COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

Inteligencia artificial y ética

Clave: 201502309

MADRID | Junio 2020

Resumen ejecutivo:

A lo largo de los últimos años hemos visto como tomaba forma la cuarta revolución tecnológica, dirigiendo nuestra atención a nuevos conceptos como eran la inteligencia artificial o el *"machine learning".* El propósito de este trabajo es discernir si todas las ventajas y beneficios que traen estos avances a la sociedad global, vienen con un coste ético mucho mayor del que la sociedad debería estar dispuesta a pagar, o, si por el contrario, ese coste ético no es tan grande. Para ello, en primer lugar, en este trabajo se analizará el contexto actual y se definirá qué es la Cuarta Revolución Industrial y sus implicaciones. En segundo lugar, se centrará en la definición del concepto de Inteligencia Artificial, desde un punto de vista tanto teórico como práctico, para acabar respondiendo una pregunta: ¿qué implica realmente para las sociedades vivir entre algoritmos que toman pequeñas decisiones por las personas? En tercer lugar, a raíz de la respuesta a la anterior pregunta planteada, analizará los posibles problemas éticos si esta tecnología sigue avanzando al mismo ritmo, y los actuales problemas de sesgos éticos. Además, de estudiar los problemas, también estudiará las posibles soluciones, tanto las que ya se están dando, como las que habrá que desarrollar en el futuro, y sus limitaciones. En conclusión, se observará hasta qué punto el ser humano controla y puede llegar a controlar la utilización y preferencias de una tecnología que, a pesar de ser creada por seres humanos, puede llegar a sobrepasarlos.

Palabras clave: Ética, inteligencia artificial, tecnología, individuo, riesgos, impacto

Abstract:

Over the last few years we have seen the fourth technological revolution taking shape, directing our attention to new concepts such as artificial intelligence or machine learning. The purpose of this work is to discern whether all the advantages and benefits that these advances bring to the global society, come with a much higher ethical cost than the society should be willing to pay, or, if, on the contrary, that ethical cost does not is so big. To do this, firstly, this paper will analyze the current context and define what the Fourth Industrial Revolution is and its implications. Secondly, it will focus on the definition of the concept of Artificial Intelligence, from both a theoretical and a practical point of view, to end up answering a question: what does it really mean for societies to live between algorithms that make small decisions for people? Thirdly, following the answer to the previous question asked, it will analyze the possible ethical problems if this technology continues to advance at the same rate, and the current problems of ethical biases. In addition to studying the problems, you will also study the possible solutions, both those that are already occurring and those that will have to be developed in the future, and their limitations. In conclusion, it will be observed to what extent the human being controls and can control the use and preferences of a technology that, despite being created by human beings, can exceed them.

Key words: Ethics, artificial intelligence, technology, individual, risks, impact

Tabla de contenido

1. Introducción y finalidad del trabajo

A lo largo de los últimos años se ha visto cómo tomaba forma la Cuarta Revolución tecnológica, dirigiendo la atención a nuevos conceptos como eran la inteligencia artificial o el "*machine learning*". El propósito de este trabajo es discernir si todas las ventajas y beneficios que traen estos avances a la sociedad global, vienen con un coste ético mucho mayor del que la sociedad debería estar dispuesta a pagar, o, si por el contrario, ese coste ético no es tan grande.

A priori parece que los campos de la ética y la inteligencia artificial no tienen nada que ver el uno con el otro, pero, según va evolucionando la inteligencia artificial se han ido planteando numerosos dilemas que mucho tienen que ver con la ética. ¿Dónde está el límite del desarrollo de la inteligencia artificial? ¿llegará un día que las máquinas superen a los humanos? ¿está bien que sea la tecnología la que tome decisiones por las personas? ¿qué es el comportamiento ético de un sistema inteligente? ¿deben tener derechos y ser sujetos de responsabilidad? ¿dónde queda la privacidad del individuo?

Estas preguntas y muchas más hacen necesaria la relación entre la ética y la inteligencia artificial, pues aunque esta última puede ser una gran fuente de ventajas y oportunidades, también puede suponer, en cierta medida, la deshumanización de los seres humanos. Si las interacciones sociales se sustituyen por tecnología, si las decisiones las toman sistemas inteligentes, si la privacidad y la libertad de pensamiento desaparece, ¿qué quedará de lo que se entiende como sociedad humana? El desarrollo tecnológico es necesario, pero también su regulación y limitación. No se deben obviar los potenciales riesgos y peligros que puede entrañar, no sólo a nivel personal, sino económico, medioambiental y político. La tecnología no es mala en sí misma, pero dado que los usuarios son seres humanos con intereses propios y sin neutralidad, la tecnología debe ser desarrollada y utilizada bajo unos principios éticos básicos.

Este trabajo tiene como objetivo, primero, explicar los conceptos que se interrelacionan, el de ética y el de inteligencia artificial. También, exponer el contexto en el que se han desarrollado y creado estos conceptos, esto es necesario para poder analizar después la relación y los riesgos. Segundo, responder a la pregunta de: ¿qué implica realmente para las sociedades vivir entre algoritmos que toman pequeñas decisiones por

las personas día a día? Tercero, establecer la relación, y, la necesidad de la misma, entre la ética y la inteligencia artificial. Analizar los potenciales retos y peligros si la tecnología sigue avanzando a un ritmo acelerado sin regulación. Por último, proponer una línea de actuación y posibles actitudes que puedan servir de preparación antes de que el desarrollo de la inteligencia artificial llegué a su punto álgido.

Personalmente, he decidido realizar este trabajo porque me parece un tema actual, que será muy relevante en un futuro, y, que fuera del campo de estudio propio se conoce poco. Me parece fundamental ser consciente, a nivel personal, de los posibles riesgos, actuar de manera responsable y saber qué está ocurriendo a mi alrededor. Ya que vivimos rodeados de tecnología y la utilizamos a diario para prácticamente cualquier cosa, creo que se debería educar más a la sociedad sobre el uso responsable y ético, además de a qué nos exponemos al utilizar estos dispositivos. Por esto, y por mi interés y curiosidad sobre el tema he decidido escoger este tema.

2. Metodología, estructura y estado de la cuestión

Este punto expondrá de forma breve la metodología seguida para la realización de este trabajo, la estructura que sigue el escrito y hará una revisión de las principales fuentes.

Metodología

La metodología de este trabajo se basa en la recopilación, síntesis y análisis de diversas fuentes históricas y académicas. El método que sigue este trabajo de fin de grado es deductivo, a través de la revisión de literatura tanto sobre ética, como sobre inteligencia artificial. El primer paso fue la búsqueda de información, he recurrido a numerosas bases de datos, con acceso libre dada la situación del Covid-19, como Dialnet, JSTOR, Annual Reviews, Proquest o Google Scholar. También he utilizado apuntes de la propia universidad de la asignatura de Ética empresarial impartida este último curso. Además, he utilizado diversos libros consultados a través de Google Books o de posesión propia. Estas fuentes me han aportado diversos artículos y secciones de libros fundamentales para el desarrollo de este trabajo. El segundo paso fue la síntesis de toda la información recabada, para quedarme exclusivamente con los puntos más importantes y que realmente podían añadir valor a este trabajo. Por último, procedí a realizar el contexto y marco

teórico y a analizar la relación entre la ética y la inteligencia artificial, que llevan a las conclusiones del mismo trabajo.

Estructura

Este trabajo está dividido en seis puntos que van profundizando poco a poco en el tema. El marco teórico está compuesto por un punto sobre el concepto de la ética en sí, necesario para luego realizar la relación entre este concepto y la inteligencia artificial. Para establecer el contexto y los antecedentes de la inteligencia artificial se realiza un punto sobre la Cuarta Revolución Industrial, sus orígenes y sus tecnologías. Antes de entrar en el análisis de la relación entre la ética y la inteligencia artificial, se expone en el tercer punto el origen de la inteligencia artificial, sus distintas definiciones y tipos y el impacto que ha tenido a todos los niveles a nivel global. Los dos siguientes puntos son el análisis en sí. Se comienza dando la razón de por qué están relacionadas la inteligencia artificial y la ética y por qué es necesaria esa relación, además, se establecen los principios éticos que debe seguir el desarrollo de la inteligencia artificial y los métodos para implementarlos. A continuación, en el punto previo a las conclusiones, se explican los principales riesgos prácticos de la introducción de la inteligencia artificial en la vida cotidiana, y, también, se realiza una reflexión sobre en qué punto se encuentra el desarrollo de estos sistemas y en qué dirección irá su evolución. Por último, el punto de conclusiones recoge las principales de este escrito y pone fin al mismo.

Estado de la cuestión

A la hora de realizar este trabajo se ha utilizado una amplia bibliografía científica, histórica, académica y periodística. Las fuentes están centradas por apartados, en la historia de la Cuarte Revolución Industrial, la historia de la inteligencia artificial, los riesgos y el impacto de la inteligencia artificial, la relación de la ética y la inteligencia artificial, los retos éticos de la misma y el concepto de ética en sí mismo. Por ello, las fuentes son de diversa procedencia, desde estudios científicos y tecnológicos, análisis éticos sobre las implicaciones, artículos periodísticos sobre el impacto, hasta fuentes históricas para el contexto.

Para el desarrollo de este trabajo ha sido fundamental la obtención de diversos informes sobre los temas abordados aquí, las fuentes principales de estos informes son la

OECD, Mckinsey&Co, la Comisión Europea, Iberdrola, la cátedra de responsabilidad social corporativa del IESE y el CSIC.

Las fuentes académicas han sido escogidas de entre los autores más relevantes en estos campos. Estos son McCarthy, Haenlein, Kaplan o Bostrom, de entre otros muchos. De McCarthy destacan los documentos sobre la inteligencia artificial, así como el documento de la Conferencia de Dartmouth, mientras que autores como Haenlein, Kaplan o Bostrom están más enfocados al estudio de la relación entre la ética y la inteligencia artificial, así como los riesgos de esta última.

Como fuentes académicas también han sido muy importantes dos libros, *"Artificial Intelligence A Modern Approach"* de los autores Russel y Norvig, y, *"The 4th Industrial Revolution. Responding to the Impact of Artificial Intelligence on Business"* de Skilton y Hovsepian. Ambos abarcan desde el estudio histórico, el científico, hasta la perspectiva social y ética.

Por último, también ha sido de mucha utilidad la convención organizada por la Asociación de mujeres abogadas europeas, en Madrid el pasado noviembre (2019), sobre la Cuarta Revolución Industrial y la ética.

Este apartado sobre el estado de la cuestión cierra la parte introductoria del trabajo para pasar al marco teórico, y, posteriormente el análisis.

3. El concepto de ética

Este escrito tiene como objetivo desarrollar la relación existente entre la ética y una de las tecnologías más importantes de la Cuarta Revolución Industrial, la inteligencia artificial. Para comenzar, en este apartado, a modo de marco teórico, se plantea el significado y definición del concepto de ética.

La ética se define de manera general como: *"conjunto de normas morales que rigen la conducta de la persona encualquier ámbito de la vida"* o *"parte de la filosofía que trata del bien y del fundamento de sus valores"* (RAE, 2020). La ética es un concepto que puede ser definido desde muchos puntos de vista y campos de estudio.

Existen tres niveles de estudio en la ética, la *metaética* enfocada a la naturaleza y al significado de la ética en sí misma, la ética normativa centrada en la búsqueda de sistemas normativos que lleven al ser humano a la mejor vida posible, y, la ética aplicada que consiste en interpretar cuestiones específicas (Raffino, 2020). En el caso de la inteligencia artificial, el nivel de estudio es el de ética aplicada y ética normativa. Busca responder a cuestiones del día a día sobre la interactuación de los hombres con las máquinas, poner límites morales a su uso, y, establecer unas normas para que no se rebasen dichos límites.

