidad, porque dicha predicción se basaría en la presunción de que los comportamientos que en el pasado han tenido otras personas en circunstancias similares a las del interesado que invoca ahora la decisión automatizada de la Administración por vía de predicción algorítmica, permitirían predecir con seguridad la conducta futura de ese interesado (Huergo Lora, 2020, 79).

Por último, respecto a la problemática que se origina a partir de la aplicación de la los sistemas expertos de IA y los algoritmos en el ámbito de la salud, conjugando su empleo con el marco legal que protege los derechos de los pacientes (consentimiento informado, privacidad y confidencialidad de los datos médicos, acceso al historial clínico, respeto de la voluntad anticipada por el paciente respecto a los cuidados y tratamientos de salud que desea recibir, derecho a presentar reclamaciones y sugerencias[2]).

[2] En España existe un amplio marco legal regulador de los derechos y obligaciones de los pacientes conformado por la siguiente normativa: el art. 43 CE, donde se reconoce el derecho a la protección de la salud; la Ley 14/1986, de 25 de abril, en cuyo art. 10. en el que se recogen los derechos de los pacientes con respecto a

En los últimos años se ha producido un aumento significativo del interés por las aplicaciones de los algoritmos de aprendizaje automático (*machine learning algorithms*) para la toma de decisiones médicas automatizadas. En los sistemas expertos aplicados a la diagnosis médica funcionan de la misma forma que el resto de sistemas expertos de la IA clásica, en la medida que contienen una base de datos de reglas deductivas mediante las cuales, a partir de un conjunto de hechos conocidos, se pueden inferir determinadas consecuencias. Concretamente, en medicina, la presencia de determinados síntomas puede hacer que el sistema experto los asocie a determinadas enfermedades relacionadas con dichos síntomas y sugiera realizar pruebas diagnósticas para llegar a conclusiones inequívocas.

Los sistemas de IA empleados en las diagnosis médicas están basados en conocimientos codificados de los expertos, como el sistema IBM Watson Oncology para la diagnosis y el tratamiento oncológico, un sistema de IA de metaanálisis que combina la extracción automática de textos de documentos clínicos con un número ingente de reglas lógicas. En el diagnóstico médico basado en imágenes, un sistema experto como éste es capaz de buscar características definidas por los expertos y codificar explícitamente las reglas de decisión definidas por los médicos (Jie-Zhiying-Li, 2021).

En suma, los algoritmos de aprendizaje automático han transformado para bien la atención sanitaria profesional, sobre todo si se compara el porcentaje de acierto y precisión de los diagnósticos realizados por los sistemas expertos de IA con los diagnósticos realizados por médicos especializados. En este sentido, según un estudio publicado en 2015 por la *National*

---

las distintas administraciones públicas sanitarias; la Ley Básica Reguladora de la Autonomía del Paciente y de Derechos y Obligaciones en Materia de Información y Documentación Clínica (Ley 41/2002, de 14 de noviembre); la Ley 16/2003, de 28 de mayo, de Cohesión y Calidad del Sistema Nacional de Salud, y el Real Decreto-Ley 7/2018, de 27 de julio, sobre el acceso universal al Sistema Nacional de Salud.

*Academy of Sciences*, alrededor de un 5% de los diagnósticos médicos realizados a adultos en centros hospitalarios estadounidenses fueron erróneos; a su vez, se estimaba que un 10% de los fallecimientos de los pacientes se debieron precisamente a errores de diagnóstico médico[3].

En descargo de los responsables de los diagnósticos médicos realizados por clínicos humanos hay que remarcar el hecho de que, a diferencia de los sistemas expertos de IA, los clínicos, en cuanto humanos, son conscientes de su falibilidad y de la responsabilidad que asumen ante sus pacientes; por el contrario, existe una presunción favorable hacia la diagnosis mecánica realizada a través del algoritmo de aprendizaje automático, en la medida que ofrece mayor certeza y precisión (aunque en realidad haya también un margen de error en los algoritmos de aprendizaje automático). Pero la diagnosis médica plantea un problema aún mayor que el del cálculo meramente estadístico y cuantificativo de la exactitud de los resultados de las diagnosis clínicas: se trata de un problema epistémico-ético.

