Cuadernos de la
Cátedra CaixaBank de
Responsabilidad Social
Corporativa

Nº 42
Septiembre de 2019

Ética e inteligencia artificial

Sergio Marín García

Cátedra CaixaBank de
Responsabilidad Social Corporativa

IESE Business School
University of Navarra

Cátedra CaixaBank
de Responsabilidad
Social Corporativa

Ética e inteligencia artificial

Sergio Marín García

Cátedra CaixaBank de
Responsabilidad Social Corporativa

DOI: https://dx.doi.org/10.15581/018.ST-522

ÍNDICE

1. INTRODUCCIÓN

La inteligencia artificial (IA) y todas las aplicaciones y servicios basados en esta tecnología han recibido una atención considerable por parte de la comunidad científica y empresarial durante los últimos años. Si bien es cierto que los medios de comunicación y otras plataformas de divulgación han contribuido a crear unas expectativas desmedidas respecto a las posibilidades de la IA, también lo es que dicha tecnología se encuentra ya presente en muchas de nuestras actividades cotidianas: desde realizar una búsqueda en Internet hasta escuchar música desde nuestros dispositivos o solicitar un préstamo bancario.

Junto con un optimismo generalizado ante las potenciales aplicaciones de la IA, el desarrollo de esta tecnología también ha estado acompañado, desde el principio, por una preocupación ante los riesgos que el uso de algunas de estas aplicaciones entraña. A primera vista, el diseño de sistemas "inteligentes" capaces de tomar decisiones autónomas introduce la problemática de conferir referentes éticos a estos sistemas. La creación de sistemas autónomos ha motivado investigaciones y estudios sobre el peligro que este tipo de tecnología entraña para la existencia humana[1]. Aunque algunas de estas perspectivas posean un carácter distópico y fatalista, la realidad es que algunas aplicaciones de la IA ya están planteando verdaderos retos éticos. Uno de los casos más discutidos quizá sea el de los coches autónomos. Diseñar un vehículo de este tipo implica dotarlo de un algoritmo lo suficientemente complejo como para hacer frente a situaciones imprevistas en la carretera. Dicho proceso debe tener en cuenta situaciones anómalas, como encontrarse de pronto con un peatón cruzando la calzada, con otro vehículo invadiendo el carril contrario o con un semáforo estropeado. En todas estas situaciones, el vehículo ha de ser capaz de "decidir" qué curso de acción tomar. Esto supone que el diseño del algoritmo debe incluir, de alguna manera, algún patrón de conducta ético: ante una colisión inevitable, ¿qué vida habrá de sacrificar el vehículo: la del conductor o la del peatón que cruza por donde no debe? (Bonnefon, Sharif y Rahwan, 2016). Junto con el problema de la autonomía, veremos también que la IA ha suscitado otros problemas de tipo ético, como los de la privacidad, la seguridad, la selección de datos o la creación de hábitos de consumo.

La IA —como el resto de las tecnologías diseñadas por el ser humano— puede derivar en aplicaciones nocivas o beneficiosas para las personas (Argandoña, 2019). La propia historia nos provee de varios ejemplos: la pólvora fue empleada originalmente en China para confeccionar fuegos artificiales, mientras que su posterior introducción en Europa derivó en la producción de armas de fuego; el conocimiento de la energía nuclear permitió, por un lado, la creación de armas de destrucción masiva y, por otro, el desarrollo de tratamientos médicos que emplean isótopos radioactivos. Ahora bien, que la tecnología se preste a diversos usos no equivale a afirmar que la tecnología en sí misma sea neutral (Martin y Freeman, 2004, p. 356) ni a que todo desarrollo técnico sea aconsejable. Una de las actitudes más frecuentes ante la tecnología es la de considerarla, precisamente, como un fenómeno irrefutablemente positivo. El desarrollo de las posibilidades técnicas supondría siempre —según esta visión— algo encomiable y beneficioso para la sociedad. Tal como veremos, algunas de las aplicaciones de la IA ponen especialmente de manifiesto que el desarrollo tecnológico no es algo inocuo ni constituye siempre un avance, sino que siempre responde a unos objetivos fijados de antemano, es fruto de un proceso de toma de decisiones, ha sido diseñado por personas concretas y, por ello mismo, puede ser perjudicial o entrañar riesgos de diverso tipo.

En el presente cuaderno, pretendemos abordar justamente la relación entre la IA y la ética. En concreto, se tratará de poner en claro por qué el diseño y desarrollo de este tipo de

[1] Véanse, por ejemplo, las investigaciones realizadas por el Future of Humanity Institute (FHI), de la University of Oxford, acerca de la seguridad y el gobierno en el marco de la IA: www.fhi.ox.ac.uk.

tecnología está sujeto también al razonamiento ético. De entrada, esto plantea una serie de retos. El primero de ellos es el de establecer con precisión en qué consiste la IA, en concreto, qué se entiende por "inteligencia" y de qué manera esta puede estar presente de manera artificial en un dispositivo, vehículo o asistente virtual. De la puesta en claro de este punto se derivan consecuencias importantes para el análisis ético: según el grado de inteligencia y autonomía que quepa adscribir a estos dispositivos, el diseño de estos sistemas tendrá aparejada una mayor o menor responsabilidad. El segundo reto —estrechamente relacionado con el anterior— es el de esclarecer de forma reconocible el proceso que abarca, desde el diseño de un algoritmo hasta la decisión "autónoma" que un sistema pueda tomar. Sin tener una idea clara de cómo se desarrolla este proceso, la adscripción de responsabilidades y obligaciones puede resultar sumamente complicada. A esto se le añade el hecho de que, por su propio diseño, algunos de estos sistemas inteligentes han sido concebidos para mejorar continuamente la toma de decisiones, pero no siempre es posible comprender el proceso que internamente ha seguido el algoritmo para dar con nuevas alternativas. Cabe añadir además que, sin un conocimiento bastante especializado en disciplinas como la computación, la teoría de la información o las matemáticas, estos procesos se vuelven prácticamente incomprensibles para el observador inexperto.

[...] según el grado de inteligencia y autonomía que quepa adscribir a estos dispositivos, el diseño de estos sistemas tendrá aparejada una mayor o menor responsabilidad.

Hemos dividido la estructura del cuaderno de la siguiente manera: en el primer apartado, tratamos de forma breve la aparición y el desarrollo de la IA a mediados del siglo pasado. El contenido de esta sección será de gran utilidad para adquirir una visión más precisa y realista que la que a menudo se difunde en la prensa acerca de la IA; en el segundo apartado, repasamos las diversas aplicaciones que la IA ha encontrado en diversas industrias y sectores, y señalaremos para cada uno de ellos algunos de los problemas éticos que salen a relucir; por último, en el tercer apartado, recogemos de forma sintética cuáles son los principales desafíos de la IA para la ética y señalamos algunos principios que pueden guiar el avance y desarrollo de esta tecnología en el futuro.

2. LA INTELIGENCIA ARTIFICIAL

A día de hoy, la IA se ha convertido en uno de esos fenómenos de los que todo el mundo ha oído hablar, pero del que pocos saben exactamente en qué consiste. A grandes rasgos, la IA se percibe como cierta capacidad o potencia de computación que permitiría crear sistemas y dispositivos dotados de las mismas capacidades cognitivas que los seres humanos (Charniak y McDermott, 1985). A esta visión algo simplificada la acompañan también unas expectativas algo desmedidas acerca de lo que esta tecnología puede llegar a lograr.

A este desconocimiento se añade, además, un fenómeno característico de la IA, el llamado "efecto IA", que se produce cada vez que un nuevo avance en el desarrollo de un programa o dispositivo es descartado por la opinión pública bajo el argumento de que no es inteligente: "Cada vez que alguien descubre cómo hacer que un equipo haga algo [...], existe un coro de críticos que dicen: 'Eso no es pensar'". (McCorduck, 2004, p. 204). Tal es el estado de la cuestión que la IA ha llegado a definirse como "todo aquello que no se haya logrado todavía" (Hofstadter, 1979, p. 601). Semejante actitud entraña el riesgo de perder de vista una definición rigurosa de este tipo de inteligencia, así como de los avances que este campo ha presenciado durante sus casi ya setenta años de vida.

2.1. ORÍGENES DE LA INTELIGENCIA ARTIFICIAL

La IA vio la luz como campo de investigación el verano de 1956 en la llamada Conferencia de Dartmouth (Russell y Norvig, 2016, p. 17). Por aquel entonces, John McCarthy, formado en la Princeton University y en la Stanford University, decidió reunir en el Dartmouth College a los principales investigadores estadounidenses del campo de la informática y

La premisa fundamental de este seminario era la creencia de que "cualquier aspecto del aprendizaje y elemento de la inteligencia puede, en principio, ser descrito de manera tan precisa que sea posible crear una máquina que lo emule".

de la psicología cognitiva para trabajar durante los meses de verano en lo que McCarthy acuñó como "*artificial intelligence*". La premisa fundamental de este seminario era la creencia de que "cualquier aspecto del aprendizaje y elemento de la inteligencia puede, en principio, ser descrito de manera tan precisa que sea posible crear una máquina que lo emule" (McCarthy, Minsky, Rochester y Shannon, 1955). El listado de temas de trabajo del seminario (redes neuronales, arbitrariedad y creatividad, ordenadores automáticos) pone de manifiesto que, en la mente de los organizadores, la IA no era una rama más de la informática, de la teoría de la decisión o de la lógica: se trataba expresamente de una nueva disciplina centrada en crear máquinas capaces de replicar la capacidad humana de emplear el lenguaje, de aprender y razonar creativamente.

A la conferencia de 1956 la siguieron décadas de optimismo generalizado respecto a los potenciales avances y aplicaciones de la IA. Herbert Simon, uno de los investigadores del Carnegie Mellon University que asistió a la Conferencia de Dartmouth, afirmó que "en veinte años las máquinas serán capaces de llevar a cabo cualquier tipo de trabajo que un hombre pueda hacer" (Simon, 1965, p. 96). Sin embargo, las dos primeras décadas de investigación no condujeron a avances significativos y, hacia 1974, los fondos destinados a financiar las investigaciones se redujeron drásticamente (Crevier, 1993, p. 115).

La IA volvió a recibir una atención considerable a partir de la década de los ochenta, principalmente a raíz del éxito comercial de los "sistemas expertos" (Russell y Norvig, 2016, pp. 22-24). Se trataba de distintos programas inteligentes que emulaban las capacidades analíticas y de toma de decisiones de los seres humanos. Algunos de estos programas comenzaron a ser empleados en el ámbito empresarial para procesar órdenes y pedidos. El programa R1, por ejemplo, llegó a suponer a la Digital Equipment Corporation un ahorro anual de cuarenta millones de dólares (McDermott, 1982). A finales de la década de los noventa del pasado siglo y comienzos del XXI, la IA comenzó a ser empleada para labores de logística, minería de datos o diagnósticos médicos. Para entonces, la mayor capacidad de computación y las sinergias entre la IA y otras disciplinas como la economía, la estadística o las matemáticas, condujeron a la expansión de esta tecnología en un diverso número de industrias (Russell y Norvig, 2016, pp. 25-26). Tal como tendremos ocasión de ver, a día de hoy, la IA es una tecnología ampliamente extendida en muchas industrias, con aplicaciones en una gran variedad de ámbitos.