Estas normas éticas llevan presentes desde el principio de la historia, evolucionando a la par que la forma de vida del ser humano. En la Grecia clásica se presentan los primeros estudios sobre la conducta humana y bajo que principios se regía. Más adelante la religión se confundió con la ética, al establecer los dogmas religiosos cómo debía comportarse un buen individuo. Es a partir de la Edad Moderna cuando llega un nuevo modelo ético más referido a la razón de la propia persona, más parecido al de hoy en día (Raffino, 2020).

A la vez que han ido evolucionando las áreas de las sociedades, la ética ha ido ocupando su lugar en cada una de ellas. Por ello, existen varios tipos de ética actualmente. Primero, la ética profesional que analiza el ejercicio de las profesiones mismas, así como adjudica responsabilidades por actuaciones poco morales. Segundo, la ética militar que regula la ética de los conflictos. Tercero, la ética económica que mide las consecuencias y el comportamiento justo dentro de los mercados. Por último, la ética religiosa que sigue la tradición que mezcla las ideas religiosas con la moral, un ejemplo muy claro es el cristianismo (Raffino, 2020). La llegada de la Cuarta Revolución Industrial y la inteligencia artificial han supuesto que la ética se implique en sus avances, puesto que estos plantean dilemas y producen grandes cambios sociales.

La moral no es lo mismo que la ética, aunque ambos conceptos están estrechamente relacionados. La ética estudia la moral, pone códigos de comportamiento para que estos sean morales, pero, la moral funciona en términos absolutos de bien o mal que la ética no entra a juzgar. La ética como rige comportamientos sociales y la moral rige los comportamientos del individuo (Ceballos, 2020). Por lo que en el ámbito de la inteligencia artificial es la ética la que analiza la interacción con la sociedad, lo

moralmente bueno o no ya es cuestión individual, no a nivel general de la inteligencia artificial.

Uno de los grandes problemas de la ética, que aplica perfectamente a la situación de la inteligencia artificial, es que el valor ético de las acciones y la validación social de las mismas no suelen coincidir. El valor ético es la relación directa entre la acción y su naturaleza, mientras que la validación social tiene que ver con la percepción que la sociedad tiene de la misma (Ceballos, 2020). Por ejemplo, la percepción sobre el uso indebido de datos es la normalización de esta acción, pero, claramente el valor ético es negativo.

Sin embargo, no es tarea de la ética juzgar, sino guiar. Los seres humanos no tienen limitación de actuación, pero, ¿cómo se decide que es lo que debe hacerse y lo que no? Por la naturaleza humana. La naturaleza humana se autorregula para no actuar de determinadas maneras que tienen consecuencias contraproducentes que llevan a la deshumanización. La ética en todos los casos plantea una guía de actuación, el cumplimiento o no de la misma es decisión individual (Ceballos, 2020).

Tras esta breve introducción sobre el concepto de ética y unas primeras relaciones con la inteligencia artificial, se procede a empezar el contexto de estas nuevas tecnologías. Para ello, el siguiente punto introduce la Cuarta Revolución Industrial y sus tecnologías.

4. La Cuarta Revolución Industrial

En este punto se hace un breve resumen de la historia de la Cuarta Revolución Industrial y sus tecnologías, para poner en contexto la inteligencia artificial y pasar en el siguiente punto a un análisis más profundo de la misma.

El término de Cuarta Revolución Industrial, o, industria 4.0, lo acuñó Klaus Schwab en 2016. El fundador del Foro Económico Mundial la define como: *"industria que genera un mundo en el que los sistemas de fabricación virtuales y físicos cooperan entre sí de una manera flexible a nivel global"* (Iberdrola, 2020). Es la fusión de tecnologías más allá de sistemas inteligentes y conectados, como la nanotecnología, la robótica o la computación cuántica. Todas estas áreas interactúan a nivel físico y digital,

este componente digital y sus nuevas tecnologías son lo que diferencian a la Cuarta Revolución Industrial de las anteriores (Iberdrola, 2020).

Cada revolución industrial ha supuesto cambios sociales y económicos a nivel global. La primera revolución industrial, a finales del siglo XVIII, introdujo la mecánica en el mundo laboral al aplicar el vapor a la producción mecánica. La segunda, a finales del siglo XIX, introdujo la electricidad a la producción creando las cadenas de montaje y provocando la aceleración industrial. La tercera, a finales del siglo XX, se adentra en la programación de las máquinas y da comienzo a la automatización. La Cuarta Revolución Industrial se puede decir que empezó en 2014, al surgir las fábricas inteligentes y la gestión online de la producción (Iberdrola, 2020).

Pero, esta revolución es diferente. En el libro de Schwab "La Cuarta Revolución Industrial" se dice*: "Estamos al borde de una revolución tecnológica que modificará la forma en que vivimos, trabajamos y nos relacionamos. En una escala de alcance y complejidad la transformación será diferente a cualquier cosa que el género humano haya experimentado antes"*. Esto ocurre por la velocidad, el alcance y el impacto que tienen las nuevas tecnologías.

Las principales tecnologías de esta cuarta revolución son: (1) la inteligencia artificial, desarrollada a fondo en siguientes apartados. (2) El internet de las cosas que conecta el mundo físico y el digital permitiendo que los dispositivos interactúen con objetos y estos se vuelvan "inteligentes". (3) Cobots, interactúan con los humanos y optimizan muchas tareas monótonas de producción. (4) Realidad aumentada y realidad virtual, combinan la realidad con el mundo digital. (5) Big data, gestión y análisis de cantidades masivas de datos. (6) Impresión 3D y 4D, referido a la impresión de prototipos físicos (Iberdrola, 2020).

Pese a todas las ventajas que ofrecen estas nuevas tecnologías, esta revolución no está exenta de debates. Una de las cuestiones que han generado más debate es si todo desarrollo tecnológico debería ser llevado a cabo. Esto plantea la duda de si aún teniendo la posibilidad de avanzar tecnológicamente en un campo, se debería hacer realidad a cualquier coste. Cualquier creación y avance del ser humano puede ser utilizado de manera beneficiosa o nociva, incluyendo todas las tecnologías desarrolladas en la Cuarta

Revolución Industrial. Existe la visión de que cualquier avance es positivo, simplemente por el hecho de la evolución que implica, pero, hay que recordar que la tecnología pierde su neutralidad en el momento en el que es programada para objetivos y usos concretos por un ser humano. Los seres humanos no son neutrales y a lo largo de la historia se ha visto como "nuevas tecnologías" que iban a revolucionar de manera positiva el mundo han acabado siendo la pesadilla de sociedades. Ejemplos de esto son la creación de la pólvora que derivó en armas de fuego, o, el conocimiento de la energía nuclear que acabó en la creación de las armas nucleares. *"Que la tecnología se preste a diversos usos no equivale a afirmar que sea en si misma neutral, ni a que todo desarrollo técnico sea aconsejable"* (Marín García, 2019) (Martin & Freeman, 2004).

Tras esta breve reflexión sobre la Cuarta Revolución Industrial, se procede a ahondar en la inteligencia artificial en el siguiente punto.

5. La Inteligencia Artificial

En este punto se analizará el origen histórico del concepto inteligencia artificial (IA), su definición y dificultades de aplicación, los tipos de inteligencia artificial que existen, y, por último, el impacto positivo y negativo que han tenido estos avances en la sociedad. El objetivo es establecer un marco teórico claro, que permita posteriormente aplicar la ética en los siguientes puntos del escrito.

5.1. Origen y concepto

La inteligencia artificial es presentada por muchos medios de comunicación como algo extraordinario y desde un punto de vista de optimismo generalizado, creando unas expectativas desmedidas de sus posibilidades actuales, y, que contribuyen a ver la IA como algo ajeno, extraño y lejano. Pero nada más lejos de la realidad, la IA influye y convive con los seres humanos a diario en muchas de sus actividades cotidianas (Marín García, 2019).

Por otro lado, están los que observan este fenómeno desde una perspectiva fatalista, en la que los riesgos y peligros superan a los beneficios que pueda aportar la inteligencia artificial. Existe un amplio campo de investigación sobre el peligro que

pueden entrañar estas tecnologías para las personas, y, aunque está claro que estas preocupaciones pueden ser exageradas, también es cierto que al mismo tiempo que avanzan las aplicaciones prácticas de la inteligencia artificial lo hacen sus retos éticos (Marín García, 2019).

"La inteligencia artificial – como el resto de las tecnologías diseñadas por el ser humano – puede derivar en aplicaciones nocivas o beneficiosas para las personas" (Argandoña, 2019). Aún así, la IA es algo formidable que muestra las grandes capacidades inventivas y creativas de los seres humanos, sirve para mejorar el nivel de vida a muchos niveles y tiene grandes beneficios sobre las sociedades. Que pueda ser utilizada "mal" tiene más que ver con que los humanos son los que deciden cómo emplearla que con un peligro intrínseco de la inteligencia artificial.

Esta "tecnología" es comúnmente definida como la capacidad de crear sistemas y dispositivos con las mismas capacidades cognitivas del ser humano, es decir, una "imitación" de la inteligencia del ser humano (Charniak & Mcdermott, 1985). En siguientes apartados se expondrán las dificultades a la hora de definir el término, puesto que no existe un consenso completo. Pero, ¿cuál es el origen del término? ¿Cuáles son los antecedentes y orígenes de esta tecnología tan disruptiva?

El primer antecedente conocido, curiosamente, no es ningún estudio informático o experimento, sino referencias en la literatura. El escritor de ciencia ficción Isaac Asimov publicó en 1942 su cuento *"Círculo vicioso"*, en el que describía robots desarrollados por ingenieros que seguían ciertas pautas. Esto sirvió como inspiración para muchas de las siguientes generaciones de científicos que estudiaban la inteligencia artificial y la robótica (Haenlein & Kaplan, 2019). Sin embargo, no todo lo relativo a este ámbito en aquella época era ciencia ficción, en 1942 Alan Turing inventó su máquina concida como "El bombe". Alan Turing, cuyo trabajo estuvo inspirado en el de Gödel y en el de Gottfried Leibniz, consiguió crear una máquina que descifraba el código "enigma" de los alemanes durante la Segunda Guerra Mundial. Esta máquina es considerada el primer ordenador electromecánico y supuso la realización de la idea de que una máquina podía llevar a cabo una tarea cuasi imposible para cualquier ser humano. Bajo esta premisa, en 1950, Turing publicó un artículo llamado *"Maquinaria informática e inteligencia"* en el que expuso como crear máquinas "inteligentes" y como testear su

inteligencia. Esto es conocido como el "test de Turing"[1], y, hoy en día, sigue siendo utilizado como punto de referencia (Skilton & Hovsepian, 2018) (Haenlein & Kaplan, 2019).

Otro de los trabajos reconocidos actualmente como precursores de la inteligencia artificial es el de Warren McCulloch y Walter Pitts. Juntos, basándose en varias de las teorías de Turing, propusieron un modelo de neuronas artificiales en el cual cada una de ellas estaría "apagada" o "encendida" según el estado de sus neuronas adyacentes. Proponían que cualquier función computable podía llevarse a cabo por una red de neuronas conectadas. Esas redes de neuronas podían crearse de manera artificial, y, llegaron a sugerir, que podrían incluso aprender por si mismas (Russel & Norvig, 2016).