En un artículo reciente sobre la ética de la toma de decisiones algorítmica en la atención sanitaria sus autores, Thomas Grote y Philipp Berens, reconocen la utilidad de los algoritmos de aprendizaje automático para mejorar la capacidad de decisión de los clínicos al aportarles una fuente adicional de información y pruebas complementarias que les ayudarán a tomar sin duda una mejor decisión médica.

Sin embargo, los clínicos se enfrentan a un serio obstáculo cada vez que intentan inferir información a partir de los resultados de un algoritmo de aprendizaje automático. El problema subyacente puede describirse del siguiente modo: tanto el médico como el algoritmo de aprendizaje automático pueden considerarse expertos. Sin embargo, han recibido una formación diferente y

[3] Cfr., *National Academies of Sciences*, "Engineering, and Medicine. Improving Diagnosis in Health Care", The National Academies Press, Washington DC, 2015.

razonan de formas muy distintas. Para el clínico, esta diversidad de formación y razonamiento plantea serios inconvenientes deontológicos y epistemológicos siempre que se produce un desacuerdo entre colegas o pares médicos considerados iguales (incluso cuando se reconoce como competentes a los algoritmos en igualdad de condiciones que sus "pares médicos").

Por lo tanto, en primera instancia, en el caso de "desacuerdo entre iguales" de dos colegas o pares se produciría una discrepancia en la diagnosis de los síntomas, la historia clínica y el examen físico del paciente. Esta discrepancia entre "pares" (clínico y algorítmico) podría inclinarse a favor del algoritmo de aprendizaje automático por simple desistimiento del médico para eludir ulteriores responsabilidades derivadas de una toma de decisión errónea; con lo cual, desde un punto de vista epistémico y ético, además de fomentarse entre los médicos el dogmatismo y la credulidad cientificista-tecnologicista, terminaría imponiéndose, antes que el mejor criterio médico, un simple mecanismo pragmático de "medicina defensiva" que protegería a los clínicos de la posible rendición de cuentas a sus pacientes (Grote-Berens, 2020, 208).

A veces la presunción de perfectibilidad e inexorabilidad del diagnóstico determinado por el algoritmo de aprendizaje autonómico puede ser la causa del desistimiento del control de humanidad por parte de los médicos respecto al funcionamiento correcto del sistema de IA aplicado al ámbito de la salud. En este sentido conviene no rebajar la exigencia del enfoque ético de la IA, pues de producirse esa renuncia por parte de la comunidad médica a la supervisión de las tomas de decisión automatizadas, se estaría abandonando (y sacrificando) injustificadamente el primer modelo de gestión para la interacción entre los humanos y las máquinas inteligentes (HITL) en el que el ser humano conserva el control sobre el sistema algorítmico, en beneficio del modelo antagónico (HOTL), en el que la decisión médica se confiaría, en última instancia, a las máquinas mientras que los clínicos se inhiben de participar en la misma para no asumir responsabilidades. Esta renuncia supondría, en última instancia, una rendición por parte los

profesionales de la medicina al mito de la perfección de la IA, además de su efectiva rendición a la falacia tecnológica inspirada por el fundamentalismo tecnológico, que no puede confundirse con la metodología empírico, la capacidad crítica y la vocación humanista que caracteriza a la ciencia médica.

En el próximo apartado, el conclusivo, se abordará precisamente esta cuestión relativa a la necesidad de supervisión ética de los algoritmos y la exigencia de responsabilidad derivada del funcionamiento erróneo o del mal uso de que pueda hacerse de los mismos.

## CONCLUSIÓN

En 2016 Google presentó el proyecto DeepMind Health, en colaboración con el Royal Free London NHS Foundation Trust. En este proyecto de la multinacional tecnológica estadounidense puso en funcionamiento dos aplicaciones: Streams y Hark, con las que pretendía mejorar los sistemas de salud (en particular el NHS, es decir, el servicio de salud británico). Así, mientras que la app Streams servía para detectar pacientes en riesgo de contraer alguna enfermedad y se reducía a unos segundos el proceso de revisión de las pruebas analíticas de sangre de los pacientes, la app Hark estaba configurada para reducir las listas de espera y el papeleo en los hospitales. Al inicio de esta colaboración de Google con el NHS, los responsables de DeepMind Health se comprometieron a que los datos de los pacientes (registros personales de 1,6 millones de pacientes) nunca se conectarían con las cuentas o servicios de Google, precisamente en aras del respeto a su derecho a la intimidad (Powles-Hodson, 2017, 351-367).