2.2. DEFINICIONES DE INTELIGENCIA ARTIFICIAL

Hemos visto de forma somera cuál ha sido el desarrollo de la IA desde su nacimiento hasta el día de hoy. Junto con esta visión histórica, se hace necesaria también una aproximación más teórica a la misma que nos permita establecer una definición y delimitar las competencias de esta tecnología.

Es preciso advertir, en primer lugar, que no existe una definición consensuada de IA. Cada definición parece asumir unos objetivos distintos para la IA o parte de concepciones distintas de qué pueda significar el término "inteligencia". En un primer intento de concreción, podríamos definir la IA como una combinación de algoritmos planteados para crear máquinas con las mismas capacidades que el ser humano (Iberdrola, 2019). Dicho lo cual, cabría matizar aún más: ¿en qué consisten dichos algoritmos y de qué manera se diseñan?, ¿en qué tipo de soporte material se insertan?, ¿qué capacidades específicas del ser humano se pretende replicar en las máquinas?

Si repasamos alguno de los textos fundacionales de la IA, encontramos este mismo grado de indefinición en algunas de las pretensiones de aquellos años. La premisa fundamental del manifiesto de Dartmouth era que la IA, en un corto espacio de tiempo, iba a ser capaz de sintetizar los procesos cognitivos y de lograr la inteligencia general en máquinas.

Pero, a lo largo de estos textos, no se encuentran apuntes más precisos acerca de la estructura de la inteligencia, su funcionamiento o su traslación en sistemas informáticos. Y, a pesar de lo indefinido, es esta premisa la que mejor refleja la naturaleza de la IA. Esta tecnología busca, por un lado, descomponer la inteligencia humana en sus procesos más simples y elementales para poder, más tarde, formalizarlos —esto es, expresarlos en el lenguaje formal de la lógica—, recomponerlos en forma de algoritmos y estrategias de programación, y reproducir así la inteligencia humana en máquinas y programas. Apoyándose en la afirmación de Leibniz[2], a saber, que "todo lo que sepamos describir de forma clara, completa, precisa e inequívoca es computable" (Leibniz, 1923, p. 174), la IA busca precisamente describir la inteligencia humana en sus elementos más simples para poder así hacerla computable y transferible mediante algoritmos y lenguajes de programación.

Con todo, la premisa del manifiesto de Dartmouth no proporciona una visión más específica acerca del alcance de la IA. Por este motivo, distintos autores han tratado de aportar una definición más concisa de la esta. Para ello, tratan de establecer, en primer lugar, qué objetivo persigue esta tecnología: imitar la forma de actuar o pensar de los seres humanos o construir sistemas que razonen o que actúen como ellos. En función de cómo se definan estos objetivos, Russell y Norvig clasifican las distintas definiciones en cuatro categorías (2016, p. 2):

- **Sistemas que actúan como humanos**: visión inaugurada por Alan Turing y su famoso test (1950). Se trata de sistemas y programas con capacidad de procesar el lenguaje natural, representar conocimiento, razonar automáticamente y aprender para adaptarse a nuevas circunstancias (Kurzweil, 1990).
- **Sistemas que piensan como humanos**: sistemas capaces de automatizar operaciones mentales, como la toma de decisiones, la resolución de problemas o el aprendizaje (Bellman, 1978).
- **Sistemas que piensan racionalmente**: sistemas que tratan de emular el pensamiento lógico racional y de alcanzar conclusiones de acuerdo a una serie de leyes universales del pensamiento definidas por la lógica (Winston, 1992).
- **Sistemas que actúan racionalmente**: sistemas que tratan de ampliar la racionalidad más allá de las leyes de la lógica e incluir así otros elementos, como la incertidumbre, la autonomía, el cambio, etc. (Poole, Mackworth y Goebel, 1998).

2.3. INTELIGENCIA ARTIFICIAL E INTELIGENCIA HUMANA

Existe, por tanto, un consenso entre autores al afirmar que la IA busca replicar la inteligencia humana en sistemas informáticos y robots, pero las definiciones divergen al especificar qué facultades cognitivas se busca emular y programar en las máquinas. La existencia de distintas definiciones de IA responde a la falta de una definición única de inteligencia. Pretender mimetizar la inteligencia humana en un programa o sistema parecería requerir, en primer lugar, saber con precisión en qué consiste dicha inteligencia.

Sin embargo, cuando se intenta concretar el significado de inteligencia, nos topamos rápidamente con una gran variedad de definiciones procedentes de campos como la neurofisiología, la filosofía, la sociología o la psicología. Dentro de esta última, por ejemplo, se propone el uso del término "inteligencias múltiples" para referirse a la diversidad de habilidades y ámbitos en los que puede desplegarse el conocimiento humano

[2] Gottfried Wilhelm von Leibniz fue uno de los grandes pensadores de los siglos XVII y XVIII. Su obra abarca una gran variedad de campos como la filosofía, la teología, la matemática, el derecho o la lógica. Leibniz se interesó por el estudio del lenguaje y la representación del conocimiento. En concreto, supuso la existencia de un lenguaje universal (*characteristica universalis*), una especie de alfabeto del pensamiento humano que sirviera para representar de forma simbólica el conocimiento contenido en la ciencia, la matemática y la metafísica.

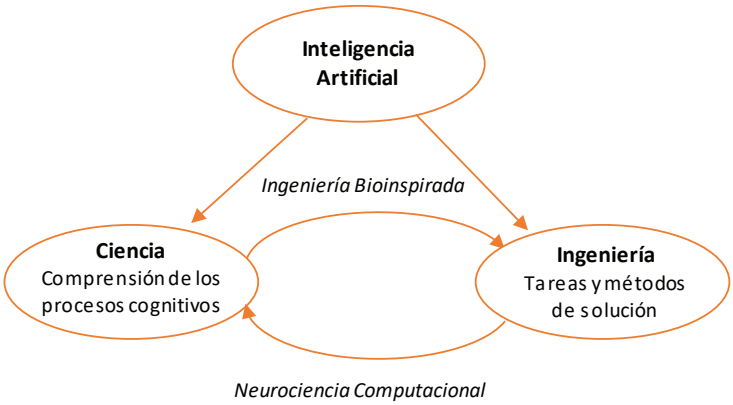
(Gardner, 2004). El término "inteligencia" parece poseer una connotación demasiado amplia, por lo que es de poca utilidad a la hora de hablar de programas diseñados para realizar tareas específicas, de ahí que la tesis de la Conferencia de Darthmouth ("Cualquier aspecto del aprendizaje y elemento de la inteligencia puede, en principio, ser descrito de manera tan precisa que sea posible crear una máquina que lo emule") sea, de hecho, ampliamente discutida por otros profesionales de la IA (Marín y Palma, 2008, p. 4).

Esto ha llevado a distintos investigadores a optar por un uso más sutil del término "inteligencia" y, con ello, por una definición más matizada de la IA. En los últimos años, ha cobrado especial relevancia la visión de la IA como un complemento, no un sustituto de la inteligencia humana. Esta concepción de la IA sostiene que los sistemas y aplicaciones dotados de IA han de poseer un carácter instrumental y deben enfocarse en complementar las deficiencias de la inteligencia humana, centrándose en aquellas tareas en las que una máquina es capaz de obtener un mejor desempeño que una persona. Esta visión ha dado lugar a estudios acerca del potencial que la IA posee en colaboración con otros agentes (personas, grupos sociales, ordenadores, etc.) (Malone, 2018).

Tras repasar brevemente las principales definiciones de la IA, resulta útil analizar con más detenimiento cómo esta trata de replicar en máquinas y sistemas informáticos algunas de las facultades cognitivas de los seres humanos. Poner en claro hasta qué punto la IA, como disciplina, ha logrado crear sistemas inteligentes será de gran ayuda a la hora de considerar las implicaciones éticas de esta tecnología. En consonancia con lo apuntado hasta ahora, se suele aceptar que el objetivo general de la IA es desarrollar (Mira, Delgado, Boticario y Díez, 1995):

1. **Modelos conceptuales**: diseñar modelos que proporcionen una mejor comprensión de los procesos cognitivos que tienen lugar en los seres humanos. Esta tarea se apoya en otras ramas, como la epistemología, la neurología u otras disciplinas relacionadas con la cognición. A esta rama de la IA se la suele denominar "neurociencia computacional".
2. **Procedimientos de reescritura formal de dichos modelos**: una vez observadas, analizadas y descritas las distintas funciones cognitivas de los diversos seres inteligentes, se busca reescribir los modelos obtenidos en un lenguaje formal, es decir, pasar de la descripción y observación del lenguaje natural a su formulación en un lenguaje formal; de un modelo conceptual del conocimiento humano a otro formal y computable.
3. **Estrategias de programación** y máquinas físicas para reproducir de la forma más eficiente y completa posible las tareas cognitivas y científico-técnicas más genuinas de los sistemas biológicos que hemos etiquetado como inteligentes.

Con base en estos objetivos, la IA suele dividirse en dos campos: la IA como ciencia y la IA como ingeniería del conocimiento (Marín y Palma, 2008, p. 5). La primera consiste en aquella rama de la IA encargada de buscar una teoría computable del conocimiento humano, es decir, "una teoría en la que sus modelos formales puedan ejecutarse en un sistema de cálculo y tener el mismo carácter predictivo que tienen, por ejemplo, las ecuaciones de Maxwell en el electromagnetismo" (Marín y Palma, 2008, p. 7). Esta rama es la encargada del primero de los objetivos generales de la IA: desarrollar modelos conceptuales. Para ello, la IA como ciencia analiza, en primer lugar, la estructura del conocer humano. Esto es, estudia todos aquellos procesos y mecanismos, tanto a nivel neuronal como subcelular, que dan lugar a las funciones de percepción, memoria, lenguaje, decisión, emoción y acción. Al hacerlo, trata de descomponer estos procesos cognitivos en subtareas hasta alcanzar el nivel de inferencias primitivas, es decir, de aquellos razonamientos que ya no necesitan de una descomposición posterior y que, por lo tanto, pueden ser fácilmente expresados en un



Fuente: Marín y Palma, 2008.

lenguaje formal. Por su lado, la IA como ingeniería del conocimiento tiene como objetivo la reescritura formal de las inferencias que se hayan obtenido en el análisis de los procesos cognitivos. Una vez elaborado el modelo formal, la IA como ingeniería se ocupa, en último lugar, de programar los operadores y sistemas.

Existe, como es lógico, un gran debate en torno al potencial de este planteamiento para explicar la totalidad de los procesos cognitivos. El nacimiento de la IA en el verano de 1956 se vio envuelto en unas expectativas poco realistas respecto a las posibilidades de replicar la inteligencia humana en máquinas y sistemas. Desde entonces, varios autores cuestionan que el pensamiento creativo, la percepción, el aprendizaje o la comprensión puedan ser totalmente conmensurables en una dinámica de modelaje, formalización y programación. No está del todo claro que el lenguaje natural pueda ser expresado en su integridad en el lenguaje formal, que la semántica pueda reducirse a la sintaxis o que el conocimiento pueda estar contenido en arquitecturas formales[3]. Con todo, no es preciso determinar con exactitud las posibilidades reales de emular la inteligencia humana ni delimitar con precisión si un sistema informático o máquina es inteligente o no. Tal como veremos a continuación, en la medida en que estos sistemas y aplicaciones son capaces de "decidir" por sí mismos, su diseño y funcionamiento adquiere relevancia ética.