El término inteligencia artificial fue acuñado en 1956 durante la Conferencia de Dartmouth. El informático John McCarthy, junto con Marvin Minsky, Nathaniel Rochester y Claude Shannon, reunió en la Conferencia de Dartmouth a los investigadores más relevantes sobre informática y psicología cognitiva de Estados Unidos con la intención de trabajar sobre un nuevo concepto, la *"artificial intelligence"* (Russel & Norvig, 2016). La propuesta del proyecto de verano de Darthmouth sobre inteligencia artificial decía:

"Proponemos que se lleve a cabo un estudio de dos meses, realizado por diez hombres sobre la inteligencia artificial durante el verano de 1956 en la Universidad de Dartmouth en Hanover. El estudio se basa en la hipótesis de que cualquier aspecto del aprendizaje o cualquier otra característica de la inteligencia puede ser, en principio, descrita de manera tan precisa que una máquina sea capaz de simularla. Un objetivo será el de averiguar como hacer que las máquinas utilicen el lenguaje y conceptos abstractos, sean capaces de resolver problemas hasta ahora reservados para los humanos y de mejorarse a sí mismas. Pensamos que se puede conseguir un avance significativo en más de uno de estas áreas si se selecciona de manera adecuada a los investigadores" (McCarthy, Minsky, Rochester, & Shannon, 2006).

[1] Establece que si un humano interactuando con otro humano y una máquina no puede distinguir al humano de la máquina, la máquina es "inteligente".

En esa misma propuesta también establecieron los principales problemas de la inteligencia artificial:

1. *Ordenadores automáticos: el problema no es la falta de capacidad del ordenador, sino la incapacidad humana de escribir los programas adecuados.*
2. *Cómo puede un ordenador ser programado para usar el lenguaje: necesidad de crear nuevas "reglas".*
3. *Redes neuronales: cómo puede un grupo de hipotéticas neuronas ser programadas para que creen conceptos.*
4. *La teoría del tamaño de un cálculo: al tener un problema definido, una manera de resolverlo es comprobar todas las posibles respuestas en orden. Esto es muy ineficiente, por lo que se tratará de crear un método que lo resuelva.*
5. *Mejora de sí mismos: probablemente una máquina inteligente podrá llevar a cabo actividades descritas como "automejora".*
6. *Abstracciones: averiguar qué métodos podría tener una máquina para construir abstracciones a partir de diversos datos.*
7. *Creatividad y arbitrariedad: la creatividad va implícita en la arbitrariedad, habría que controlar dicha arbitrariedad e intuición para crear pensamiento creativo.* (McCarthy, Minsky, Rochester, & Shannon, 2006)

Para este grupo de académicos e investigadores, la inteligencia artificial no era una rama más de la informática, de la psicología, o de ninguna de sus especialidades, era un nueva área de estudio. Su propósito era el de volcar la inteligencia humana en una máquina, para que pudiera replicar la toma de decisiones, el lenguaje, la capacidad de aprendizaje, e, incluso, de razonar. Este nuevo ámbito iba más allá de cualquier avance previo tecnológico de los hombres, querían crear lo, hasta el momento, imposible.

Después de la conferencia de Dartmouth se extendió un optimismo generalizado respecto a los avances que se podían lograr en el campo de la inteligencia artificial. Historias de éxitos como el desarrollo del programa ELIZA[2], entre 1964 y 1966, por Joseph Weizenbaum, o, el programa Solucionador General de Problemas[3] creado por

[2] Este programa era una herramienta que procesaba el lenguaje para simular una conversación humana. Fue uno de los primeros programas en conseguir superar el test de Turing.
[3] Era capaz de resolver automáticamente ciertos problemas.

Herbert Simon, llevaron a que los fondos destinados a este tipo de investigaciones crecieran drásticamente. Este optimismo fue tal, que incluso Marvin Minsky llegó a afirmar que entre tres y cinco años serían suficientes para desarrollar una máquina que incorporase la inteligencia general de un ser humano medio (Haenlein & Kaplan, 2019).

Herbert Simon dijo en 1957: *"no es mi intención sorprender a nadie, pero la manera más simple de resumirlo es decir que hoy en día existen máquinas que pueden pensar, aprender y crear. Su habilidad para realizar estos actos va a incrementarse rápidamente hasta que, en un futuro "visible", el rango de problemas que podrán manejar será igual al que se ha aplicado la mente humana"* (Russel & Norvig, 2016). Esta frase es una muestra más del optimismo mencionado, donde el futuro visible se refería a un horizonte temporal de unos diez años. Sin embargo, no fue hasta cuarenta años más tarde cuando comenzaron a hacerse realidad estas predicciones. La sobre confianza de los académicos estaba justificada por los avances que conseguían los nuevos sistemas de inteligencia artificial, pero, no tuvieron en cuenta que esos éxitos estaban basados en muestras muy pequeñas y problemas simples. La mayoría de aquellos sistemas fallaban cuando intentaban ampliar la muestra o complejizar los problemas (Russel & Norvig, 2016).

Se encontraron tres principales dificultades a la hora de aplicar los sistemas "inteligentes" a problemas más complejos y amplios. La primera fue la falta de conocimientos que tenían dichos sistemas sobre el asunto que iban a resolver, sólo funcionaban mediante simples manipulaciones sintácticas del hombre. Un ejemplo muy conocido fue el del programa para traducir el ruso al inglés, para conocer el contenido de documentación rusa referida al lanzamiento de Sputnik en 1957. Este programa se pensó de manera que las palabras fueran reemplazadas una a una por su significado en inglés, sin embargo, falló estrepitosamente. El intento dio lugar a la traducción de la frase en ruso *"the spirit is willing but the flesh is weak"*, a *"the vodka is good but the meat is rotten"* en inglés. El problema era que el programa de traducción necesitaba conocimientos sobre la ambigüedad de las palabras y las expresiones para poder establecer el contexto y significado de las frases. Este fracaso dio lugar a la desconfianza del gobierno estadounidense en los avances sobre este campo, y, en un informe en 1966, llegó a decirse que *"no existía una máquina para traducir textos científicos y no había ninguna esperanza de que ocurriera pronto"* (Russel & Norvig, 2016). La segunda

dificultad se refería a la amplitud de la muestra o del problema a resolver. Muchos de los sistemas resolvían los problemas mediante combinaciones de "pasos"[4] hasta que encontraban la solución, los académicos pensaban que esto era infinito y que lo único que se necesitaba para resolver problemas más amplios era una mayor potencia de hardware. Pero, estaban equivocados y fallaban cuando intentaban resolver problemas con más de una docena de elementos. Por último, la tercera dificultad era las limitaciones que existían sobre las estructuras básicas que utilizaban para crear inteligencia artificial. Por ejemplo, Minsky y Papert, dos académicos, probaron que los perceptrones[5] podían aprender cualquier cosa que ellos representaran, pero carecían de las herramientas para representar de manera exacta muchos elementos (Russel & Norvig, 2016) (Minsky & Papert, 1969).

Estas dificultades y fracasos erradicaron aquel optimismo generalizado sobre la inteligencia artificial, y, durante la década de los setenta la financiación y atención a este campo cayeron en picado. El gobierno británico encargó al matemático James Lighthill un informe sobre progresos conseguidos en el campo de la inteligencia artificial hasta el momento. En aquel informe Lighthill criticó duramente la visión optimista que reinaba sobre los posibles avances de la inteligencia artificial y estableció que las máquinas "inteligentes" podrían, como mucho, emular el comportamiento de un principiante en juegos como el ajedrez. Hablaba del concepto de "explosión combinatoria" refiriéndose a la imposibilidad que tenían los sistemas inteligentes de gestionar números demasiado grandes de posibilidades. Además, expuso que el razonamiento y el sentido común en máquinas era algo que se encontraba fuera de sus posibles habilidades. El gobierno británico retiró por completo los fondos a este campo salvo en tres universidades, la de Edimburgo, Sussex y Essex. Al mismo tiempo, ante fracasos como el sistema de traducción de 1957, el congreso de Estados Unidos se quejaba de la alta adjudicación de fondos para el estudio de la inteligencia artificial. Ambos hechos dieron lugar a que la financiación de este campo se estancara y los gobiernos frenaran en seco su apoyo. La realidad era menor que las expectativas generadas (Simon, 1965) (Crevier, 1993) (Haenlein & Kaplan, 2019).

[4] Este tipo de método para resolver problemas es conocido como método débil, puesto que no es capaz de escalar a problemas más complejos.
[5] Neuronas artificiales creadas por Frank Rosenblatt que podían formar redes neuronales artificiales.

Durante la década de los 80 se dio el éxito de ciertos sistemas expertos que podían emular las capacidades analíticas de los seres humanos y tomar decisiones. La definición de sistema experto es descrita como colecciones de normas que asumen que la inteligencia humana puede ser formalizada y reconstruida desde una perspectiva vertical como una serie de "y síes" (Haenlein & Kaplan, 2019). Uno de los ejemplos fue el programa R1, creado por John McDermott *"RI es un programa. que configura los sistemas informáticos VAX-11/780. Dado el pedido de un cliente, determina qué modificaciones, si es que hay alguna, se deben hacer al pedido por razones de funcionalidad del sistema y produce una serie de diagramas que muestran cómo se asociarán los diversos componentes del pedido... R1 se implementa como un sistema de producción. Tiene suficiente conocimiento del dominio de configuración y de las peculiaridades de las diversas restricciones de configuración que, en cada paso del proceso de configuración, simplemente reconoce qué hacer. En consecuencia, se requiere poca búsqueda para poder configurar un sistema informático"* (McDermott, 1980). El programa R1 utilizado por DEC (Digital Equipment Corporation) supuso que, para 1986, la empresa había ahorrado unos 40 millones de dólares al año. Prácticamente todas las grandes compañías estadounidenses empezaron a tener sus propios equipos de desarrollo de inteligencia artificial, y, poco a poco, esta área volvió a recibir atención (Russel & Norvig, 2016).

En 1981 los japoneses se sumaron a la carrera de la inteligencia artificial anunciando su proyecto "Quinta generación", Estados Unidos formó el MCC (Corporación de Microelectrónica y Tecnología de Computadoras), y, Gran Bretaña reinstauró la financiación a la inteligencia artificial después del informe Alvey. *"La industrial de la inteligencia artificial pasó de representar unos pocos millones de dólares en 1980 a billones en 1988, incluyendo cientos de compañías especializadas"* (Russel & Norvig, 2016). Aunque muchas de estas nuevas compañías y proyectos no consiguieron sus ambiciosos objetivos y cayeron en el olvido, supuso un gran avance.

Estos sistemas expertos durante la época de los noventa consiguieron grandes éxitos en áreas que permitieran la formalización que requerían, como el programa Deep Blue de IBM que venció al campeón mundial de ajedrez Gary Kasparov[6] que desmintió

[6] Ronda de partidas desde 1996 a 1997.

la afirmación de Lighthill sobre la "explosión combinatoria". Deep Blue procesaba 200 millones de posibles movimientos por segundo y determinaba el movimiento óptimo analizando 20 movimientos por delante con el método de árbol (Campbell, Hoane, & Hsu, 2002). El problema era que fallaban en áreas donde la formalización no cabía, como reconocimientos faciales o distinción de elementos dentro de imágenes (Hutson, 2018). Para realizar esas tareas era necesaria la interpretación de datos externos y aprender de ellos, para poder usarlos a la hora de conseguir los objetivos marcados. Necesitaban poder adaptarse. La inteligencia artificial se caracteriza por estos rasgos, dado que los sistemas expertos carecían de ellos no podrían considerarse puramente inteligencia artificial. La verdadera inteligencia artificial estaba más cerca de las redes artificiales neuronales de Minsky y Papert, aunque su trabajo quedara estancado por la falta de estructura básica, como se ha mencionado en párrafos anteriores (Haenlein & Kaplan, 2019).