En 2019 Google rompería esta promesa inicial al vincular DeepMind Health, la empresa filial de atención médica a Google, la empresa matriz, incorporando a la misma millones de datos personales de los pacientes del NHS sin contar con su previo consentimiento, a la vez que desmantelaba el comité de revisión

ética de los algoritmos aplicados en el ámbito sanitario y cuyo panel de revisores ya había manifestado pocos meses antes del cierre del mismo su preocupación ante el riesgo potencial de que DeepMind Health pudiera utilizar su acceso a los datos personales de los pacientes del NHS para propiciar las ganancias monopolísticas para la empresa matriz de Google, Alphabet (Murgia, 2018).

A propósito de la revisión institucional de la transparencia y buen funcionamiento de los algoritmos, hay autores que ponen en duda que un solo comité ético sea capaz por sí mismo de garantizar el rigor, la utilidad y la integridad de los *big data*. Se trataría, según esta doctrina escéptica respecto a la viabilidad del control de humanidad sobre la IA, de una pretensión poco realista (Lipworth-Mason-Kerridge-Ioannidis, 2017, 489-500). Otra línea doctrinal sostiene que el problema de la responsabilidad y de la pública rendición de cuentas por parte de quienes hacen uso de los algoritmos potencialmente lesivos de derechos y libertades de los ciudadanos debería ser planteada como una cuestión de acuerdo o consenso entre las partes interesadas (Floridi, 2022, 173).

Como solución alternativa a la posible falta de acuerdo sobre la revisión de la transparencia y el buen funcionamiento de los algoritmos, existe un código ético propuesto por la *Association for Computing Machinery* (ACM) en el que se enuncian siete principios éticos fundamentales para los profesionales de la Informática y la Computación[4]. El objetivo principal que se

[4] Los siete principios éticos generales enunciados por el Código de Ética y Conducta Profesional de la ACM y que todo buen profesional de la Informática y la Computación debería seguir son los siguientes: 1.- Contribuir a la sociedad y al bienestar humano, reconociendo que todas las personas son partes interesadas en la Informática; 2.- Evitar daños o consecuencias negativas, especialmente cuando son significativas e injustas; 3.- La conducta de un buen profesional debe ser siempre transparente, honesta y fiable; 4.- Respeto a los valores de igualdad (no discriminación), tolerancia, respeto a los demás y justicia; 5.- Crear oportunidades para que los miembros de la organización o grupo crezcan como profesionales; 6.- Respetar la intimidad (*privacy*), por lo que los profesionales de

persigue este código deontológico profesional es, precisamente, el beneficio de todas las partes interesadas en el mantenimiento de un debate abierto entorno a las cuestiones éticas que promueven la responsabilidad y la transparencia (Buhmann-Passmann-Fieseler, 2019).

Por último, entre quienes apuestan por generar las condiciones necesarias para el cultivo de una "cultura algorítmica" más responsable, destaca Luciano Floridi, quien propone repartir una "responsabilidad moral distributiva" entre todos los agentes morales, es decir, personas humanas o jurídicas, es decir, sociedades constituidas por seres humanos (Floridi, 2016, 2). En sentido análogo, Mark Coeckelbergh, inspirándose en la Declaración de Montreal para un desarrollo responsable de la IA (2018), sostiene que la espiral de la Ética de la IA no debe circunscribirse solamente al ámbito de la política internacional y la gobernanza global; en efecto, también la comunidad académica debe asumir una posición activa en la propuesta de iniciativas que busquen una interacción equilibrada e inocua entre los humanos y las máquinas desarrolladas con IA, de tal modo que sea posible concebir la armonía entre el avance tecnológico con la salvaguarda del legado humanista y los valores que son inherentes y necesarios para la defensa de la dignidad, la autonomía de la voluntad y los derechos fundamentales del ser humano (Coeckelbergh, 2020, 157).