2.4. INTELIGENCIA ARTIFICIAL, AUTONOMÍA Y ÉTICA

Tal como ha quedado demostrado, la IA persigue —como principal objetivo— replicar la inteligencia humana en máquinas y sistemas informáticos. Hasta qué punto esto sea posible constituye un tema de debate abierto entre los expertos en IA, científicos y filósofos. Según hemos visto, dicho debate responde a la existencia de múltiples definiciones de inteligencia, lo que origina, a su vez, diversas maneras de concebir la naturaleza y el propósito de la IA.

La relevancia ética del desarrollo de la IA no permanece ajena a este debate. El campo de la ética se extiende allá donde encuentra agentes dotados de autonomía e inteligencia, es decir, sujetos capaces de tomar decisiones y actuar de forma racional. En este sentido, el término "autonomía"[4] suele emplearse en este contexto para designar la capacidad de escoger un curso de acción de forma libre, capacidad que tradicionalmente se ha identificado como un rasgo distintivo y exclusivo de los seres humanos (European Group on Ethics in Science and New Technologies, 2018, p. 9). Esta autonomía no ha de ser confundida con la capacidad que presentan otros seres vivos de dirigir sus actos. El león,

[...] el término "autonomía" suele emplearse en este contexto para designar la capacidad de escoger un curso de acción de forma libre, capacidad que tradicionalmente se ha identificado como un rasgo distintivo y exclusivo de los seres humanos.

[3] El lector interesado en esta discusión, y en el antiguo debate filosófico entre mente y cerebro, puede consultar las obras de Clancey (1997, 1999), Dreyfus (1972, 1994), Searle (1987) o Edelman (1987).

[4] La propia etimología de la palabra confirma este significado. Del griego antiguo, αὐτονομία se compone del término *autós* ("propio, mismo") y *nómos* ("ley").

por ejemplo, caza a su presa movido por su instinto animal, pero es incapaz de hacerse cargo de los motivos que guían su conducta. Por esta misma razón, la autonomía humana viene siempre ligada a la responsabilidad: las personas, al ser capaces de dar cuenta de sus propios motivos y razones, responden también por ello ante los actos y decisiones tomadas.

[...] emplear el término "autónomo" para referirse a los dispositivos provistos de IA puede llevar a confusión [...] puesto que hasta la fecha ningún sistema o artefacto inteligente es capaz de dar cuenta de sus propios actos y decisiones de la manera en la que las personas son capaces, resulta erróneo calificar a estos dispositivos de autónomos en el sentido ético de la palabra.

A la vista de estas consideraciones, emplear el término "autónomo" para referirse a los dispositivos provistos de IA puede llevar a confusión. Se suele afirmar que muchas de las aplicaciones, sistemas y máquinas dotados de IA son capaces de operar de forma "autónoma", es decir, de realizar operaciones y procesos por sí mismos. Sin embargo, puesto que hasta la fecha ningún sistema o artefacto inteligente es capaz de dar cuenta de sus propios actos y decisiones de la manera en la que las personas son capaces, resulta erróneo calificar a estos dispositivos de autónomos en el sentido ético de la palabra. Es cierto que en el ámbito tecnológico ha proliferado el uso de dicho término para referirse simplemente a aplicaciones capaces de operar sin supervisión humana, pero, desde el punto de vista del análisis ético, el término adecuado para referirse a estos dispositivos sería "automático", no "autónomo".

Con todo, la existencia de aplicaciones capaces de tomar distintos cursos de acción por sí mismas no deja de plantear algunas dudas sobre la creencia de que solo los seres humanos son capaces de actuar de forma autónoma. Aclarar este punto es de una importancia capital, pues la relevancia ética de los sistemas provistos de IA será una u otra en función de quién sea verdaderamente responsable de los razonamientos y las decisiones de estos dispositivos: las personas encargadas de su diseño o los propios dispositivos. En esta línea, algunos autores proponen distinguir entre una IA débil y una fuerte (López de Mántaras, 2015). Esta distinción entre IA débil e IA fuerte fue introducida inicialmente por el filósofo John Searle en el año 1980. Según esta visión, la IA débil sería la ciencia que permitiría diseñar y programar ordenadores capaces de realizar tareas de forma inteligente, mientras que la IA fuerte sería la ciencia que permitiría replicar en máquinas la inteligencia humana. En otras palabras, la IA débil permitiría desarrollar sistemas con inteligencia especializada —ordenadores que juegan al ajedrez, diagnostican enfermedades o resuelven teoremas matemáticos—, mientras que la IA fuerte permitiría desarrollar ordenadores y máquinas dotados de inteligencia de tipo general. La presencia de inteligencia general —la que presentan los seres humanos— sería la condición suficiente para inferir que un dispositivo puede actuar de forma autónoma, no solo automática.

La relevancia ética de los dispositivos y sistemas dotados de una IA débil sería fácil de determinar. Estos dispositivos pueden operar de forma automática, pero no poseen inteligencia de tipo general y, por tanto, carecen de autonomía en sentido estricto. Se trata de aplicaciones programadas de antemano para realizar una única tarea y, en muchos casos, han demostrado superar con creces la pericia humana. Pero se trata de sistemas incapaces de actuar de forma racional, por lo que su actividad no es éticamente imputable. En los sistemas de este tipo, la responsabilidad ética recae en su totalidad sobre las personas encargadas de su diseño y funcionamiento. Puesto que estas aplicaciones solo tienen la posibilidad de operar de acuerdo a un único curso de acción, la reflexión ética en torno a este tipo de tecnología ha de centrarse en la programación y el diseño de estos sistemas.

La consideración ética de las aplicaciones dotadas de una IA fuerte resulta algo más compleja, en primer lugar, porque no resulta claro que sea posible crear máquinas con inteligencia de tipo general. Hoy sabemos que la complejidad del cerebro humano es prácticamente imposible de replicar (López de Mántaras, 2015) y que los intentos de desarrollar máquinas capaces de mostrar un comportamiento inteligente en varios ámbitos no han tenido mucho éxito. Un ordenador capaz de vencer al ajedrez a cualquier ser humano es incapaz de aplicar ese conocimiento a un juego similar como las damas, cuando cualquier persona podría inferir rápidamente las reglas y comenzar a jugar.

La práctica imposibilidad de crear un sistema dotado de inteligencia y sentido común hace igualmente dudosa la propuesta de la *singularidad* tecnológica (inteligencias artificiales dotadas de inteligencia general capaces de automejorarse progresivamente hasta superar con creces la inteligencia humana) y otorga un carácter distópico a las reflexiones en torno a los potenciales riesgos derivados de esta tecnología.

A pesar de que una IA fuerte parezca impracticable, es cierto que ya existen aplicaciones dotadas de IA con capacidad de tomar decisiones y escoger entre diversos cursos de acción. Los vehículos automáticos o robots domésticos son un buen ejemplo de ello. Aun así, es preciso afirmar que incluso en estos sistemas la capacidad de toma de decisiones es solo aparente: están programados para realizar un cálculo de probabilidades teniendo en cuenta una diversidad de factores y circunstancias, pero, dados un escenario determinado y un conjunto de datos limitado, el curso de acción óptimo que estos sistemas seguirán siempre será uno. Con todo, es cierto que dentro de este cálculo probabilístico ya entran en consideración factores de relevancia ética. Tal como veremos en el último apartado, en los sistemas de este tipo la responsabilidad ética sigue recayendo en su totalidad sobre las personas encargadas de su diseño y funcionamiento, pero al ser mayor el grado de automatización, la reflexión ética ha de tener en cuenta principios de diseño más específicos, como la trazabilidad de los algoritmos, la explicabilidad, la rendición de cuentas o la supervisión humana. En el siguiente apartado tendremos ocasión de analizar algunas de estas aplicaciones dotadas de IA y examinaremos con más detenimiento cuáles son los riesgos éticos que plantean.

[...] en los sistemas de este tipo la responsabilidad ética sigue recayendo en su totalidad sobre las personas encargadas de su diseño y funcionamiento, pero al ser mayor el grado de automatización, la reflexión ética ha de tener en cuenta principios de diseño más específicos, como la trazabilidad de los algoritmos, la explicabilidad, la rendición de cuentas o la supervisión humana.

3. APLICACIONES DE LA INTELIGENCIA ARTIFICIAL: RETOS ÉTICOS

En el apartado anterior presentamos distintas definiciones de IA y repasamos los principales objetivos que este tipo de tecnología pretende alcanzar. Asimismo, ha quedado explicado cómo la relevancia ética de la IA se encuentra siempre aparejada al nivel de automatización que los dispositivos inteligentes son capaces de alcanzar: a mayor automatización y capacidad operativa, mayor importancia cobra el diseño y programación de los sistemas y máquinas provistos de IA.

Ahora bien, que el desarrollo de la IA sea éticamente relevante aún nos dice poco sobre la conveniencia o inconveniencia de su desarrollo. En otras palabras, es preciso examinar los avances en el campo de la IA desde la perspectiva de la moralidad. La ética se entiende habitualmente como el estudio de la moralidad, es decir, como el estudio y discusión de los bienes, las normas y las conductas que contribuyen al desarrollo y florecimiento de la vida humana (Sterba, 2009, p. 1). Comúnmente se suele aceptar dentro de estos bienes la protección de la vida humana, la libertad y la dignidad humanas, la religión, etc. (Finnis, 2011). No es nuestra intención aquí evaluar el desarrollo de la IA desde este punto de vista normativo. Sin embargo, sí es preciso subrayar que la perspectiva de la moralidad ha de ser tenida en consideración a la hora de evaluar con rigor los potenciales riesgos y beneficios de las distintas aplicaciones dotadas de IA, pues no basta con afirmar que esta tecnología posee relevancia ética, es decir, no es suficiente reconocer que la IA puede contribuir a o perjudicar la vida y las relaciones humanas. Es necesario establecer, primero, en qué puede consistir esa vida buena y qué papel puede desempeñar la tecnología en su consecución.

Tampoco resultaría prudente dejar los desarrollos en el campo de la IA al margen del análisis ético: en primer lugar, por la potencial capacidad de autonomía que estos dispositivos pueden llegar a alcanzar, la cual, como ya vimos, reclama un atento análisis ético; y, en segundo lugar, por el rápido desarrollo y penetración que la IA está experimentando.

A partir de la década de los noventa del pasado siglo, la IA comenzó a emplearse en un número más amplio de aplicaciones, lo que contribuyó a una mayor difusión entre distintas industrias. Los estudios realizados acerca de la penetración de la IA confirman la utilidad que esta tecnología brinda a las empresas en distintas fases de la cadena de valor y señalan la tendencia creciente a incorporar nuevos dispositivos inteligentes en los próximos años. En un estudio realizado en el 2018 por McKinsey & Company entre más de 2.000 directivos, el 47% afirmó que sus compañías ya contaban con al menos un sistema dotado de IA en algún segmento de la cadena valor. Al preguntarles acerca de la inversión destinada a este tipo de tecnología, el 71% de los encuestados afirmó que esta crecería exponencialmente durante los próximos años (McKinsey & Company, 2018).