Sin embargo, también durante los 90 la inteligencia artificial tuvo una gran expansión y se descubrieron múltiples aplicaciones y nuevas sinergias con muchas industrias. Además, fue en ese momento cuando la IA se integró en el método científico para desarrollarse. Para ser aceptadas sus hipótesis debían ser testeadas y estaban sujetas a rigurosos experimentos empíricos, sus resultados debían ser comprobados estadísticamente. Esto dio un giro a las aplicaciones de la IA, puesto que ya no valía con que la nueva gran idea funcionase en algo concreto y simple, sino que debía demostrar su utilidad y aplicabilidad de manera más amplia (Cohen, 1995). Esta tendencia hacia los métodos empíricos pudo verse en muchas de las ramas de la inteligencia artificial como los sistemas de reconocimiento del habla, de traducción o las redes neuronales que después dieron lugar a la tecnología de minería de datos. La inteligencia artificial dejó de estar aislada del resto de ciencias tecnológicas como la teoría de la información o la estadística, y, aunque, las áreas de visión o la robótica quedaron excluidas, por el nuevo grado de formalización y especialización del método empírico, poco a poco han ido reintegrándose (Russel & Norvig, 2016). El método empírico permitió a la IA desarrollarse en la medida en la que los problemas se convirtieron en más comprensibles, y, con ayuda del análisis matemático los métodos se hicieron más robustos y los objetivos más realistas. Además, según ha ido pasando el tiempo, los avances en el campo de la IA han ido basándose cada vez más en teorías ya existentes y en evidencias probadas para mostrar las aplicaciones reales en el mundo, en vez de en la proposición de nuevas ideas y simplemente prototipos con poca utilidad (Russel & Norvig, 2016).

A partir del año 2000 las tendencias siguieron cambiando el en área de la IA, muchos de los académicos más importantes en este campo, como Minsky o McCarthy, sugirieron que se estaba "desvirtuando" la inteligencia artificial, y, que esta debía dejar de centrarse en mejorar aplicaciones para determinadas tareas que ya funcionaban bien, para volver a sus raíces. El objetivo primero de la inteligencia artificial era el de crear máquinas que piensen, aprenden y crean. Los esfuerzos debían redirigirse hacia desarrollar una inteligencia artificial al nivel del humano *("human-level AI")*, también conocido como AGI (Inteligencia Artificial General). Esto planteará distintos desafíos y retos éticos, como si es inteligencia *"friendly"*, que se analizaran en este escrito más adelante (Russel & Norvig, 2016) (Yudkowsky, 2008). Otra de las tendencias que cambió fue el foco de encontrar el "algoritmo perfecto" a centrarse más en los datos, su disponibilidad, la manera de tratarlos y su cantidad. Empezaron a existir grandes bases de datos que estaban disponibles, y, se comprobó que era más eficiente un algoritmo mediocre con 100 millones de datos que un algoritmo perfecto que sólo contara con un millón de datos (Banko & Brill, 2001) (Russel & Norvig, 2016).

Además, las redes neuronales artificiales, consideradas como la "verdadera" inteligencia artificial volvieron con la forma del aprendizaje automático (más conocido como *"machine learning"*). El mayor ejemplo fue la creación del programa de Google AlphaGo, este programa conseguía batir al campeón del mundo del juego Go, más complejo que el ajedrez por los 361 posibles movimientos al empezar, en comparación a los 20 del ajedrez. Se creía que era imposible, pero basándose en el aprendizaje automático se hizo real. Actualmente, muchas de las aplicaciones consideradas como inteligencia artificial se basan en el aprendizaje automático y en las redes neuronales artificiales (Haenlein & Kaplan, 2019) (Silver, y otros, 2016).

Finalmente, la inteligencia artificial ha llegado a convertirse en una herramienta fundamental que forma parte del día a día de las sociedades. La IA no sólo impacta en las compañías, sino también en las personas y sus rutinas. Transforma la manera de tomar decisiones e incluso la manera de interactuar con los demás, sea desde el punto de vista más personal, negocio – cliente o empresa – empleado (Haenlein & Kaplan, 2019). Está ampliamente extendida en todas las industrias y sus aplicaciones pueden encontrarse en numerosos ámbitos, lo que ha acabado planteando ciertos dilemas. Estos dilemas, que serán expuestos más adelante, se refieren sobre todo a la pregunta de: ¿hasta qué punto

deben sustituir las máquinas inteligentes a los humanos en sus tareas?, o, ¿qué decisiones pueden ser tomadas por un programa por su cuenta?¿bajo qué criterio?

Para dar respuesta a estas preguntas y exponer los dilemas que causa la inteligencia artificial hoy en día a todos los niveles de la sociedad, en los siguientes apartados se tratará primero de ofrecer una definición consistente, así como de explicar los problemas de su definición, segundo, de exponer los distintos tipos de IA, y, tercero, de analizar el impacto real, positivo y negativo, que tiene la inteligencia artificial en la vida diaria de las personas.

5.2. Definiciones y dificultades del término

En este punto se ofrecerá la definición teórica de la inteligencia artificia, para después exponer las principales dificultades a las que se enfrenta el concepto, a la hora de aplicarlo y establecer sus límites.

No existe una definición concreta y consensuada sobre la inteligencia artificial, es un concepto líquido y en constante revisión. Tiene sentido que esto ocurra dada la naturaleza cambiante del mismo concepto, algo que está sujeto a una permanente evolución y de lo que no se puede establecer un límite "finito" de descubrimientos y avances, no puede ser catalogado en un marco estricto e inamovible. Sin embargo, los principales problemas de definición no hacen referencia a la definición teórica del término, sino a sus límites, aplicaciones prácticas y objetivos futuros.

El mayor problema con la definición práctica de la inteligencia artificial radica en dos razones principales. Primero, que cada definición establece unos objetivos distintos, no existen unos objetivos consensuados y concretos para la IA. Segundo, que según la perspectiva desde que se analice el término "inteligencia" este implica, y, significa, cosas totalmente distintas, por lo que varían las tecnologías incluidas dentro del área dela inteligencia artificial (Marín García, 2019). Antes de analizar en mayor profundidad estas dificultades, se presentarán las principales definiciones, pilares y características de la inteligencia artificial.

A lo largo de la historia de la inteligencia artificial sus estudiosos han ido dando ciertas definiciones teóricas. En 1973, durante el debate "Lighthill[7]" que descalificaba la IA, John McCarthy ofreció una definición a aquel término que había acuñado en 1955: *"La inteligencia artificial es una ciencia, es el estudio de la resolución de problemas y tiene como objetivo conseguir soluciones en situaciones complejas"*. Añade además: *"Es una ciencia básica, como las matemáticas o la física, y tiene problemas distintos de las aplicaciones, y, distintos de cómo funciona el cerebro del ser humano y de los animales. Requiere de experimentos para avanzar e incluye un gran número de partes"* (Haenlein & Kaplan, 2019). Los cuatro principales pilares o áreas que destacó McCarthy de la IA fueron: (1) los procesos de búsqueda que lidiaban con el problema de la "explosión combinatoria", (2) la representación de información de manera interna en la máquina, referido a la información sobre situaciones concretas que la máquina tendría que gestionar, y, la representación de procedimientos y leyes naturales, (3) aconsejar, cómo instruir al ordenador e influenciarlo, y, (4) la programación automática, entendido como una intuición más allá de la usada normalmente por un ordenador (Haenlein & Kaplan, 2019).

Su definición evolucionó con el tiempo y en 2007 estableció que la IA era: *"La ciencia e ingeniería de hacer máquinas inteligentes, especialmente programas de ordenador inteligentes. Está relacionada con la tarea de utilizar ordenadores para comprender la inteligencia humana, pero, la inteligencia artificial no se confina a métodos biológicamente observables"* (McCarthy, 2007). Otra de las aportaciones de John McCarthy al término fueron los componentes o ramas de la IA, él propuso doce ramas: (1) la IA lógica, referida al uso del conocimiento general y situaciones específicas para conseguir sus objetivos, (2) programas de búsqueda que lidian con muchas posibilidades diferentes, (3) patrones de reconocimiento que observan tendencias y las comparan con patrones conocidos, (4) representación de hechos utilizando la lógica matemática, (5) inferencia a partir de otros hechos, (6) sentido común y razonamiento, aunque es lo que está menos desarrollado en comparación al ser humano, (7) aprendizaje de experiencias del propio programa, (8) sistemas de planificación de objetivos, (9) epistemología como el estudio del tipo de conocimiento necesario para resolver

[7] El debate "Lighthill" hace referencia al informe Lighthill pedido por el gobierno británico en 1973. En este informe, como se comentaba en apartados anteriores, el matemático Lighthill daba una visión pesimista sobre la IA y cuestionaba sus posibles avances.

problemas del mundo real, (10) ontología como el estudio de varios tipos de objetos y sus propiedades, (11) heurística como una manera de descubrir una idea dentro de un programa, y, (12) la programación genética como técnica para conseguir que los programas resuelvan problemas a lo largo de generaciones, inspirada en la evolución biológica y en los algoritmos evolutivos (Skilton & Hovsepian, 2018).

Se puede asumir, que lo generalmente aceptado sobre lo que implica la inteligencia artificial es la creación de máquinas con la inteligencia del ser humano. La Real Academia Española (RAE) define la inteligencia artificial como: *"disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico"* (Real Academia Española , 2019). Kaplan y Haenlein la definen como: *"la habilidad de un sistema para interpretar datos externos de manera correcta, aprender de esos datos, y, utilizar dichos aprendizajes para conseguir objetivos concretos a través de una adaptación flexible"* (Kaplan & Haenlein, 2019). Otra definición que se acerca mucho es: *"la IA es la combinación de algoritmos planteados para crear máquinas con las mismas capacidades que el ser humano"* (Iberdrola, 2019). Por último, el actual director científico de Microsoft, Eric Horvitz define la inteligencia artificial como: *"el estudio científico de los principios de la computación detrás del pensamiento y el comportamiento inteligente"*. Establece también los cuatro pilares principales de la IA, similares a los de McCarthy, como: (1) la percepción, (2) el aprendizaje, (3) el procesamiento del lenguaje natural, y, (4) el razonamiento.

Queda claro que hay muchos puntos de convergencia en todas las definiciones de la inteligencia artificial. Básicamente, todos los académicos están de acuerdo en que la IA se basa en sistemas tecnológicos que puedan replicar el comportamiento de la mente humana, interpretar de manera "lógica" datos externos, y, aprender para conseguir objetivos. También, se acepta de manera general que la IA pretende sacar los procesos más elementales de la inteligencia y formalizarlos en forma de algoritmos y sistemas (Marín García, 2019). El problema llega a la hora de establecer qué entra en esta categoría, ¿dónde empieza y acaba la inteligencia dentro de una tecnología? La definición teórica del término inteligencia artificial no es tan confusa, lo es su puesta en práctica y contexto.