## BIBLIOGRAFÍA

BALASUBRAMANIAN, R.-LIBARIKIAN, A.-McELHANEY, D. (2021), "Insurance 2030-The Impact of AI on the future of insurance", *McKinsey*

la Informática solo deben usar la información personal para fines legítimos y sin violar los derechos y libertades de los individuos; 7.- Respeto de la confidencialidad de los secretos comerciales, los datos de los clientes, las estrategias comerciales no públicas, información financiera, datos de la investigación, artículos académicos previos a la publicación y solicitudes de patentes. Cfr., https:// www.acm.org/code-of-ethics

*& Company*, March 12, 2021, https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance (última consulta, 25 de abril de 2023)
BENON, H.-DECKER, M. (2021), "Insuring Commercial Drones? Liability or Opportunity? *Insurance Journal*, September 6, 2021, https://www.insurancejournal.com/magazines/mag-features/2021/09/06/630181.htm (última consulta, 25 de abril de 2023)
BUHMANN, A.-PASSMANN, J.-FIESELER, C. (2019), "Managing algorithmic accountability: Balancing reputational concerns, engagement strategies, and the potential of rational discourse", *Journal of Business Ethics*, nº 163, pp. 265-280, htttps://link.springer.com/article/10.1007/s10551-019-04226-4 (última consulta, 25 de abril de 2023)
BOIX PALOP, A. (2022), "Transparencia en la utilización de Inteligencia Artificial por parte de la Administración", *El Cronista del Estado social y democrático de derecho*, nº 100, pp. 90-105.
CERRILLO I MARTÍNEZ, A. (2021), "La transparencia de los algoritmos que utilizan las administraciones públicas", *Anuario de Transparencia Local 3/2020*, Fundación Democracia y Gobierno Local, pp. 41-78.
COECKELBERGH, M. (2020), *AI Ethics*, The MIT Press, Cambridge (Massachusetts)-London.
FITZEK, F. H. P.-LI, S.-Ch.-SPEIDEL, S.-STRUFE, T. (2021), "Tactile Internet with Human-in-the-Loop: New Frontiers of Transdisciplinary Research", *Tactile Internet with Human-in-the-Loop*, Academic Press, San Diego (California).
FLORIDI, L. (2016), "Faultless responsability: On the nature and allovation of moral responsability for distributed moral actions", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374 (2083), https://royalsocietypublishing.org/doi/10.1098/rsta.2016.0112 (última consulta, 25 de abril de 2023)
(2022), *Etica dell´intelligenza artificiale. Sviluppi, opportunità, sfide*, Rafaello Cortina Editore, Milano.
GARZÓN VALDÉS, E. (1989a), "Representación y democracia", *Doxa. Cuadernos de Filosofía del derecho*, nº 6, 1989, pp. 143-164.
(1989b), "Algo más acerca del *coto vedado*", *Doxa. Cuadernos de Filosofía del derecho*, nº 6, 1989, pp. 209-213.
GROTE, T.-BERENS, P. (2019), "On the Ethics of Algorithmic Decision-Making in Healthcare", *Journal of Medical Ethics*, nº 46, 20th November