El empleo de la IA presenta grandes diferencias según la industria y el sector. Asimismo, las distintas aplicaciones de la IA muestran distintos niveles de uso entre las empresas, según el entorno y la ventaja competitiva que el empleo de esta tecnología confiera a cada compañía[5]. En la actualidad, las aplicaciones de IA más extendidas son:

- **Vehículos automatizados**: robots, vehículos u otro tipo de dispositivos móviles con capacidad para desplazarse de forma autónoma sin la dirección de una persona.
- **Reconocimiento de voz**: dispositivos con capacidad de identificar el lenguaje humano, procesarlo e interactuar con él.
- **Planificación autónoma**: dispositivos con capacidad de organizar y planificar tareas y operaciones de acuerdo a unos objetivos previamente establecidos.
- **Visión artificial**: tecnología capaz de procesar, analizar y comprender las imágenes del mundo real y formalizarlas para que puedan ser tratadas por un ordenador.
- ***Machine learning***: capacidad de aprender que poseen determinados computadores y sistemas de aprendizaje, es decir, la habilidad de mejorar su desempeño en una determinada tarea a base de experiencia.

En este contexto, sería prácticamente imposible evaluar con detenimiento la moralidad de cada una de estas aplicaciones. Con todo, sí queremos repasar a continuación los principales usos de la IA que, a nuestro juicio, presentan unos beneficios claros, así como aquellos que plantean unos riesgos evidentes.

3.1. BENEFICIOS POTENCIALES DE LA INTELIGENCIA ARTIFICIAL

Son muchos los indicios que apuntan a que la IA estará presente en numerosos ámbitos de nuestras vidas (Comisión Europea, 2018a). A la espera de futuros avances y desarrollos, son ya bastante numerosos los beneficios que las aplicaciones y sistemas inteligentes nos están proporcionando. Los sistemas de reconocimiento de imagen se están empleando para identificar posibles anomalías en las radiografías. Esta tecnología ha demostrado ser, en concreto, entre un 62% y un 97% más eficaz que un panel de radiólogos a la hora de identificar nódulos en los pulmones (Koo et al., 2012).

La IA está permitiendo también a las empresas conocer mejor a sus clientes y desarrollar nuevas estrategias de marketing y comunicación. El banco suizo SEB emplea un asistente virtual llamado Aida para atender las llamadas de los clientes. El asistente es capaz de interactuar con ellos, de acceder a toda su información y de gestionar peticiones como abrir una cuenta o realizar una transferencia (Wilson y Daugherty, 2018), Con base en los datos almacenados, algunas compañías están utilizando asistentes inteligentes que, en función del perfil y los datos almacenados de cada cliente, aconsejan un determinado

[5] Para un análisis más detallado de la penetración de la IA en las distintas industrias y sus predicciones de uso para el futuro, pueden consultarse los siguientes informes: McKinsey & Company, 2018; PwC, 2019; Deloitte, 2019; Shoham, Y., Perrault, R. Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J. C., … Bauer, Z., diciembre del 2018.

producto o servicio (Rosenberg, 2018). Asistentes virtuales como Siri, Alexa o Google Home ofrecen al usuario la posibilidad de realizar infinidad de tareas mediante un comando de voz: desde hacer la compra desde casa hasta organizar la agenda o enviar un correo electrónico.

Las aplicaciones dotadas de IA están permitiendo también mejorar la eficiencia en distintos procesos de logística y transporte. Algunas empresas de almacenaje han comenzado a desarrollar fábricas inteligentes que no precisan de ningún operador humano (European Parlamentary Research Service, 2016, p. 6). En algunas granjas a lo largo de Europa, la IA se está empleando para registrar los desplazamientos, la temperatura y el consumo de alimentos de los animales. En España, por ejemplo, la empresa procesadora de alimentos El Dulze emplea robots para seleccionar y recoger las hojas de lechuga de la banda transportadora. Estos resultan ser mejores que las personas a la hora de escoger las hojas de más calidad: tras implantarlos, la tasa de desperdicio disminuyó del 20% al 5% (Fanuc Robotics Europe, 2012).

Asimismo, el creciente uso de vehículos automatizados presenta un gran potencial para contribuir a una drástica reducción del número de accidentes mortales que ocurren en desplazamientos por carretera. La Sociedad de Ingenieros de Automoción (SAE) establece seis niveles para medir la capacidad de conducción autónoma de un vehículo: cero es el nivel en el cual el coche no puede ser empleado por otra persona y seis, el nivel en el cual la conducción es totalmente automática. Compañías como General Motors, Renault-Nissan o Daimler y Bosch actualmente emplean vehículos con nivel cuatro de autonomía para fines comerciales y logísticos.

Todas estas aplicaciones presentan claros beneficios para multitud de usuarios. Para las empresas suponen, en muchos casos, un importante ahorro económico, al poder incrementar de forma exponencial la productividad en distintas fases de la cadena de valor. Se estima que, para el 2025, el impacto económico de las distintas aplicaciones de la IA estará comprendido entre los 6,5 y los 12 trillones de euros anuales (McKinsey Global Institute, 2013). Para millones de personas supone también un aumento en el nivel de vida, a través de la reducción de riesgos sanitarios, alimenticios o de transporte.

En España, CaixaBank lleva desde el 2013 trabajando en aplicaciones de IA en el sector bancario a partir de Watson (IBM) y sus tecnologías.

Entre las tecnologías desarrolladas cabe destacar los *chatbots*, gracias a la cual las máquinas pueden interactuar con las personas utilizando el lenguaje natural. En el 2017, CaixaBank lanzó **Gina**, el *chatbot* de imaginBank para atender consultas de los clientes. Poco después, se puso en marcha **Neo** (2018), integrado en la aplicación CaixaBankNow.

Gina y Neo son capaces de resolver por voz o texto las dudas del cliente y de ayudarle a encontrar las opciones que busca. Ambos están basados en una avanzada IA que entiende y responde más de 450 preguntas en diferentes idiomas.

Entre los servicios que facilitan, Gina y Neo permiten a los clientes realizar consultas sobre productos y servicios, pedir ayuda para contratar un producto o recibir recomendaciones sobre novedades relacionadas con sus intereses. A día de hoy, los asistentes de CaixaBank atienden más de 500.000 conversaciones al mes.

3.2. RIESGOS POTENCIALES DE LA INTELIGENCIA ARTIFICIAL

Junto con las ventajas que todas estas aplicaciones traen consigo, es bueno advertir también los riesgos que el uso de la IA lleva aparejados. Para las mismas aplicaciones mencionadas arriba, existen múltiples escenarios en los que el empleo de la IA da lugar a situaciones éticamente problemáticas[6]. Algunos de estos riesgos son comunes al uso de otros tipos de tecnología, mientras que otros constituyen riesgos específicos del campo de la IA.

Al igual que otros desarrollos tecnológicos como Internet, los ordenadores personales, los teléfonos móviles o la televisión, la IA plantea un tipo de riesgos derivados de la introducción de un nuevo avance técnico. Algunos de estos riesgos son:

1. **Destrucción de puestos de trabajo**: a lo largo de la historia, los cambios tecnológicos siempre han estado acompañados de profundos cambios sociales, lo que a menudo se ha traducido en la desaparición de muchos puestos de trabajo (Lin, Abney y Bekey, 2011). Parte del debate en torno a la IA se centra en las preguntas de si esta tecnología es equiparable a los grandes cambios tecnológicos del pasado y de si, por tanto, la penetración de los sistemas inteligentes se traducirá en la desaparición masiva de empleo (Quinn, 2015, p. 24). Aunque no dispongamos de suficientes datos, algunos estudios estiman que entre el 21% y el 38% del empleo en los países desarrollados podría desaparecer a causa de la digitalización y la automatización de la economía (Berriman, Hawsworth y Goel, 2017, p. 1). Al mismo tiempo, esta tecnología está propiciando la aparición de otras formas de empleo y competencias profesionales (World Economic Forum, 2016). El desafío ético que esta transformación supone no radica tanto en la mayor o menor conveniencia de estos cambios, sino en la capacidad de adaptación que las empresas, los trabajadores y los Gobiernos exhiban.

[...] algunos estudios estiman que entre el 21% y el 38% del empleo en los países desarrollados podría desaparecer a causa de la digitalización y la automatización de la economía (Berriman, Hawsworth y Goel, 2017, p. 1).

2. **Manipulación, seguridad y vulnerabilidad**: al igual que otros sistemas informáticos o aplicaciones tecnológicas, la IA se compone de elementos de software y hardware susceptibles de funcionar erróneamente. A su vez, muchas de las aplicaciones provistas de IA operan mediante algoritmos basados en modelos estadísticos y grandes cantidades de información, lo que también puede llevar a decisiones sesgadas o conclusiones incompletas. Esta tecnología, además, puede ser manipulada para distintos fines, como manipular unas elecciones (Polonski, 2017) o modificar el precio de distintos productos y servicios (OCDE, 2017).

3. **Transformación de las relaciones humanas**: existe ya un consenso en torno al perjuicio que el uso prolongado de pantallas, dispositivos móviles y redes sociales produce sobre nuestras habilidades cognitivas (Greenemeier, 2011), nuestra estabilidad emocional (Heid, 2018) y nuestra salud física (Goldhill, 2015). La proliferación de dispositivos dotados de IA en los que delegar muchas de nuestras interacciones y procesos sociales —toma de decisiones, comunicación, planificación— podría derivar en una pérdida significativa de habilidades personales (Groth, Nitzberg y Esposito, 2018).

4. **Erosión de la sociedad civil**: episodios recientes como el de Cambridge Analytica[7] han demostrado que los espacios de diálogo abiertos por los nuevos medios tecnológicos esconden, en muchos casos, intereses políticos y económicos concretos. De forma similar, existe el riesgo de que la introducción de sistemas inteligentes

[6] Para un estudio más detallado de los potenciales riesgos derivados del uso de la inteligencia artificial, puede consultarse el informe realizado por el FHI de la University of Oxford: Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, ... Anderson, H., 2018, en colaboración con otros centros de investigación.

[7] Se conoce por Cambridge Analytica al escándalo que involucró a la consultora británica de análisis de datos Cambridge Analytica y a Facebook. En marzo de 2018, un antiguo empleado de Cambridge Analytica publicó un artículo en el que explicaba cómo la empresa había estado empleando y analizando desde 2015 datos personales millones de usuarios de Facebook al servicio de distintas campañas políticas. Tras destaparse el escándalo, Facebook perdió 100 millones de dólares de capitalización bursátil y su CEO, Mark Zuckerberg, tuvo que comparecer ante el congreso para dar explicaciones.

en los medios de información y comunicación pueda distorsionar la opinión pública y cercenar la pluralidad de puntos de vista. Este es el caso de los chatbots programados para compartir información y noticias bajo un sesgo determinado e influir en la toma de decisiones.