Surgen varias dudas con respecto a las afirmaciones dadas por las definiciones de la IA, ¿qué algoritmos se usan y en qué se basan?, ¿a qué capacidades del ser humano se refieren?, ¿qué es la inteligencia y cómo se implanta en una máquina?, ¿qué máquinas son consideradas como inteligentes?, ¿cómo diferenciar un mero avance tecnológico de un sistema inteligente? El grado de poca especificación existente, desde los orígenes de la inteligencia artificial, es el que provoca los problemas de definición y aplicación. Todos están de acuerdo en que la IA son máquinas inteligentes, pero no concretan aspectos fundamentales. A su vez, tiene sentido que haya aspectos que no se concreten, teniendo en cuenta los rápidos avances que se producen en este campo cambiándolo todo. Por ejemplo, no se puede establecer una lista exacta que incluya cuales son las tecnologías consideradas como inteligentes, porque según el punto de vista una misma tecnología puede ser considerada como IA o no. Otro ejemplo sería la imposibilidad de especificación de las capacidades que puede tener una máquina inteligente, este campo está en constante evolución y es imposible determinar en una lista cerrada las capacidades existentes y posibles.

Uno de los principales obstáculos a la hora de pasar de la teoría a la práctica las hipótesis y definiciones de la inteligencia artificial es el "efecto IA". Este efecto se refiere a la negación de la existencia del pensamiento autónomo de una máquina cada vez que se descubre cómo hacer que esta haga algo nuevo. Esto quiere decir que en muchos casos, el descubrimiento de una nueva modalidad o funcionalidad autónoma no es considerado como "pensar". Niega la existencia de inteligencia en las máquinas, es considerado como avances que realiza un ser humano sobre la misma. Así se relega a la inteligencia artificial a todo aquello que queda aún por descubrir y conseguir, no real y no presente (Hofstadter, 1979) (McCorduck, 2004) (Marín García, 2019). El gran problema radica en que esta hipótesis pierde de vista la definición comúnmente aceptada, rechazando que la IA exista en presente y relegándola al futuro, a la vez que deja de lado todos los avances conseguidos en este campo. Este "efecto IA" está relacionado con el problema mencionado previamente, sobre qué está considerado como inteligencia artificial y qué es, simplemente, un avance técnico conseguido por el ser humano.

Otra de las grandes dificultades, es la amplitud del término inteligencia. John McCarthy definió la inteligencia como: *"La parte computacional de la habilidad de conseguir objetivos en el mundo. Varios tipos y grados de inteligencia ocurren en las*

personas, en muchos animales, y, en las máquinas" " (McCarthy, 2007). Así mismo, la RAE define inteligencia como: *"capacidad de entender o comprender", "capacidad de resolver problemas", "conocimiento, comprensión, acto de entender"*, o*, "habilidad, destreza y experiencia"* (Real Academia Española, 2019). El problema es que la inteligencia puede definirse desde muchos campos de estudio distintos, como la filosofía, la psicología, la sociología… todas estas áreas plantean definiciones y conceptos distintos sobre la inteligencia, y, todavía no se puede hablar de un concepto de inteligencia separado y sin hacer referencia a la inteligencia humana (McCarthy, 2007) (Marín García, 2019). Esta amplitud del término provoca que la tesis de la Conferencia de Dartmouth que establecía que cualquier aspecto de la inteligencia podía ser descrito de manera tan precisa que podría crearse una máquina que lo emulase, o, que la afirmación de Leibniz sobre que todo lo que se sepa describir de forma clara, completa e inequívoca es computable, queden obsoletas (Palma & Marín, 2008). *"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"* (Marín García, 2019).

Aún así, sí que existe un consenso sobre los tipos de inteligencia existentes, los cuales intenta replicar la inteligencia artificial. El primer tipo y más básico es la inteligencia mecánica, se refiere a la habilidad de realizar tareas automáticas y llevar ciertas rutinas a cabo. Aplicada en máquinas, los procesos no requieren mucha creatividad y son muy repetitivos, se limitan a aprender y adaptarse al mínimo. El segundo tipo es la inteligencia analítica, entendida como la habilidad para procesar información, aprender de ella, y, en consecuencia, resolver problemas. Las máquinas con este tipo de inteligencia son capaces de aprender y adaptarse sistemáticamente según los datos que tengan disponibles. El tercer tipo es la inteligencia intuitiva, esta es la habilidad de pensar con creatividad y ajustarse constantemente a nuevas situaciones de las que no se tiene referencia. Las máquinas inteligentes con intuición pueden aprender y adaptarse en base a lo que entienden de la situación. Por último, el cuarto tipo es la inteligencia empática, la habilidad de reconocer y entender emociones de otros sujetos y reaccionar adecuadamente a las mismas. Es la generación más avanzada de IA y puede aprender y adaptarse de manera empática basándose en la su propia experiencia (Huang & Rust, 2018).

Pese al consenso sobre los tipos de inteligencia y los tipos de IA correspondientes a cada uno, el desconocimiento real de qué es la inteligencia y cómo se puede describir de manera inequívoca, hacen que muchos académicos discutan que sea incluso posible. Esto lleva a una nueva visión de la IA donde esta no emula la inteligencia humana, sino que imita ciertos elementos y puede servir como complemento al razonamiento de los seres humanos. Este complemento supliría a la persona en las tareas en las que una máquina puede ser más eficiente, como un cálculo muy complejo. Las interacciones que surgirían de esa complementación serían el antecedente de una colaboración de la IA con muchos agentes distintos, que podrían tener *outputs* muy positivos (Malone, 2018).

Habiendo revisado las principales definiciones y dificultades a las que se enfrente la inteligencia artificial como concepto, se puede llegar a la conclusión de que existe un consenso sobre la definición teórica pero hay una gran falta de concreción y diversas dificultades a la hora de aplicar y ponerle límites al término. La inteligencia artificial trata de replicar el comportamiento de la mente humana, para poder interpretar, razonar y resolver problemas. Pero, más allá de esos parámetros generales faltan muchas cuestiones por especificar, como qué capacidades exactas quiere formalizar en máquinas o qué sistemas cumplen de manera rigurosa los requisitos para ser considerados inteligentes. Además, hasta que no se concrete una definición de inteligencia, más allá de sus numerosas acepciones, el concepto es tan sumamente amplio que es imposible afirmar que este se pueda describir de forma tan clara como para traspasarlo a una tecnología.

En el siguiente apartado, que cerrará el punto sobre el concepto de la inteligencia artificial, se tratarán los tipos y clasificaciones de la IA.

5.3. Tipos y clasificaciones de la inteligencia artificial

La inteligencia artificial aunque comparte una definición general como área de estudio y nueva tecnología, puede ser clasificada según distintos puntos de vista y fijando distintos parámetros de comparación. En este apartado se pretende mostrar las principales clasificaciones y categorías de la IA, para cerrar el análisis sobre el concepto y pasar al impacto que conlleva su uso.

La IA puede ser clasificada según distintos criterios. Uno de ellos es según el tipo de inteligencia que muestre la máquina, las categorías de inteligencia son: (1) cognitiva, (2) emocional, y, (3) social. Según estas categorías, los distintos tipos de inteligencia artificial serían analítica, inspirada en el humano y humanizada. Otro criterio sería el de categorizar la inteligencia artificial según su etapa de desarrollo, la clasificación sería: (1) inteligencia artificial muy concreta (con muy poca amplitud de aplicación), (2) inteligencia artificial general, y, (3) súper inteligencia artificial (Kaplan & Haenlein, 2019). Todos los tipos y clasificaciones tienen en común que están afectados por el ya comentado "efecto IA". Una vez la tecnología es entendida y se comprende el proceso y el funcionamiento, la "magia" de la inteligencia artificial, o, de la tecnología súper avanzada, desaparece (Haenlein & Kaplan, 2019).

En general la inteligencia artificial persigue unos "mismos objetivos", que son imitar la forma de pensar del ser humano y construir equipos que puedan razonar. Pero, según se definan estos objetivos, la inteligencia artificial puede ser de cuatro tipos (Russel & Norvig, 2016):

- Tecnología que piensa como los humanos: referida a máquinas "con mente", que literalmente pueden pensar como un humano (Haugeland, 1985). También hace referencia a máquinas que puedan realizar actividades relacionadas con el pensamiento humano, como la toma de decisiones (Bellman, 1978).
- Tecnología que actúa como los humanos: aquellas máquinas que llevan a cabo funciones que requieren de la inteligencia cuando las realizan personas (Kurzweil, 1992). Estudia cómo hacer que los ordenadores hagan cosas en las que, hoy por hoy, los humanos son mejores (Rich & Knight, 1991).
- Tecnología que piensa de manera racional: modelos tecnológicos que tienen facultades mentales, como la percepción, el razonamiento o la capacidad de reaccionar y actuar (Charniak & Mcdermott, 1985) (Winston, 1992).
- Tecnología que actúa de manera racional: dentro del campo de la inteligencia computacional, estudia el diseño de agentes inteligentes. Tiene que ver con el comportamiento inteligente dentro de un artefacto (Poole & Goebel, 1998) (Nilsson, 1998).

Los objetivos compartidos de la IA mencionados previamente pueden resumirse en tres: (1) desarrollar modelos conceptuales que puedan comprender los procesos cognitivos que ocurren en el cerebro de las personas. Está relacionado con la neurociencia computacional. (2) desarrollar procedimientos de reescritura formal de dichos modelos, es decir, intentar replicar y formalizar los procesos cognitivos observados, y, (3) desarrollar estrategias de programación para reproducir de la manera más completa e inequívoca dichos procesos cognitivos (Mira & Delgado, 1995). Estos objetivos provocan una nueva división de la IA en dos áreas o categorías distintas, la IA como ciencia y la IA como ingeniería del conocimiento (Palma & Marín, 2008). La IA como ciencia busca una teoría computable del conocimiento humano y está más relacionada con el primer objetivo de desarrollar modelos conceptuales. La IA, primero, como ingeniería intenta formalizar los procesos cognitivos, es decir, está relacionada con el segundo objetivo. Por último, la IA como ingeniería del conocimiento se refiere a la programación del lenguaje formal en sistemas (Marín García, 2019).

Por último la clasificación más reciente y la más utilizada es la dada por el profesor Arend Hintze. Hoy en día aún no convivimos con esos humanoides prometidos por la IA desde sus orígenes, ni existe una tecnología que supere cognitivamente a las personas. La inteligencia artificial tiene dos herramientas fundamentales que son las que, algún día, conseguirán el objetivo de crear esas máquinas soñadas. Esas herramientas son el aprendizaje automático y el aprendizaje profundo. El aprendizaje automático o *"machine learning"* otorga la capacidad de aprender a través de la identificación de patrones de comportamiento en conjuntos de datos, permite modificar el comportamiento actuando en consecuencia de sus análisis. El aprendizaje profundo o *"deep learning"* es una subcategoría del aprendizaje automático y se basa en el uso de redes neuronales que permiten que la máquina "piense" de manera más independiente (IAT, 2020) (Banafa, 2016). Aunque existen más herramientas propias de la IA, estas dos son las más relevantes. Ambas ayudan a establecer los cuatro tipos distintos de IA existentes, que suponen también ciertos límites a superar (Redacción APD, 2019):

- Máquinas reactivas: el tipo más básico de IA. No pueden basarse en la experiencia para tomar decisiones, pueden predecir la mejor opción entre las opciones que tienen. Un ejemplo es Deep Blue.

- Máquinas con memoria limitada: tienen algún tipo de memoria y pueden reaccionar según la experiencia. Pero es limitada, tanto en duración como en espacio, y en situaciones específicas. Un ejemplo son los coches autónomos.
- Máquinas con teoría de la mente: no solo forman representaciones sobre el mundo, sino sobre otros agentes. Comprende a las personas y objetos, permite la interacción social. Se acercan más a lo que se entiende como IA, al objetivo a lograr.
- Máquinas con autoconciencia: representan el paso final, sistemas que puedan realizar representaciones sobre sí mismos. Serán conscientes de su propia existencia y podrán entender y predecir los sentimientos ajenos. Este avance está aún muy lejos de lo conseguido hasta ahora.