2019, pp. 205-211, https:// 10.1136/medethics-2019-105586 (última consulta, 25 de abril de 2023)
HABERMAS, J. (1981), *Theorie des Kommunikativen Handelns* (2 Bände), Suhrkamp, Frankfurt am Main.
HERBERT, B.-ANDERSON, K. J. (2011[5]), *Dune. La Yihad butleriana*, trad. esp., Eduardo G. Murillo, Debolsillo, Barcelona.
HUERGO LORA, A. (2020), "Una aproximación a los algoritmos desde el Derecho administrativo" (pp. 23-87), en *La regulación de los algoritmos* (Director: Alejandro Huergo Lora; coordinador: Gustavo Manuel Díaz González), Thomson Reuters Aranzadi, Cizur Menor (Navarra).
JIE, Z.-ZHIYING, Z.-LI, L. (2021), "A Meta-Analysis of Watson for Oncology in Clinical Application", *Nature*, 11:5792, https://www.nature.com/articles/s41598-021-84973-5 (última consulta, 25 de abril de 2023)
LARSON, E. J. (2022), *El mito de la Inteligencia Artificial. Por qué las máquinas no pueden pensar como nosotros lo hacemos*, trad. esp., Milo J. Krmpotić, Shackleton Books, Barcelona.
LIPWORTH, W.-MASON, P. H.-KERRIDGE, I.-IOANNIDIS, J. P. A. (2017), "Ethics and Epistemology in Big Data Research", *Journal of Bioethical Inquiry*, 14 (4), pp. 489-500, https:// link.springer.com/article/10.1007/s11673-017-9771-3
LOSANO, M. G. (1969), *Giuscibernetica. Macchine e modelli cibernetici del diritto*, Einaudi, Torino.
MEDINA GUERRERO, M. (2022), "El derecho a conocer los algoritmos utilizados en la toma de decisiones. Aproximación desde la perspectiva del derecho fundamental a la protección de datos personales", *Teoría y realidad constitucional*, nº 49,
NINO, C. S. (1989), "Autonomía y necesidades básicas", *Doxa. Cuadernos de Filosofía del derecho*, nº 6, 1989, pp. 21-34.
MURGIA, M. (2018), "DeepMind´s move to transfer health unit to Google stirs data fears", *Financial Times*, 13th November, https://www.ft.com/content/f4a73450-e771-11e8-8a85-04b8afea6ea3 (última consulta, 25 de abril de 2023)
NUSSBAUM, M. C. *Frontiers of Justice. Disability, Nationality. Species Membership* (2007), The Belknap Press of Harvard University Press, Cambridge (Massachusetts)-London.
(2012), *Crear capacidades. Propuesta para el desarrollo humano*, trad. esp., A. Santos Mosquera, Paidós, Barcelona.
PONCE SOLÉ, J. (2019a), *La lucha por el derecho a una buena administración: el estándar jurídico de diligencia debida y el buen*

*gobierno en las políticas públicas*, Cuadernos de la Cátedra de Democracia y Derechos Humanos, núm. 15, Universidad de Alcalá de Henares- Defensor del Pueblo, Alcalá de Henares.
(2019b), "Inteligencia artificial, Derecho administrativo y reserva de humanidad, algoritmos y procedimiento administrativo debido tecnológico", Revista General de Derecho Administrativo, nº 50, pp. 141-171.
(2022), "Reserva de humanidad y supervisión humana de la Inteligencia Artificial", *El Cronista del Estado social y democrático de derecho*, nº 100, pp. 58-67.
POWLES, J.-HODSON, H. (2017), "Google DeepMind and Healthcare in Age of Algorithms", *Health and Technology*, 7/2017, pp. 351-367, https://link.springer.com/article/10.1007/s12553-017-0179-1
PRESNO LINERA, M. (2022), "Derechos fundamentales e Inteligencia Artificial en el Estado social, democrático y *digital* de Derecho", *El Cronista del Estado social y democrático de derecho*, nº 100, pp. 48-57.
RAO, D. M.- CHERNYAKHOVSKY, A.-RAO, V. (2011), "Analyzing Global Epidemiology of Diseases Using Human-in-the-Loop Bio-Simulations" (153-174), *Human-in-the-Loop Simulations* (eds. L. Rothrock-S. Narayanan), Springer Verlag, London.
RAWLS, J. (1971), *A Theory of Justice*, The Belknap Press of Harvard University Press, Cambridge (Massachusetts)-London.
ROSS, M.-TAYLOR, J. (2021), "Managing AI Decision-Making Tools", *Harvard Business Review*, https://hbr.org/2021/11/managing-ai-decision-making-tools (última consulta, 25 de abril de 2023)
SHACKLE, S. (2020), "The Mistery of the Gatwick Drone", *The Guardian*, 1 December 2020, https://www.theguardian.com/uk-news/2020/dec/01/the-mystery-of-the-gatwick-drone (última consulta, 25 de abril de 2023)
ŽIŽEK, Slavoj (2023), *Hegel y el cerebro conectado*, trad. esp., Fernando Borrajo, Paidós, Barcelona.