La IA se encuentra, en mayor o menor medida, presente en todos estos riesgos en tanto que muchos de los actuales sistemas de información, programas informáticos o dispositivos incorporan algún tipo de funcionalidad o software provisto de IA. Junto con estos riesgos, la IA presenta también una serie de peligros más específicos relacionados con el desarrollo y el funcionamiento de esta tecnología:

1. **Rendición de cuentas**: los dispositivos y sistemas dotados de IA interactúan cada vez más con las personas y su entorno. Esta interacción suscita la pregunta de quién es responsable de los daños producidos en caso de que alguno de estos dispositivos opere erróneamente o tome una decisión de forma autónoma que resulte en algún tipo de perjuicio (Comisión Europea, 2019). A primera vista, parecería que cualquier daño que un sistema inteligente pueda ocasionar es responsabilidad de las personas encargadas del diseño y la programación de los algoritmos. Sin embargo, esta distinción se vuelve más difusa en la medida en que aumenten la autonomía y la capacidad de decisión de estos sistemas. Si un vehículo autónomo decide estrellarse contra una vivienda para evitar un accidente mortal, ¿quién es responsable de los daños producidos sobre una propiedad ajena, el conductor del vehículo, el fabricante o el propio vehículo que tomó esa decisión?
2. **Explicabilidad**: estrechamente relacionado con la problemática anterior se encuentra el riesgo de la explicabilidad. Lo que en muchas situaciones puede dificultar una clara asignación de daños, perjuicios y responsabilidades es, precisamente, la falta de una explicación clara de por qué un sistema inteligente tomó una decisión en particular. Los dispositivos dotados de IA plantean el riesgo de acabar tomando decisiones impredecibles o inexplicables para un ser humano. Debido a la complejidad de los algoritmos, existe el peligro de que estos sistemas lleguen a conclusiones y resultados inexplicables para los usuarios (Bostrom y Yudkowsky, 2014, p. 1). Este es el principal problema de los actuales sistemas basados en *deep learning*[8]: son cajas negras cuyo proceso de toma de decisiones no puede ser trazado o explicado (UK, 2017).

El caso de los vehículos autónomos

Una de las aplicaciones de la IA que más debate ha producido ha sido la aparición de los vehículos automáticos. El empleo de este tipo de automóviles implica supeditar la seguridad de pasajeros y peatones al funcionamiento del sistema informático que controla el vehículo, lo que en principio supondría un descenso dramático del número de accidentes mortales que se producen en las carreteras.

Sin embargo, la introducción de este tipo de vehículos también plantea una serie de problemas. El fallo de estos sistemas ya ha ocasionado la muerte de varias personas en los últimos años (Wakabayashi, 2018). Más aún, la elaboración de estos vehículos exige el diseño de un algoritmo capaz de establecer prioridades a la hora de distribuir el daño entre distintos escenarios. Por ejemplo, ante un escenario de siniestro total, el algoritmo ha de "escoger" qué vida humana posee más valor y, por lo tanto, debe ser salvada (Bonnefon *et al.*, 2016). El diseño de este tipo de algoritmos suscita varias cuestiones éticas: ¿es posible comparar el valor de dos vidas humanas?, ¿de qué manera debería ser programado dicho algoritmo?, ¿resulta prudente poner en circulación vehículos de este tipo?

[8] El *deep learning*, o aprendizaje profundo, es una rama del m*achine learning* que investiga cómo operar con algoritmos ideados para aprender automáticamente. Se basa en el empleo de redes neuronales artificiales y su premisa fundamental es el uso de varias capas de datos en las que cada nivel aprende a extraer información más abstracta o elaborada a partir de la información de la que se dispone.

3. **Imparcialidad**: los sistemas dotados de IA, especialmente aquellos que operan con grandes cantidades de datos, pueden contener en su programación algún tipo de sesgo o prejuicio que los lleve a alcanzar conclusiones parciales o injustas (Mittelstadt, Allo, Taddeo, Watcher y Floridi, 2016, p. 7). El uso de esta tecnología en actividades de ventas o marketing puede conducir a un efecto aumentado de estos sesgos. Empresas que empleaban sistemas dotados de IA para realizar transacciones comerciales han comprobado cómo, en determinadas ocasiones, los algoritmos de aprendizaje automático pueden discriminar en función de la raza o el género (Buolamwini y Gebru, 2018). Distintos estudios han mostrado cómo la asignación de una hipoteca mediante programas informáticos inteligentes puede dar como resultado precios más elevados según la raza o el color de piel del cliente (Barlett, Morse, Stanton y Wallace, 2019). Al mismo tiempo, para muchas de estas aplicaciones se plantea la problemática de cómo dar con muestras lo suficientemente grandes y representativas o de cómo modificar manualmente el conjunto de datos para que no altere significativamente los resultados.
4. **Privacidad**: gran parte de las aplicaciones provistas de IA basan su funcionamiento en el acceso a grandes cantidades de información. En este contexto, existe una preocupación ante el uso y la gestión que pueda hacer de esta información, especialmente de la que posee carácter personal. Algunos sistemas inteligentes, como los asistentes virtuales —Alexa, Google Now o Siri—, se encuentran presentes en miles de hogares, en los cuales recogen y procesan —aun cuando el dispositivo se encuentre apagado— millones de conversaciones, muchas de las cuales contienen información sensible para el usuario (Fussell, 2019). Toda esta información es suficiente para que la empresa correspondiente infiera nuestro estado de ánimo o salud física a partir de las conversaciones que mantenemos (Shulevitz, 2018). La cantidad de información que estas empresas acumulan pone encima de la mesa cuestiones como la capacidad de manipulación de muchos sectores de la población, la erosión de distintas instituciones democráticas o la creación de hábitos psicológicamente dañinos para las personas (Eyal, 2017).

Resulta evidente que la IA trae consigo numerosos beneficios para las empresas y demás grupos que componen la sociedad. Son muchas las ventajas que las aplicaciones provistas de IA —y las que se irán desarrollando durante los próximos años— pueden aportar a ámbitos como el educativo, el sanitario, el comercial o el del transporte. El enorme potencial que esta tecnología posee pide a su vez un ejercicio de la prudencia ante los posibles riesgos que su uso pueda desencadenar, de ahí que, junto con los peligros inherentes a la introducción de una nueva forma de tecnología, se haga necesario también un análisis pormenorizado de los riesgos específicos que plantea la IA.

Queda pendiente, pues, la tarea de definir en qué consiste el uso prudente de estas aplicaciones, mediante la elaboración de una serie de principios que guíen su diseño y desarrollo, y la acotación de los ámbitos en los que puedan emplearse de forma segura y fiable. A lo largo de estas páginas se ha incidido, en varias ocasiones, en la importancia que tiene el diseño de esta tecnología a la hora de asegurar un empleo seguro y prudente. Junto con la fase del diseño, en el siguiente apartado repasaremos también qué otras herramientas pueden emplear las empresas, los Gobiernos y los organismos reguladores para garantizar que la IA se emplee de manera responsable y para asegurar que se siga desarrollando de manera prudente.

4. LA ÉTICA EN EL DISEÑO DE LA INTELIGENCIA ARTIFICIAL

Ante el potencial que la IA exhibe, urge establecer unos principios y proponer una serie de pautas para que esta tecnología se emplee y siga desarrollándose de manera responsable y segura. La existencia de unos principios y estándares correctamente delimitados puede contribuir sobremanera a potenciar los beneficios derivados de la IA y a minimizar los riesgos que su uso implica. En el presente apartado proponemos una serie de principios básicos que pueden servir de guía para orientar los posibles usos y desarrollos de la IA. Tal como hemos visto, las aplicaciones de la IA son de diverso tipo y su uso varía enormemente. Esto hace que sea prácticamente imposible establecer un único procedimiento o código de conducta específico para la IA en general. Los principios aquí enumerados poseen, por tanto, un carácter orientativo y su aplicación habrá de concretarse de manera distinta según el contexto y el uso de la IA que se estén considerando. En esta misma línea, recogemos también en este apartado una serie de métodos que pueden ser de utilidad para llevar a la práctica los principios aquí propuestos.

Conviene mencionar que son muchos los organismos e instituciones que ya se han apresurado a elaborar una lista de principios y recomendaciones para asegurar un uso adecuado de la IA. Entre los documentos disponibles, merece la pena destacar los publicados por distintos órganos de la Unión Europea, que abordan también la dimensión regulatoria de los robots y sistemas autónomos; las recomendaciones del Foro Económico Mundial, en las que se proponen distintos métodos para asegurar un buen uso de la IA por parte de empresas y Gobiernos; y los informes publicados por el UNICRI Centre for AI and Robotics de las Naciones Unidas y por la Unesco. En los últimos años, también han visto la luz diversas iniciativas de la mano de empresas, asociaciones profesionales o académicos que han contribuido a formular criterios de buenas prácticas para el uso de la IA. Entre ellas destacan los estándares formulados por la IEEE (Institute of Electrical and Electronics Engineers); los principios Asilomar sobre IA propuestos por el Future of Life Institute; o la declaración para el desarrollo responsable de la IA elaborada por el Forum on the Socially Responsible Development of Artificial Intelligence, celebrado en la Université de Montréal en el 2017.

En el ámbito español, merece la pena destacar la *Declaración de Barcelona sobre la inteligencia artificial*, firmada en el 2017 por distintos expertos en IA y promovida por la Obra Social ”la Caixa”. El lector podrá encontrar, al final de este cuaderno, un extenso listado con los documentos más relevantes publicados por estos organismos. En este apartado hemos tratado, simplemente, de recoger los puntos más importantes contenidos en dicha bibliografía y presentarlos de manera más accesible.

4.1. PRINCIPIOS ÉTICOS PARA EL DISEÑO Y DESARROLLO DE LA INTELIGENCIA ARTIFICIAL

Un primer punto de partida para garantizar una IA segura y robusta es proponer una serie de principios éticos que abarquen las distintas aplicaciones y sistemas provistos de IA (vehículos autónomos, asistentes virtuales, robots, software, etc.). Estos principios no poseen una formulación tan específica como cabría esperar de un código de conducta o de una serie de normas. Se trata, más bien, de una serie de imperativos en los que se presenta un bien humano que debe ser respetado en toda situación y promovido mediante nuestras acciones.

En el ámbito de la ética y la moralidad, es frecuente encontrar principios formulados de esta manera: por ejemplo, "no matarás"; o el conocido imperativo categórico de Immanuel Kant: "Obra de tal modo que uses a la humanidad, tanto en tu persona como en la persona de cualquier otro, siempre al mismo tiempo como fin y nunca simplemente como medio" (Kant, 1785). De manera similar, el campo de la robótica y de la IA es susceptible también de albergar formulaciones semejantes. Muchos de los principios que es posible encontrar en los documentos listados se inspiran previamente en las leyes de la robótica elaboradas por Isaac Asimov (*Rundaround*, 1942, p. 94):

> I. Un robot no puede hacer daño a un ser humano o, por inacción, permitir que un ser humano sufra daño. II. Un robot debe obedecer las órdenes dadas por los seres humanos, excepto si estas órdenes entrasen en conflicto con la primera ley. III. Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la primera o la segunda ley.

Aunque estas últimas pertenezcan al ámbito de la ciencia ficción, constituyen un buen modelo en el que inspirarse: identifican un bien en particular —en este caso, la vida y la autonomía de los seres humanos— que debe ser protegido y respetado —en el universo de Asimov, por los robots—.