Este apartado sobre los distintos tipos de inteligencia artificial y posibles clasificaciones cierra el análisis del concepto de la IA y permite el paso al impacto que esta tiene sobre las sociedades.

5.4. Impacto

La inteligencia artificial ya ha ocupado un lugar en las vidas cotidianas de las personas, tanto a nivel privado como a nivel empresarial. Un 47% de 2000 empresarios afirman que sus empresas contaban con algún sistema de IA, el 71% decían que la inversión en este área crecería de manera exponencial en los próximos años (McKinsey&Company, 2018). Además, de entre las tecnologías emergentes es la que mayor contribución supondrá para la economía mundial para 2030. Mientras que la nanotecnología se espera que contribuya unos 125 billones de dólares o el quantum unos 65 billones de dólares, la contribución de la IA se espera que sea de unos 15.7 trillones de dólares para 2030 (Tanguay, 2019). Esta estimación es muy superior a los 2000 millones que representaba en 2015 (Iberdrola, 2019). La IA ya ha llegado, aún queda mucho por desarrollar, pero, ha venido para quedarse.

El número de aplicaciones de la inteligencia artificial es inmenso, está presente en multitud de áreas distintas. Algunas de las más extendidas son (Russel & Norvig, 2016):

- Los vehículos autónomos: referidos a aquellos vehículos que no necesitan de un humano para ser dirigidos, cuentan con sensores, cámaras y radares especiales. Un ejemplo es el Volkswagen Touareg STANLEY.
- Reconocimiento de voz: estos sistemas reconocen el lenguaje humano y pueden interactuar con él. Cualquier contestador automático con el que se puede mantener una conversación para obtener un servicio es un ejemplo.
- Planificación autónoma: dispositivos capaces de planificar qué actividades son necesarias y en qué orden para conseguir unos objetivos previamente fijados. El primero fue el programa *Remote Agent* de la NASA que planificaba un programa para controlar las operaciones de una astronave. Pueden detectar problemas, diagnosticarlos y solucionarlos.
- Participación en juegos: humano contra máquina, el primero fue el programa *Deep Blue* contra el campeón mundial de ajedrez Garry Kasparov. Gracias a este sistema, IBM llegó a valer 18 billones de dólares.
- Lucha contra el *spam*: algoritmos capaces de distinguir el *spam* y desecharlo, además, aprenden y evolucionan puesto que las técnicas de los *spammers* evolucionan constantemente.
- Planificación de logística: sistemas que consiguen organizar vehículos, personas, transporte de objetos o cualquier tipo de actividad en un plan concreto. Los Estados Unidos utilizaron esta técnica durante la Guerra del Golfo de 1991 para programar los transportes y personas. Sin el programa DART hubieran tardado semanas en realizar el plan.
- Robótica: todos los sistemas automáticos de limpieza, como aspiradoras autónomas, entran en esta categoría. Otro ejemplo son los *bots* capaces de transportar objetos pesados y peligrosos.
- Traducción: algoritmos y sistemas que permiten la traducción instantánea de un idioma a otro.
- Visión artificial: tecnología que puede procesar y comprender imágenes y traducirlas a lenguaje computacional.
- Aprendizaje automático: máquinas que pueden aprender a través de la identificación de patrones en los datos.

- Aprendizaje profundo: máquinas que usan redes neuronales para “pensar” de manera independiente, sin la necesidad de los parámetros establecidos por el ser humano.

Estas aplicaciones está claro que traen muchos beneficios a las vidas de los seres humanos, además de un gran crecimiento económico. La IA cambia la manera de relacionarse de las personas, de hacer las cosas, e, incluso, de gestionar las empresas y la relación con el cliente. Se predice que para finales del 2020 el 85% de la relación con los clientes será llevada por sistemas IA, además, los Estados invertirán en esta tecnología por el crecimiento económico que puede suponer, con China y Estados Unidos a la cabeza de la carrera (Iberdrola, 2019). La carrera por conseguir los mayores avances dentro de la IA ha comenzado, los grandes beneficios que supone esta tecnología a la vida diaria la hacen merecedora de atención e inversión.

Los beneficios quedan reflejados, más allá de las aplicaciones comentadas, en ejemplos concretos. Por ejemplo, los sistemas de visión artificial ya pueden identificar anomalías en radiografías y adelantar el diagnóstico de enfermedades. Se ha demostrado que esta tecnología es entre un 67% y 97% más eficiente que un panel de médicos a la hora de diagnosticar ciertas condiciones pulmonares (Koo, y otros, 2012). Otro gran avance muy beneficioso son los asistentes virtuales. Estos se usan en áreas como marketing donde atienden las llamadas de los clientes y gestionan sus peticiones, o, también en análisis de datos de clientes para aconsejar el servicio o producto que mejor se adecua a su perfil. Así se mejora la relación negocio – cliente, además de la vida privada puesto que existen asistentes virtuales disponibles para el propio uso del consumidor como Alexa o Siri (Wilson & Daugherty, 2018). También, las ventajas del uso de la IA se ven reflejadas en el área de logística y transporte. Existen almacenes inteligentes sin necesidad de recursos humanos o granjas que utilizan sistemas inteligentes para gestionar los desplazamientos, la calidad de la materia prima o el consumo (European Parliamentary Research Service, Scientific Foresight Unit (STOA), 2016). Supone un gran ahorro en costes y evita los desperdicios de recursos, la empresa española El Dulze utiliza robots para seleccionar las mejores hojas de lechuga y su tasa de desperdicio cayó del 20% al 5%. Por último, uno de los avances más sonados es el de los coches autónomos, tienen sus riesgos, pero se prevé que los accidentes caerán en picado gracias a estos coches (Marín García, 2019).

Signo de todos estos beneficios es el impacto que ha tenido la IA en la inversión a nivel mundial. Cada vez más *start-ups* generan ideas innovadoras relacionadas con la inteligencia artificial, la inversión ha crecido exponencialmente en los últimos años alcanzando el 12% en 2018 del total del capital privado invertido, mientras que en 2011 sólo llegaba al 3%. Entre 2016 y 2017 la inversión privada se duplicó, llegando a los 16000 millones de dólares en 2017. Gracias a estas inversiones y a los resultados obtenidos la IA va camino de su implantación mundial a todos los niveles (OECD, 2019).

Pero, ¿qué implica realmente el uso de la IA en la vida de las personas? En primer lugar, permite el acceso al conocimiento y millones de datos lo que ha reducido la desigualdad a nivel mundial en términos de educación y percepciones reales del mundo y los mercados. Por ejemplo, miles de agricultores han podido conocer los precios justos y los productos que podrían aumentar su rentabilidad (Bryson, 2019). Puede suponer una oportunidad para mejorar las sociedades y aumentar la calidad de vida de las personas, esto es posible bajo la condición de que la información dada sobre las implicaciones de la IA sea clara y completa. Como se verá en próximos apartados la información sobre la IA y la regulación gubernamental es necesaria siempre para optimizar las ventajas que esta puede aportar. La educación a los individuos sobre los potenciales riesgos, los límites éticos y el buen uso de los sistemas inteligentes son fundamentales para minimizar daños. Si se utiliza con cabeza, responsabilidad y sentido común la vida será mucho más sencilla. Las jornadas laborables se reducirán, las tareas serán realizadas de manera más eficiente, se reducirá el tiempo perdido… *"La inteligencia artificial puede ser el catalizador de nuevos cambios, profundos y positivos, en el empleo y las relaciones laborales, en la desubicación de los centros de trabajo, y, en general, en una mejora de las condiciones de vida. Pero, para garantizar los efectos positivos, los ciudadanos deben tener sentido crítico y exijan a los poderes públicos una gestión adecuada de los beneficios"* (CSIC, 2019).

La IA consigue reunificar muchas áreas de estudio que se han considerado históricamente como independientes y no correlacionadas, esto permite avances a todos los niveles como ya se ha visto en este apartado. Desde desarrollo tecnológico para la medicina, logística, hasta avances en tratamientos psicológicos. No obstante, no todo son puntos positivos, el uso y desarrollo de la IA conlleva muchos riesgos y retos a solucionar como puede ser el balance del empleo al desaparecer ciertos tipos de oficios y crearse

otros nuevos. En próximos apartados se analizarán más en profundidad estos potenciales peligros, tanto a nivel ético como económico y personal.

Pese a los desajustes en las sociedades por los nuevos cambios, la perdida de privacidad, el potencial aislacionismo y distintos riesgos, los seres humanos están lejos de convertirse en obsoletos. Nick Bostrom, filósofo sueco de la Universidad de Oxford, afirma que entre 2075 y 2090 hay un 90% de posibilidades de que existan sistemas tan inteligentes como los humanos. Stephen Hawking decía que las máquinas superarán a los hombres en menos de 100 años. Estas afirmaciones dan vértigo y miedo, pero, no hay que perder la visión de que la inteligencia artificial complementará a los seres humanos, no los sustituirá. Acciones que antes quedaban fuera del alcance de las personas serán posibles y la eficiencia irá ganando terreno. *"¿Te imaginas explorar partes del universo totalmente hostiles para el ser humano? Gracias a la IA, un día será posible"* (Iberdrola, 2019).

Tras este apartado sobre el impacto de la IA en las sociedades, se cierra el punto sobre el contexto de la inteligencia artificial para dar paso a su relación con la ética, los potenciales riesgos que conlleva y hacia dónde se dirige este campo lleno de posibilidades.

6. Inteligencia artificial y ética

Este punto abre la discusión sobre el papel que debe tener la ética en la inteligencia artificial, a nivel de nuevos avances, límites, regulación y uso. Pretende responder a preguntas como hasta dónde debe llegar la IA, quién es responsable de los sistemas inteligentes, y, a qué retos éticos hace frente esta área de estudio.

6.1. ¿Por qué la ética?

En apartados anteriores se comentaba como era el potencial mal uso que hacen los humanos de las aplicaciones de la inteligencia artificial lo que suponía un mayor peligro, la ética se debe aplicar precisamente por este motivo. Aunque la ética y la inteligencia artificial puedan parecer a priori dos realidades totalmente separadas, están muy conectadas. En este apartado se establece la relación entre estas dos áreas,

respondiendo a la pregunta de: ¿por qué la ética? Para luego exponer los principales retos éticos a los que se enfrenta este nuevo campo con tanto por delante por explorar.

La inteligencia artificial y sus aplicaciones no son fruto de avances tecnológicos fortuitos, dichos avances se producen en base a los objetivos fijados de antemano por personas concretas. Que sean las personas, con intereses y planes propios, las que toman las decisiones y sellan el destino de las aplicaciones que crean, pone en evidencia la no neutralidad de la inteligencia artificial y el potencial riesgo que puede representar. Así se hace clara la relación entre la ética y la inteligencia artificial, necesaria dado que los "creadores" están sujetos al razonamiento ético, y, por ende, sus creaciones (Marín García, 2019). *"La ética es voluntaria pero no opcional: es una exigencia de la excelencia"* (Argandoña, 2019). Por ello, como al profesional se le exige excelencia, y, sus creaciones tecnológicas, en este caso, son una extensión de su profesión, caen dentro de las mismas exigencias y necesidades éticas.