El desarrollo actual de la IA plantea un escenario mucho más realista y apremiante que cualquier otro de ciencia ficción. En vista de los riesgos que los sistemas y dispositivos inteligentes poseen, se hace necesario identificar qué bienes humanos entran en peligro en este escenario y formular, como consecuencia, una serie de principios que orienten el uso de la IA hacia su defensa y promoción. Tras haber repasado en el apartado anterior algunos de estos riesgos, y tomando en consideración los principios propuestos por los organismos citados anteriormente, proponemos el siguiente listado:

1. **Respeto de la autonomía humana**: los sistemas inteligentes deben respetar en todo momento la autonomía y los derechos fundamentales de las personas. Su diseño y programación debe respetar, por tanto, la vida y los derechos humanos sin ningún tipo de discriminación.
2. **Transparencia**: en el caso de los sistemas provistos de IA, la transparencia atañe principalmente a la explicabilidad y la trazabilidad de dichos sistemas. Dado que el diseño de estos dispositivos contempla que tomen decisiones automáticamente con base en distintos cálculos y proyecciones, debe ser posible en todo momento trazar el razonamiento seguido por el sistema y explicar las consecuencias alcanzadas. En concreto, ha de ser posible trazar el conjunto de datos empleados en el razonamiento, el funcionamiento del algoritmo y los pasos seguidos para alcanzar los resultados. Todo este proceso debe ser, además, explicable desde los puntos de vista técnico de la programación y humano del diseño. El diseño y el empleo de una tecnología impredecible son incompatibles con la defensa de la autonomía humana. Es absolutamente necesario que la actividad de todos estos dispositivos sea de fácil comprensión y acceso.
3. **Responsabilidad y rendición de cuentas**: estrechamente relacionado con el principio anterior, el diseño y el empleo de sistemas inteligentes deben estar precedidos por una clara asignación de responsabilidades ante los posibles daños y perjuicios que estos puedan ocasionar. La presunta autonomía de estos sistemas no puede servir de pretexto para la dilución de responsabilidades. Al contrario, será preciso incluir los mecanismos adecuados (auditoría, informe de errores, penalizaciones, etc.) para asegurar que las responsabilidades y obligaciones en relación con el funcionamiento de estos sistemas queden bien definidas.

La presunta autonomía de estos sistemas no puede servir de pretexto para la dilución de responsabilidades [...] será preciso incluir los mecanismos adecuados [...] para asegurar que las responsabilidades y obligaciones en relación con el funcionamiento de estos sistemas queden bien definidas.

4. **Robustez y seguridad**: la fiabilidad de la IA exige que los algoritmos sean suficientemente seguros, fiables y sólidos para operar de manera precisa y segura, y para resolver errores o incoherencias durante todas las fases del ciclo de vida útil de los dispositivos. Este principio exige, además, que los sistemas se diseñen y desarrollen contemplando la posibilidad de ciberataques y fallos técnicos.
5. **Justicia y no discriminación**: el diseño de estos sistemas debe contar con la participación de todos los grupos de interés con los que cada aplicación provista de IA se relacione. Además, estos dispositivos deben garantizar un empleo justo de los datos disponibles para evitar posibles discriminaciones hacia determinados grupos o distorsiones en los precios y en el equilibrio de mercado.

Estos principios son de especial relevancia durante la fase de diseño —pues es en este momento cuando queda configurada y programada la práctica totalidad de las funcionalidades de las que cada aplicación es capaz—, pero atañen también a las fases de desarrollo, introducción y adopción de esta tecnología. Durante esta fase inicial, se pueden abordar con efectividad algunos de los principales riesgos que plantea la IA, de ahí que en muchos de estos documentos se hable con frecuencia de mecanismos *value-by-design* (*security-by-design, privacy-by-design*, etc.), es decir, de métodos que permitan materializar estos principios éticos en el diseño y la programación específica de los algoritmos.

Junto con la posibilidad de introducir parámetros éticos durante esta fase, el diseño de estos sistemas es de una importancia crucial debido a la propia naturaleza de la IA. ha quedado mencionado cómo los sistemas dotados de IA son "autónomos" en un sentido restringido de la palabra: el razonamiento y el proceso de toma de decisiones que llevan a cabo responden siempre a una programación previa que, por muy amplia y compleja que pueda ser, queda también siempre dentro de los límites establecidos por el diseño, de ahí que la responsabilidad ética ligada a los diversos usos de la IA recaiga directamente sobre las personas, no sobre las máquinas. Por esto mismo, las consideraciones éticas y sociales deben comparecer desde la primera fase de diseño de la tecnología (Buchholz y Rosenthal, 2002).

[...] los sistemas dotados de IA son "autónomos" en un sentido restringido de la palabra: el razonamiento y el proceso de toma de decisiones que llevan a cabo responden siempre a una programación previa [...] de ahí que la responsabilidad ética ligada a los diversos usos de la IA recaiga directamente sobre las personas, no sobre las máquinas.

Incorporar criterios éticos durante la fase de diseño reviste una gran importancia, pero no es suficiente para garantizar que las aplicaciones provistas de IA sean seguras y se empleen de manera responsable. La IA, como cualquier otra tecnología, ha sido desarrollada por personas y organizaciones que encarnan los objetivos, las normas y los valores de la cultura a la que pertenecen (Bijker, Hughes y Pinch, 1987). A la vez, la tecnología influye en nuestra manera de comunicarnos, de desplazarnos, de relacionarnos y, en definitiva, de vivir (Verbeek, 2011, p. 4). En resumidas cuentas, la IA no es un simple producto del que deban ocuparse los diseñadores: es un producto social en el que se ven involucrados distintos grupos de interés y en el que entran en juego fuerzas de origen social, político y económico (Craglia *et al.*, 2018). Por ello, a la preocupación por el diseño deben acompañarla:

- La participación activa de la sociedad civil en la discusión de los valores, los objetivos y los beneficios de la IA (Janasoff, 2013).
- Un espacio de diálogo entre los distintos grupos de interés involucrados en el desarrollo de la IA en el que puedan ser debatidos qué valores éticos y sociales deben ser implantados y cuál es la manera más adecuada de hacerlo (Cath, Watcher, Mittelstadt, Taddeo y Floridi, 2018).

En definitiva, las consideraciones éticas durante la fase de diseño de aplicaciones deben estar acompañadas también de una serie de mecanismos concretos que permitan a los distintos grupos de interés (empresas, investigadores, reguladores, Gobiernos, consumidores, etc.) incorporar principios y normas al funcionamiento de estas nuevas

aplicaciones. Mediante un enfoque semejante, que subraye la importancia del diseño, pero que haga partícipes de este proceso a todos los actores pertinentes, resulta posible alcanzar un uso más robusto y seguro de la IA.

4.2. MÉTODOS TÉCNICOS Y NO TÉCNICOS

Junto con la formulación de estos principios, se hace también necesario disponer de los métodos adecuados para implementarlos de manera eficaz en los dispositivos provistos de IA. De nuevo, los métodos aquí propuestos comprenden todo el ciclo de vida del producto, pero atañen de forma especial a la fase de diseño (Comisión Europea, 2019, p. 20).

4.2.1. Métodos técnicos

Los siguientes métodos tienen como objetivo traducir los principios éticos de la IA en un diseño y una programación específicos que garanticen que el comportamiento de estos sistemas se desarrolle por defecto en consonancia con dichos requerimientos. Puesto que los dispositivos inteligentes operan con base en un determinado algoritmo y conjunto de datos, programar su arquitectura conforme a una serie de parámetros éticos puede garantizar que estos sistemas se comporten siempre de la manera deseada.

- ***Ethics by design***: se trata de uno de los métodos que más atención ha recibido en los últimos años. Mediante el diseño previo de los algoritmos que controlan los sistemas inteligentes, se podría garantizar el comportamiento ético de estos. Dentro de esta categoría es posible encontrar distintos planteamientos y terminologías como *security-by-design* o *privacy-by-design*. Los mecanismos propuestos para lograrlos son varios (Etzioni y Etzioni, 2017). En primer lugar, se podría hacer que los sistemas y robots provistos de IA adquieran patrones éticos de conducta "observando" el comportamiento humano en situaciones concretas. Un segundo planteamiento consistiría en establecer una serie de normas que el sistema siempre debería seguir o enumerar los comportamientos y acciones que siempre debería evitar. Por último, los sistemas inteligentes podrían adquirir los principios éticos para orientar su conducta según la situación en la que se encontrasen. En algunas situaciones, prevalecerían los principios de carácter universal, mientras que, en otras, se daría más importancia al comportamiento que a las circunstancias específicas demandan.
- **IA explicable**: en los últimos años, ha adquirido también bastante relevancia el campo de investigación denominado XAI (siglas inglesas de *Explainable AI*) (Murdoch, Singh, Kumbler, Abbasi-Asi y Yu, 2019). En este campo, se han propuesto distintos métodos para convertir muchos de los actuales sistemas de IA en arquitecturas transparentes que dispongan de mecanismos para mostrar de forma clara su funcionamiento y su razonamiento internos. Entre ellos, destacan la investigación en árboles de decisiones y redes bayesianas, en los que se estudia cómo distintos parámetros de información pueden conducir a diferentes conclusiones (Bostrom y Yudkowsky, 2014). El objetivo general es disipar la opacidad de muchos de los actuales sistemas de IA, especialmente de los dispositivos dentro del campo de *deep learning*.
- **Prueba y validación del producto**: otro método para garantizar la seguridad de los dispositivos y asegurar que su diseño y programación responda adecuadamente a lo planeado es someter los distintos dispositivos a mecanismos de examen y validación. Este tipo de pruebas permiten comprobar que los diversos componentes funcionen correctamente y sirven para detectar de forma temprana posibles fallas o consecuencias imprevistas. Estas pruebas ya han desempeñado un papel clave en

el desarrollo de algunas aplicaciones de IA. Por ejemplo, los vehículos automáticos desarrollados por varias compañías han sido sometidos durante los últimos años a distintos exámenes y comprobaciones. Los fallos y accidentes ocurridos durante estas pruebas han permitido a los ingenieros desarrollar sistemas de navegación más precisos y seguros.

4.2.2. Métodos no técnicos

Los medios técnicos de programación y diseño desempeñan un papel central a la hora de incorporar de forma eficaz y exacta todos los principios y normas necesarios para que un dispositivo inteligente opere de forma adecuada y segura en la totalidad de los escenarios en los que pueda encontrarse. Con todo, el diseño ético de algoritmos no es un método infalible y no deja de plantear diversas dificultades (Bonnefon *et al*., 2016; Hunt, 2016): la adopción de patrones éticos mediante la observación del comportamiento humano no garantiza que dichos patrones sean necesariamente los adecuados, sino simplemente los comunes y habituales; la programación de algoritmos de acuerdo a una serie de normas fijas puede, asimismo, generar conflictos en situaciones particulares que exijan una ponderación más refinada de los bienes en juego. Un vehículo autónomo programado con la máxima de no dañar a un ser humano podría perfectamente encontrarse en un escenario en el que cualquier curso de acción derivara en la pérdida de una vida humana. En escenarios de este tipo, los principios universales son de escasa utilidad.