En múltiples ocasiones las perspectivas fatalistas y distópicas, mencionadas en apartados anteriores, tienen un punto de razón en lo que se refiere a los retos éticos que presenta la aplicación práctica de la inteligencia artificial en la vida cotidiana. Un ejemplo que presenta esta idea de manera muy clara es el uso de coches autónomos. Estos coches están diseñados de manera que incluyen un algoritmo extremadamente complejo que pueda hacer frente a situaciones imprevistas durante la conducción. Estas situaciones pueden ser un semáforo que no funciona, un peatón que cruza mal o una señal de tráfico caída. Esto implica que dicho algoritmo debe tener la capacidad de "decidir por si mismo" cómo actuar, pero, ¿en base a qué parámetros? Parece evidente que los parámetros deberían ser establecidos bajo patrones de conducta éticos, sin embargo, ¿qué es lo ético? ¿quién decide lo que es ético? Ante un accidente inevitable, ¿qué vida debería priorizar el vehículo?, la del pasajero o la del viandante. Decida lo que decida, existe un reto ético en el sentido de la autonomía del coche, puesto que no está claro cuál sería la opción "correcta" (Bonnefon, Shariff, & Rahwan, 2016). Este reto ético no es intrínseco al coche autónomo, sino que está directamente relacionado con el creador y programador del mismo. Desde el momento en el que aparecen estos retos o situaciones es necesario que la regulación ética y la capacidad de responsabilizar éticamente a los sujetos, estén presentes.

Aunque es cierto que la responsabilidad ética no es la misma en unos casos u otros, sí que existen retos a los que la inteligencia artificial se enfrenta de manera conjunta. El primero se refiere a delimitar qué es la inteligencia artificial, qué es "inteligencia" dentro de un dispositivo, según el grado de inteligencia y autonomía la responsabilidad ética será mayor o menor. Por otro lado, se debe esclarecer el proceso desde el diseño del algoritmo hasta su toma de decisiones de manera independiente. En muchos casos es imposible llegar a saber cómo es el proceso de toma de decisiones puesto que los dispositivos están programados para ir mejorando constantemente, este es el motivo por el que adscribir responsabilidades es sumamente complicado, por la falta de entendimiento (Marín García, 2019). Además, se debe recordar que detrás de toda decisión, o, proceso de toma de decisiones, de una máquina hay un ser humano. Como se mencionaba previamente, los seres humanos están sujetos a la responsabilidad ética, y, por lo tanto sus creaciones también. *"Los problemas morales los tienen las personas, no las máquinas ni el software"* (Argandoña, 2019). Por mucho que la falta de entendimiento del proceso pueda dificultar la adjudicación de responsabilidades, no pueden entenderse de manera independiente las decisiones "autónomas" de cierta tecnología y la intención de su creador. Teniendo en cuenta esta afirmación, la ética sirve para vigilar cuáles son los efectos de las decisiones tomadas por la inteligencia artificial sobre las personas (Argandoña, 2019).

La IA supone nuevos retos que hacen peligrar la ética, estos peligros están relacionados con las nuevas preocupaciones que provoca la IA, su desarrollo y su funcionamiento (OECD, 2019). Pueden resumirse en cuatro principales (Marín García, 2019):

1) La rendición de cuentas: este peligro hace referencia a la adjudicación de responsabilidades ante cualquier actuación dañina de un sistema inteligente. Estos sistemas cada vez están más presentes en la vida de las personas, ante su creciente autonomía es difícil saber si la responsabilidad es de la persona detrás del diseño y desarrollo o del sistema en sí (Comisión Europea, 2019). Esta duda suscita un debate, que se mencionará más adelante, sobre sí la autonomía de las máquinas implica derechos y responsabilidades para las mismas.
2) La explicabilidad: en línea con las ideas expuestas previamente, la falta de explicación y entendimiento de decisiones tomadas por una máquina de manera

independiente dificulta saber quién es responsable. Hay un alto riesgo de que sistemas con IA tomen decisiones impredecibles e inexplicables, un ejemplo serían las máquinas con aprendizaje profundo (Bostrom & Yudkowsky, 2014). Esto hace fundamental que se diseñen sistemas transparentes en cuanto al uso de la IA, para poder explicar los procesos y que funcionen de manera segura (OECD, 2019).

3) La imparcialidad: la IA gestiona y funciona con grandes cantidades de datos, pero las muestras de datos o el diseño mismo del sistema pueden llevar a algún tipo de sesgo que influya en las decisiones (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016). Estos sesgos quedan reflejados en sistemas de ventas, marketing, asignación de hipotecas o selección de candidatos para puestos de trabajo. Provocan discriminación sobre género y raza. Para que esto no ocurra son necesarias muestras de datos más representativas o sistemas de modificación manual, ambas opciones son muy complejos.
4) La privacidad: en relación al peligro anterior, las mismas grandes cantidades de datos disponibles para los sistemas con IA provocan la pérdida de privacidad progresiva. Muchos sistemas como Siri o Alexa están presentes en la vida cotidiana de las personas, recabando información personal constantemente, incluso apagados (Fussell, 2019). Esos datos acaban estando disponibles para las empresas detrás de los dispositivos, lo que puede llevar a la manipulación de la población, erosión de las instituciones o la creación de hábitos dañinos psicológicamente hablando (Eyal, 2017).

La IA está cambiando las sociedades, la forma de vida, la manera de relacionarse de las personas, de los tipos de trabajo y la concepción de privacidad y responsabilidad. La ética es fundamental por todo lo expuesto en este apartado, para garantizar que la transición de un mundo sin IA a un mundo con ella sea lo más natural y fácil posible (OECD, 2019).

Este apartado da paso a los principios éticos bajo los que debe desarrollarse y funcionar la IA y los métodos para implementar los mismos.

6.2. Principios éticos en la IA y métodos de implementación

En el apartado anterior se exponía la importancia de la ética en la inteligencia artificial, así como los principales peligros a los que hace frente. Pero, ¿cómo desarrollar una IA correcta y dentro de los límites? Para que esta pueda ser fiable y su uso no provoque conflictos éticos ni tenga consecuencias perjudiciales, es necesario que la IA se desarrolle bajo unos principios éticos que potencien sus beneficios y minimice sus riesgos. Estos principios no son un conjunto de normas específicas que se apliquen siempre de la misma manera, dependiendo del contexto y de la aplicación se ajustarán. Pueden resumirse en cinco imperativos principales (Marín García, 2019):

1) El respeto de la autonomía humana: entendido como el respeto en todo momento a la autonomía y los derechos básicos de las personas desde fases iniciales del desarrollo de la tecnología.
2) La transparencia: es necesario que cualquier decisión tomada por una máquina inteligente pueda ser trazada, es decir, que se entienda el razonamiento seguido y que se puedan identificar los datos utilizados y los pasos seguidos. Esto ataja el problema de la explicabilidad mencionado anteriormente, es incompatible que un sistema pueda tomar decisiones impredecibles con la defensa de la autonomía humana.
3) La responsabilidad y rendición de cuentas: se deben designar las responsabilidades en caso de perjuicios desde la fase de diseño. No es excusa la autonomía de la máquina para diluir las responsabilidades. Por autónoma que sea una máquina, esa autonomía viene dada por una programación humana que puede ser comprendida (Buchholz & Rosenthal, 2002).
4) La robustez y seguridad: los algoritmos que conforman la IA deben ser seguros y fiables, para poder resolver cualquier tipo de error o incoherencia. El diseño de los mismos debe contar con posibles ciberataques o fallos.
5) La justicia y no discriminación: se debe prever con qué grupos va a interactuar el sistema inteligente y que todos estén incluidos en el mismo. Para que haya un uso justo de los datos y evitar discriminaciones la participación total es necesaria.

Estos principios se focalizan principalmente en la fase de diseño dado que es donde queda configurada la máquina en sí, pero, también deben estar presentes en las siguientes fases de desarrollo para que sean efectivos. La pregunta es ¿cómo implementarlos? Se plantean aquí métodos técnicos y métodos no técnicos para ello (Marín García, 2019).

Métodos técnicos

Tienen como objetivo implementar los principios éticos de la IA en el diseño de los algoritmos que la conforman. Programando la arquitectura de los sistemas inteligentes en base a unos parámetros éticos, se podrá garantizar la seguridad y comportamiento adecuado de los mismos.

- *Ethics by design,* se trata de diseñar los algoritmos de manera que se garantice el futuro comportamiento ético. Los mecanismos son tres, que la máquina observe y aprenda de los humanos los comportamientos éticos, establecer normas que rijan que clase de conducta debería tener el dispositivo, y, que los sistemas adaptasen su comportamiento según la situación y contexto.
- IA explicable, concepto más comúnmente conocido como XAI (*Explainable AI*). Para que la IA sea transparente se proponen métodos para mostrar su funcionamiento, como la investigación de árboles de decisiones donde se estudia como dependiendo de la información y parámetros establecidos varían las decisiones.
- Prueba y validación del producto, establecer exámenes y pruebas de validación a los dispositivos para observar cualquier posible fallo y poder corregirlo a tiempo. Ayudan al posterior diseño de otros sistemas, que no cometan los errores observados.

Métodos no técnicos

Los métodos no técnicos se hacen necesarios porque el simple diseño de algoritmos bajo unos parámetros no es infalible, además de que presenta dificultades. Pueden presentarse situaciones en las que el simple diseño no sea suficiente para evitar conflictos. Además, en métodos como la observación del comportamiento humano cabe

la duda de si lo observado es realmente ético o simplemente lo común. Estos mecanismos contribuyen a garantizar el buen uso y la seguridad más allá del diseño de los dispositivos con inteligencia artificial.

- Regulación, los gobiernos y agencias internacionales deben legislar y regular el desarrollo y el uso de la IA. Se pueden establecer parámetros de seguridad de uso, normas, contratos…
- Certificaciones, las empresas productoras de dispositivos inteligentes podrían emitir certificaciones sobre su seguridad, fiabilidad y transparencia, fomentando la confianza de los usuarios.
- Educación y sensibilización, para crear conciencia de los potenciales riesgos de la IA y de sus beneficios. Debe alcanzar a cualquier grupo que interactúe a cualquier nivel con sistemas de IA.
- Investigación, los gobiernos deben fomentar la investigación de manera segura y fiable, para asegurar que los resultados son acordes a lo esperado y hagan frente a los desafíos.

Estos métodos, técnicos y no técnicos, sirven para desarrollar una IA acorde a la sociedad y a las personas, que tenga en cuenta los valores morales y los límites que no deben traspasarse. No es suficiente que la programación tenga en cuenta los principios éticos, sino que debe haber implicación por parte de las sociedades para garantizar el buen uso y minimizar los riesgos. La IA es un producto social que influye en la manera de comunicarse, desplazarse, relacionarse, de vivir, de los humanos (Verbeek, 2011).

Este apartado acaba de definir en qué consiste el uso prudente de la IA y qué mecanismos emplear para conseguirlo. Se cierra así el punto sobre la relación entre la ética y la IA, para pasar a un último punto sobre la situación en la actualidad.

7. La actualidad

Tras haber hecho un recorrido a lo largo de los conceptos más importantes relacionados con la ética y la inteligencia artificial, y, haber establecido qué relación existe entre estas dos áreas tan distintas, este punto pretende exponer cuál es la situación

actual. ¿Qué riesgos supone en la vida cotidiana el uso de la IA? ¿En qué dirección va la evolución de este campo? ¿Qué debe hacerse?

7.1. Riesgos de la inteligencia artificial

Este primer apartado responde a la pregunta de cuáles son los riesgos a nivel práctico de un mal uso de la IA. Como se ha mencionado previamente la IA en sí conlleva ciertos peligros, pero, cuáles son las consecuencias de dichos peligros a nivel personal y de sociedad. Está claro que las ventajas, como se veía en el apartado de impacto, son inmensas, sin embargo, los nuevos avances técnicos siempre plantean nuevos riesgos al cambiar la forma de vida general.