Por ello mismo, se hace necesario también el recurso a mecanismos no técnicos que contribuyan a mantener la seguridad y fiabilidad de la IA. Dentro de esta categoría, los más relevantes son:

- **Regulación**: la labor regulatoria y legislativa puede contribuir a establecer unos parámetros de seguridad y operatividad claros y definidos. Dentro de esta categoría, los Gobiernos y agencias de regulación cuentan con medios como tratados internacionales y resoluciones, procesos de estandarización, directrices y normas no vinculantes, o contratos.
- **Certificaciones**: otra manera de garantizar la fiabilidad y seguridad de las aplicaciones provistas de IA, y de fomentar, al mismo tiempo, la confianza de los usuarios, es la de emitir certificaciones específicas por parte de las empresas desarrolladoras. Estas certificaciones podrían traducir los distintos estándares en materia de seguridad, transparencia o fiabilidad.
- **Educación y sensibilización**: la educación y la comunicación en materia de IA puede contribuir a crear una mayor conciencia en torno a los potenciales riesgos que esta tecnología entraña. Esta labor ha de alcanzar a todos los grupos de interés (diseñadores, consumidores —ya sean individuos o empresas—, reguladores, etc.) y puede hacer, a su vez, que todos ellos adopten un papel más activo en la creación y el empleo de dispositivos inteligentes.
- **Investigación**: otra manera de garantizar que la IA se siga desarrollando de manera fiable y segura es la de asegurar que la ética y el buen gobierno acompañen siempre a los temas de investigación en IA (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019, p. 199). Esto puede alcanzarse dando prioridad a estos temas de investigación en la asignación de presupuestos o incentivando el trabajo de grupos y centros de investigación que analicen qué desafíos plantea la IA a la ética, al gobierno y a la responsabilidad social de las empresas.

[...] lograr que la IA se desarrolle y emplee de manera adecuada y beneficiosa para todos exige la consideración de diversos puntos de vista y el diálogo continuo entre los distintos grupos de interés que participan en el proceso de ideación, desarrollo y empleo de esta tecnología.

La IA es un fenómeno más complejo que el simple lanzamiento o introducción de un nuevo producto en el mercado. En el desarrollo de esta tecnología se encuentran implícitos una serie de valores, objetivos y maneras de entender la sociedad y las relaciones humanas. Por ello, lograr que la IA se desarrolle y emplee de manera adecuada y beneficiosa para todos exige la consideración de diversos puntos de vista y el diálogo continuo entre los distintos grupos de interés que participan en el proceso de ideación, desarrollo y empleo de esta tecnología. De igual manera, incorporar parámetros éticos en los dispositivos inteligentes requiere la comparecencia de mecanismos de diverso tipo: además de las herramientas técnicas que los diseñadores puedan emplear, se precisa también el recurso propio de otros agentes sociales, como Gobiernos, reguladores, consumidores, centros de investigación, etc.

5. CONCLUSIONES

En este cuaderno hemos tenido la oportunidad de acercarnos al fenómeno de la IA y a su relación con la ética. En el primer apartado se ha ofrecido un breve repaso del origen y el desarrollo histórico de esta tecnología, desde su nacimiento en 1956 como campo de investigación hasta nuestros días. En el desarrollo de esta exposición salió a relucir el objetivo principal de la IA, a saber, el de replicar la inteligencia humana en máquinas y sistemas informáticos. Para llevarlo a cabo, se busca, en primer lugar, descomponer los procesos cognitivos en sus procesos más simples y elementales para poder, más tarde, expresarlos en el lenguaje formal de la lógica en forma de algoritmos y lenguajes de programación. Vimos también cómo —aun compartiendo este mismo objetivo— existen distintas definiciones de IA, las cuales divergen entre sí a la hora de especificar qué facultades cognitivas se busca emular y programar en las máquinas.

Junto con la evolución histórica de esta disciplina y con las principales definiciones dadas hasta el día de hoy, dedicamos también varias páginas a poner en claro qué relación guarda la IA con la ética. Al hacerlo, explicamos cómo la ética solo se encuentra presente allí donde comparecen agentes autónomos, es decir, individuos capaces de escoger racionalmente un curso de acción determinado. La autonomía, en este sentido ético, es por tanto un rasgo exclusivo de los seres humanos, pues solo ellos son capaces de escoger con libertad y guiar sus acciones. Los animales, en cambio, actúan por instinto. Los robots y máquinas, por otro lado, actúan de forma automática, operando siempre de acuerdo a su programación. Incluso los dispositivos más sofisticados con aparente capacidad de decisión operan en todo momento dentro de los parámetros previamente establecidos. Un vehículo autónomo parece, en apariencia, estar tomando una decisión al controlar el vehículo y escoger qué ruta seguir o qué obstáculos evitar. Pero, en realidad, dicha autonomía consiste simplemente en un cálculo —más o menos complejo— de probabilidades realizado con base en distintos parámetros, todos ellos previamente definidos y establecidos en el diseño. Los dispositivos y sistemas inteligentes poseen una clara relevancia ética, pero no por ellos mismos, sino en la medida en que han sido diseñados y programados por personas para operar de una determinada manera. La responsabilidad ante el funcionamiento de estos sistemas recae siempre sobre los distintos grupos de interés implicados en su diseño y elaboración.

En el segundo apartado hemos repasado cuáles son los principales beneficios derivados de la IA y los posibles riesgos que entraña su uso. Dentro de estos últimos, destacamos algunos más genéricos y propios de cualquier otro desarrollo tecnológico y otros específicos de la IA. Entre estos últimos, señalamos la rendición de cuentas, la explicabilidad, la imparcialidad y la privacidad como los principales desafíos que las aplicaciones provistas de IA plantean en la actualidad.

Para hacer frente a estos retos y asegurar que la IA se emplee y desarrolle de manera segura y fiable, hemos propuesto, en el último apartado, una serie de principios éticos. Se trata de principios que señalan un bien humano de especial importancia, cuya defensa y promoción debe ser tenida especialmente en cuenta durante la fase de diseño de los sistemas inteligentes. Estos principios abogan por el respeto a la autonomía humana, la transparencia, la clara asignación de responsabilidad, el diseño de sistemas seguros y robustos, y el empleo justo de estos. Para llevar a la práctica todos estos principios, existen diversos métodos y herramientas. Las relativas al proceso de diseño han quedado clasificadas como métodos técnicos, pues atañen de forma especial a los ingenieros y programadores. Junto con estos mecanismos, hemos destacado también otras herramientas, como la actividad regulatoria, las certificaciones, la investigación y la educación.

La IA ha demostrado ya poseer numerosos beneficios para las empresas, los Gobiernos, los sistemas sanitarios, los consumidores, los investigadores, etc. Es de esperar que los avances y desarrollos de los próximos años permitan a los dispositivos inteligentes adquirir nuevas funcionalidades y mejorar muchas de las tareas que actualmente realizamos. Junto con el optimismo que suscita el enorme potencial de la IA, se hace necesaria una actitud prudente que contribuya a diseñar y emplear estos dispositivos de una manera más justa, inclusiva y responsable.

6. BIBLIOGRAFÍA

Para saber más

AI Now Institute:

AI NOW REPORT 2018: ainowinstitute.org/AI_Now_2018_Report.pdf.

Comisión Europea:

THE EUROPEAN COMMISSION'S SCIENCE AND KNOWLEDGE SERVICE, *Artificial Intelligence: A European Perspective*: ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/artificial-intelligence-european-perspective.

EUROPEAN GROUP ON ETHICS IN SCIENCE AND NEW TECHNOLOGIES, *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*: ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

GRUPO INDEPENDIENTE DE EXPERTOS DE ALTO NIVEL SOBRE INTELIGENCIA ARTIFICIAL (2019), *Directrices éticas para una IA fiable*: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60423.

POLICY AND INVESTMENT RECOMMENDATIONS FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE: ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence.

Future of Life Institute:

ASILOMAR AI PRINCIPLES: utureoflife.org/ai-principles/?cn-reloaded=1.

Obra Social "la Caixa" – B·Debate:

BARCELONA DECLARATION FOR THE PROPER DEVELOPMENT AND USAGE OF ARTIFICIAL INTELLIGENCE IN EUROPE: www.iiia.csic.es/barcelonadeclaration.

The Global Initiative on Ethics of Autonomous and Intelligent Systems:

IEEE STANDARDS ASSOCIATION, *Ethically Aligned Design*: standards.ieee.org/industry-connections/ec/autonomous-systems.html.

UNICRI, Centre for Artificial Intelligence (AI) and Robotics:

ARTIFICIAL INTELLIGENCE AND ROBOTICS FOR LAW ENFORCEMENT: www.unicri.it/in_focus/on/interpol_unicri_report_ai.

Université de Montréal:

DÉCLARATION DE MONTRÉAL IA RESPONSABLE: https://www.declarationmontreal-iaresponsable.com/.

World Economic Forum:

AI GOVERNANCE: A HOLISTIC APPROACH TO IMPLEMENT ETHICS INTO AI: weforum.my.salesforce.com/sfc/p/#b0000000GycE/a/0X000000cPl1/i.8ZWL2HIR_kAnvckyqVA.nVVgrWIS4LCM1ueGy.gBc.

Fuentes consultadas

ARGANDOÑA, A. (2019). Ética e inteligencia artificial (I). I*ESE Blog Network: Economía, Ética y RSE*. Recuperado de blog.iese.edu/antonioargandona/2019/03/25/etica-e-inteligencia-artificial-i.

ASIMOV, I. (1942). Runaround. *Astounding Science Fiction*, 29(1), pp. 94-103.

BARLETT, R., MORSE, A., STANTON, R. y WALLACE, N. (2019). Consumer-Lending Discrimination in the FinTech Era. *National Bureau of Economic Research*. doi: 10.3386/w25943.

BELLMAN, R. E. (1978). *An Introduction to Artificial Intelligence: Can Computers Think?* San Francisco: Boyd & Fraser Publishing Company.

BERRIMAN, R., HAWKSWORTH, J. y GOEL, S. (2017). Will robots really steal our jobs? An international analysis of the potential long term impact of automation. PWC. Recuperado de www.pwc.com/hu/hu/kiadvanyok/assets/pdf/impact_of_automation_on_jobs.pdf.

BIJKER, W., HUGHES, T. y PINCH, T. (eds.). (1987). *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge: MIT Press.

BONNEFON, J.-F., SHARIF, A. y RAHWAN, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293), pp. 1573-1576. doi: 10.1126/science.aaf2654.

BOSTROM, N. y YUDKOWSKY, E. (2014). The ethics of artificial intelligence. *The Cambridge Handbook of Artificial Intelligence*, pp. 316-334. Machine Intelligence Research Institute, Cambridge University Press. doi: 10.1017/CBO9781139046855.020.

BRUNDAGE, M., AVIN, S., CLARK, J., TONER, H., ECKERSLEY, P., GARFINKEL, B., … ANDERSON, H. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv preprint arXiv:1802.07228.

BUCHHOLZ, R. A. y ROSENTHAL, S. B. (2002). Technology and Business: Rethinking the Moral Dilemma. *Journal of Business Ethics*, 41(1-2), pp. 45-50.

BUOLAMWINI, J. y GEBRU, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification*. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, *PMLR 81*, pp. 77-91.

CATH, C., WACHTER, S., MITTELSTADT, B., TADDEO, M. y FLORIDI, L. (2018). Artificial Intelligence and the `Good Society´: the US, EU, and UK approach. *Science and Engineering Ethics*, 24(2), pp. 505-528. doi: 10.1007/s11948-017-9901-7.

CHARNIAK, E. y MCDERMOTT, D. (1985). *Introduction to Artificial Intelligence*. Reading (Massachusetts): Addison-Wesley.

CLANCEY, W. J. (1997). *Situated Cognition: On Human Knowledge and Computer Representations*. Cambridge: Cambridge University Press.

CLANCEY, W. J. (1999). *Conceptual Coordination: How the Mind Orders Experience in Time*. Nueva Jersey: Lawrence Erlbaum Associates.

COMISIÓN EUROPEA, COM (2018a). A*rtificial intelligence for Europe*. 237 final, 1. Recuperado de ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe.

COMISIÓN EUROPEA, SWD (2018b). *Liability for emerging digital technologies*. 137 final, 1. Recuperado de ec.europa.eu/digital-single-market/en/news/european-commission-staff-working-document-liability-emerging-digital-technologies.

COMISIÓN EUROPEA, Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial (2019), *Directrices éticas para una IA fiable*: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60423

CRAGLIA, M. (ed.), ANNONI, A., BENCZUR, P., BERTOLDI, P., DELIPETREV, P., DE PRATO, G., FEIJOO, C.,… VESNIC, L. (2018). *Artificial Intelligence: A European Perspective*, EUR 29425 EN. Luxemburgo: Oficina de Publicaciones. doi: 10.2760/11251.

CREVIER, D. (1993). *The Tumultuous History of the Search for Artificial Intelligence*. Nueva York: Basic Books.

DELOITTE (2019). T*MT Predictions 2019: What does the future hold for technology, media, and telecommunications?* Recuperado de www2.deloitte.com/insights/us/en/industry/technology/technology-media-and-telecom-predictions.html?id=gx:2el:3pr:4di7888:5awa:6di:MMDDYY:predictions2019&pkid=1005674.

DREYFUS, H. L. (1972). *What Computers Can't Do: The Limits of Artificial Intelligence*. Nueva York: Harper & Row Publishers.

DREYFUS, H. L. (1994). *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge (Massachusetts): MIT Press.

EDELMAN, G. M. (1987). *Neural Darwinism: The Theory of Neural Group Selection*. Nueva York: Basic Books.

EUROPEAN GROUP ON ETHICS IN SCIENCE AND NEW TECHNOLOGIES (2018). *Statement on Artificial Intelligence, Robotics and `Autonomous´ Systems*. Recuperado de ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

EUROPEAN PARLIAMENTARY RESEARCH SERVICE, SCIENTIFIC FORESIGHT UNIT (STOA) (2016). *Ethical Aspects of Cyber-Physical Systems*. Recuperado de www.europarl.europa.eu/RegData/etudes/STUD/2016/563501/EPRS_STU%282016%29563501_EN.pdf.

EYAL, N. (2017). Here's How Amazon's Alexa Hooks You. *INC*. Recuperado de www.inc.com/nir-eyal/heres-how-amazons-alexa-hooks-you.html.

ETZIONI, A. y ETZIONI, O. (2017). Incorporating ethics into artificial intelligence. The *Journal of Ethics*, 21(4), pp. 403-418.

FANUC ROBOTICS EUROPE, S. A. (2012). *68 Robots Perform Farmer's Work*. Estudio de caso, International Federation of Robotics.

FINNIS, J. (2011). *Natural Law and Natural Rights* (2.ª edición). Nueva York: Oxford University Press.

FUSSELL, S. (2019). Consumer Surveillance Enters Its Bargaining Phase. *The Atlantic*. Recuperado de www.theatlantic.com/technology/archive/2019/06/alexa-google-incognito-mode-not-real-privacy/590734.

GARDNER, H. (2004). *Frames of Mind: The Theory of Multiple Intelligences*. Nueva York: Basic Books.

GOLDHILL, O. (2015). Why smartphones are making you ill. *The Telegraph.* Recuperado de www.telegraph.co.uk/technology/news/11532428/Why-smartphones-are-making-you-ill.html.

GREENEMEIER, L. (2011). Piece of Mind: Is the Internet Replacing Our Ability to Remember? *Scientific American*. Recuperado de www.scientificamerican.com/article/internet-transactive-memory.

GROTH, O., NITZBERG, M. y ESPOSITO, M. (2018). Rules for Robots. *The Digital Future*. Recuperado de www.kas.de/c/document_library/get_file?uuid=1a5564f5-77ea-a5bc-228b-279d885c313b&groupId=252038.

HEID, M. (2018). There's Worrying New Research About Kids' Screen Time and Their Mental Health. *Time*. Recuperado de time.com/5437607/smartphones-teens-mental-health.

HOFSTADTER, D. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. Nueva York: Basic Books.

HUNT, E. (2016). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. *The Guardian*. Recuperado de www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter.

IBERDROLA (2019). ¿Somos conscientes de los retos y principales aplicaciones de la inteligencia artificial?. *Innovación*. Recuperado de www.iberdrola.com/innovacion/que-es-inteligencia-artificial.

JASANOFF, S. (2013). Technologies of Humility: Citizen Participation in Governing Science. *Minerva*, 41(3), pp. 223-244. doi: 10.1023/A:1025557512320.

KANT, I. (1785). *Grundlegung zur Metaphysik der Sitten*. Riga

KOO, C. W., ANAND, V., GIRVIN, F., WICKSTROM, M. L., FANTAUZZI, J. P., BOGONI, L. … KO, J. P. (2012). Improved Efficiency of CT Interpretation Using an Automated Lung Nodule Matching Program. *American Journal of Roentgenology*, 199(1), pp. 91-95. doi: 10.2214/ajr.11.7522.

KURZWEIL, R. (1990). T*he Age of Intelligent Machines*. Cambridge (Massachusetts): MIT Press.

LEIBNIZ, G.W. (1923). *Sämtliche Schriften und Briefe*, Berlin: Akademie Verlag.

LIN, P., ABNEY, K. y BEKEY, G. (2011). Robot ethics: Mapping the issues for a mechanized world. *Artificial Intelligence*, 175(5-6), pp. 942-949. doi: 10.1016/j.artint.2010.11.026.

LÓPEZ DE MÁNTARAS, R. (2015). Algunas reflexiones sobre el presente y futuro de la Inteligencia Artificial. *Novática*, 234(4), pp. 97-101. Recuperado de hdl.handle.net/10261/136978.

MALONE, T. W. (2018). *Superminds: The Surprising Power of People and Computers Thinking Together*. Nueva York: Little, Brown & Company.

MARTIN, K. E. y FREEMAN, R. E. (2004). The Separation of Technology and Ethics in Business Ethics. *Journal of Business Ethics*, 53(4), pp. 353-364. doi: 10.1023/b:busi.0000043492.42150.b6.

MCCARTHY, J., MINSKY, M. L., ROCHESTER, N. y SHANNON, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI magazine*, 27(4), p. 12. doi: 10.1609/aimag.v27i4.1904.

MCCORDUCK, P. (2004). *Machines Who Think* (2.ª edición). Natick (Massachusetts): A K Peters/CRC Press.

MCDERMOTT, D. (1982). R1: A rule-based configurer of computer systems. *Artificial Intelligence*, 19(1), pp. 39-88. doi: 10.1016/0004-3702(82)90021-2.

MCKINSEY & COMPANY (2018). *AI adoption advances, but foundational barriers remain*. Recuperado de www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain.

MCKINSEY GLOBAL INSTITUTE (2013). *Disruptive technologies: Advances that will transform life, business, and the global economy*. Recuperado de www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Disruptive%20technologies/MGI_Disruptive_technologies_Full_report_May2013.ashx.

MIRA, J., DELGADO, A. E., BOTICARIO, J. G. y DÍEZ, F. J. (1995). *Aspectos básicos de la inteligencia artificial*. Sanz y Torres.

MITTELSTADT, B. D., ALLO, P., TADDEO, M., WATCHER, S. y FLORIDI, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 205395171667967. doi: 10.1177/2053951716679679.

MURDOCH, W. J., SINGH, CH., KUMBLER, K., ABBASI-ASI, R y YU, B. (2019). *Interpretable machine learning: definitions, methods, and applications*. En prensa. arXiv:1901.04592.

OCDE (2017). *Algorithms and collusion: Competition policy in the digital age*. Recuperado de www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.htm.

POLONSKI, V. (2017). *Artificial intelligence can save democracy unless it destroys it first*. Oxford Internet Institute. Recuperado de www.oii.ox.ac.uk/blog/artificial-intelligence-can-save-democracy-unless-it-destroys-it-first.

POOLE, D., MACKWORTH, A. K. y GOEBEL, R. (1998). *Computational Intelligence: A Logical Approach*. Nueva York: Oxford University Press.

PwC (2019). 2019 *AI predictions: Six AI priorities you can't afford to ignore*. Recuperado de www.pwc.com/us/en/services/consulting/library/artificial-intelligence-predictions-2019.

QUINN, M. (2015). *Ethics for the information age* (6.ª edición). Harlow: Pearson.

ROQUE, M. y PALMA. J. (2008). *Inteligencia Artificial. Técnicas, métodos y aplicaciones*. Madrid: McGraw-Hill.

ROSENBERG, D. (2018). How Marketers Can Start Integrating AI in Their Work. *Harvard Business Review*. Recuperado de hbr.org/2018/05/how-marketers-can-start-integrating-ai-in-their-work.

RUSSELL, S. y NORVIG, P. (2016). *Artificial Intelligence: A Modern Approach* (3.ª edición). Malasia: Pearson Education Limited.

SEARLE, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), pp. 417-457. doi: 10.1017/S0140525X00005756.

SEARLE, J. (1987). Minds and Brains Without Programs. En BLAKEMORE, C. y GREENFIELD, S. (eds.), *Mindwaves*. Oxford: Basil Blackwell.

SHOHAM, Y., PERRAULT, R. BRYNJOLFSSON, E., CLARK, J., MANYIKA, J., NIEBLES, J. C., … BAUER, Z. (diciembre del 2018). *The AI Index 2018 Annual Report*, AI Index Steering Comittee, Human-Centered AI Initiative, Standford University, Standford, CA. Recuperado de cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf.

SHULEVITZ, J. (2018). Alexa, Should We Trust You? *The Atlantic*. Recuperado de www.theatlantic.com/magazine/archive/2018/11/alexa-how-will-you-change-us/570844.

SIMON, H. A. (1965). T*he Shape of Automation for Men and Management*. Nueva York: Harper & Row.

STERBA, J. P. (ed.). (2009). *Ethics: The Big Questions*. Reino Unido: John Wiley & Sons.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. Recuperado de standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf.

UK INFORMATION COMMISSIONER'S OFFICE (2017). *Big data, artificial intelligence, machine learning and data protection*, versión 2.2., rec. 8. Recuperado de ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

VERBEEK, P.-P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press

WAKABAYASHI, D. (2018). Self-driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam. *The New York Times*. Recuperado de www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

WINSTON, P. H. (1992). *Artificial Intelligence*. Wilmington (Delaware): Addison-Wesley.

WILSON, J.H. y DAUGHERTY, P.R. (2018), Collaborative Intelligence: Humans and AI Are Joining Forces. *Harvard Business Reviews*, 96(4), pp. 114-123.

WORLD ECONOMIC FORUM (2016). *The Future of Jobs: Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution*. Recuperado de www3.weforum.org/docs/WEF_FOJ_Executive_Summary_Jobs.pdf.

IESE Business School
University of Navarra

Cátedra CaixaBank
de Responsabilidad
Social Corporativa

Av. Pearson, 21
08034 Barcelona
Tel.: 93 253 42 00
Fax: 93 253 43 43

www.iese.edu