Uno de los principales riesgos a los que se enfrenta la sociedad al introducir el uso de la IA en la empresa, es la destrucción de muchos oficios y la creación de empleos nuevos. Esto produce un gran desajuste, al no estar preparada la clase trabajadora para afrontar estos nuevos empleos más técnicos. La destrucción de empleo masiva ha sido un efecto muy común de los cambios disruptivos en términos tecnológicos, dado que estos siempre vienen acompañados de grandes cambios sociales (Lin, Abney, & Bekey, 2011). Aún está por ver hasta que nivel penetrará la IA en el ámbito laboral, y, de ser así, hasta que punto significará la destrucción de empleos (Quinn, 2016). Sin embargo, ya se estima que entre el 21% y 38% del empleo en países desarrollados desaparecerá a causa de la automatización, la penetración de la IA y la digitalización (Hawksworth, Berriman, & Goel, 2017). Estos cambios, más allá de si son éticos o no, son inevitables y el verdadero reto será la capacidad de adaptación de las empresas, de las personas y de los gobiernos.

El concepto de desempleo tecnológico fue introducido por John Keynes en 1930, desde ese momento se defendía que la tecnología destruía empleos pero también creaba otros distintos. Esto no será distinto con la introducción de la IA al mundo laboral, donde los trabajadores realizan cientos de tareas diarias de las que sólo unas pocas puede realizar, a día de hoy, una máquina. Se predice que el 47% del empleo de Estados Unidos estará sujeto a la automatización, pero, el coste actual de implementar tecnologías con IA es demasiado alto como para sustituir por completo al ser humano (Kaplan & Haenlein, 2019).

Existen visiones más pesimistas que afirman que en un mundo donde las máquinas harán todo mejor que cualquier humano, ¿por qué contratar personas? Sin embargo, la realidad más cercana es que en los próximos 5 años la IA creará más de 50 millones de puestos de trabajo. Es necesario una nueva preparación de los empleados y formación, cambiarán los trabajos, no desaparecerán. Además, existirán nuevas oportunidades de realizar tareas más allá de lo puramente mecánico y rutinario. En el corto plazo no hay peligro de que la IA sustituya al ser humano en su totalidad, aunque se debe gestionar con cuidado el cambio dentro del ámbito laboral (Kaplan & Haenlein, 2019).

Otro de los grandes riesgos es el de la manipulación y los posibles sesgos presentes en la inteligencia artificial. La IA está compuesta de elementos como el software o hardware muy fáciles de manipular y con una alta probabilidad de error en su funcionamiento. Además de que sus propios componentes sean objeto de manipulación malintencionada, al funcionar basándose en amplias cantidades de datos sus decisiones pueden estar sesgadas en muchas ocasiones (Marín García, 2019).

Este era un problema ya mencionado anteriormente pero a un nivel menos teórico y más práctico los sesgos se pueden observar en muchas ocasiones. Un ejemplo muy claro es el de Google Photos, al confundir a una pareja negra con gorilas. El sistema de IA de la aplicación etiqueta imágenes con lo que son, por el sesgo existente y la falta de información que tenía, etiquetó de manera errónea a la pareja (BBC Mundo, 2015). También, ha habido un incremento exponencial de las *fake news*, manipulación de contenidos, modificación de rangos de precios justos, e, incluso distorsión de la información.

También el uso de sistemas con IA puede suponer perjuicios a las personas, como la perdida de habilidades personales. El sobreuso de las tecnologías a diario ya está provocando muchos cambios en las formas de relacionarse y de comportarse de las personas. Fomenta el aislacionismo y la sensación de soledad, lo que puede derivar a problemas psicológicos más complejos. Además, se están perdiendo progresivamente gran parte de las habilidades personales y sociales, pues al tener una máquina a través de la cual relacionarse, las personas ya no necesitan realizar esfuerzos. Las habilidades cognitivas de los seres humanos, su estabilidad emocional e incluso la salud física pueden verse altamente perjudicadas por un uso excesivo de herramientas tecnológicas (Goldhill,

2015). Los sistemas con IA llevan este problema a un paso más allá, ya no son sistemas que sirvan solo para comunicarse, sino que pueden tomar decisiones por los individuos. La reducción de interacciones sociales, de esfuerzos y de preocuparse por decidir están haciendo mella en las sociedades (Marín García, 2019).

La IA tendrá un gran impacto tanto en las desigualdades sociales como en la sensación de aislamiento y soledad. Relacionado con el riesgo del desempleo, un aumento de la productividad producido por la IA puede que no sea beneficioso del mismo modo para todo el mundo, lo que puede aumentar las desigualdades y no en un crecimiento económico general (Piketty, 2015). Respecto al aislamiento, al sustituir muchas tareas realizadas hoy en día por personas las interacciones sociales se reducirán. Ya existen robots semi-humanoides que pueden interactuar con las personas, entendiendo sus sentimientos y reaccionando adecuadamente, pero no pueden sustituir el contacto humano (Kaplan & Haenlein, 2019).

Por último, uno de los mayores peligros es el de la erosión civil y de las instituciones. Las herramientas con inteligencia artificial abren un nuevo abanico de posibilidades para la manipulación de la opinión pública, los intereses políticos y económicos o la distorsión de la realidad. Resta credibilidad a la sociedad civil, a los medios de comunicación y a los propios gobiernos o empresas que lo utilizan en su beneficio. Un ejemplo es el caso de *Cambridge Analytica*, la empresa británica que utilizaba datos privados de los individuos a través de Facebook para saber que contenido debían incluir las campañas políticas (BBC Mundo, 2018). Esto demuestra que el mundo de las noticias falsas, la manipulación de la opinión y la pérdida de privacidad individual, está más cerca que nunca.

Otro de los problemas políticos más importantes es el potencial uso militar y armamentístico de sistemas con IA. La robótica puede crear drones, robots militares e incluso exoesqueletos que aumenten la fuerza y protección de los soldados. Por otro lado, la IA puede utilizarse en la toma de decisiones y la planificación estratégica, pero, también puede ser *hackeada* por el oponente. Está claro que el problema no es el uso de robots en actividades peligrosas, sino el uso que pueda darse de los mismos (Kaplan & Haenlein, 2019).

En muchos casos se cree que la IA está a salvo de sesgos y de corrupción y que por eso las decisiones serían más lógicas si las tomaran estos sistemas, pero, se olvida que las máquinas tienen personas detrás que las diseñan y desarrollan, nunca están libres de sesgos ni de ideología, nunca son neutrales.

En conclusión, la IA puede ampliar y mejorar las capacidades del ser humano, pero presenta grandes retos de seguridad, derechos humanos y económicos que no deben dejarse a un lado. La limitación de la IA es necesaria, igual que su control y regulación para un buen uso. Tras haber expuesto los principales riesgos de la aplicación de la IA se pasa a un último apartado sobre cuáles son los retos actuales del desarrollo de la IA, en qué punto se encuentra y hacia dónde se dirige la sociedad global de la mano de estos nuevos agentes.

7.2. Hacia dónde se dirige

Está claro que la ética y la inteligencia artificial son dos realidades que se complementan, como ya se ha expuesto, la IA no es mala en sí misma, pero, al ser fruto de seres humanos con prejuicios y que cometen errores, está sujeta a los mismos *bias* que sus creadores. Este escrito ha tratado de presentar de manera general ambos conceptos y encontrar el hilo que los une. Para terminar con el análisis sobre qué hacer respecto a los retos éticos presentados, ante una futura realidad en la que la IA está cada vez más presente en cualquier actividad cotidiana, se pretende aquí exponer cuál es la situación actual y cómo deberían evolucionar las medidas aplicadas al uso de la inteligencia artificial.

Este apartado aborda desde los retos actuales de desarrollo que tiene la IA, qué posición con respecto a la misma se está tomando y hacia dónde va su evolución desde un punto de vista social y ético. Como se ha visto a lo largo de este escrito, junto con todas las ventajas que plantea la IA, esta también plantea cuestiones políticas y sociales que deben tenerse en cuenta.

Desde un punto de vista micro, cualquier sistema de inteligencia artificial está expuesto a sesgos en los datos que analiza. Estos sesgos se amplificarán en el momento en el que el sistema actúe en consecuencia a los mismos y estos se materialicen en

acciones discriminatorias de cualquier índole (Haenlein & Kaplan, 2019). Existen muchos ejemplos que reafirman esta hipótesis. Los sensores de los coches autónomos tienen mayor facilidad para detectar individuos de piel clara, porque el *input* de datos que utilizan contiene más imágenes de ese tipo (Wilson, Hoffman, & Morgenstern, 2019). Los sistemas que pueden ser utilizados por jueces para decidir sobre el futuro de un preso están sesgados por las anteriores resoluciones con tintes racistas (Angwin, Larson, Mattu, & Kirchner, 2016). No es la IA la que provoca estos sesgos, es los datos que utiliza y cómo es su programación. Esto, en última instancia, es responsabilidad humana y según se vaya avanzando, la responsabilidad irá creciendo de manera proporcional.

La tendencia futura no es la de tratar de regular la IA en sí misma, sino la de establecer algún tipo de regulación o código, que sirva de guía para las actuaciones de los ingenieros y programadoras de la IA. Igual que profesiones como la abogacía o la medicina están sujetas a las normas de los colegios, este nuevo campo de profesiones no debería estar exento de normas que lo controlen (Haenlein & Kaplan, 2019). La parte controlable, sin embargo, no será el cien por cien de la IA desarrollada. El aprendizaje profundo no depende tanto de las personas, en un futuro será prácticamente un ente autónomo, por lo que la regulación de la profesión de los ingenieros de la IA no será suficiente en este caso. Este es uno de los futuros retos más aterradores, cómo regular algo que piensa y toma decisiones por sí mismo, y, que, además, tendrá unos procesos totalmente desconocidos e incomprensibles para el ser humano (Burrell, 2016). Un ejemplo que aterrice esta idea es el uso de tecnología de reconocimiento facial, hoy en día los sistemas reconocen rostros y etiquetar un nombre a los mismos, pero, ¿qué pasará cuando estos sistemas tengan aprendizaje profundo y puedan reconocer síntomas de enfermedades o embarazos? (Haenssle, y otros, 2018). Este desconocimiento sobre qué será capaz de hacer un sistema con un aprendizaje profundo desarrollado, planteará riesgos como un aumento de la corrupción al poder esconder ciertos asuntos, sesgos o pérdida de la privacidad. Se habla ya hoy en día de estos riesgos, pero aún no se ha llegado al punto de enfrentarlos de verdad y un plan de actuación debe estar preparado.

Además de un plan de actuación preparado, la humanidad debe preguntarse cuál es el límite. La tecnología sigue desarrollándose sin pensar en los límites, las implicaciones, los riesgos. Se debe pensar en proteger la privacidad de los datos antes de que sistemas como los mencionados sean comercializados ¿qué son las personas sin sus

secretos? Individuos sin identidad ni dignidad (Ziegler, 2019). Es fundamental que a nivel micro existan políticas nacionales que promuevan sistemas de IA fiables, que fomenten la investigación ética y un desarrollo responsable (OECD, 2019).

Desde un punto de vista macro, por otro lado, cabe cuestionar quién vigila a los vigilantes. Es decir, se propone regular la actividad de la gente que está detrás de la IA y del sector privado, pero, ¿qué ocurre con los gobiernos y agentes internacionales? La inteligencia artificial según el camino hacia el desarrollo que lleva reestructurará las sociedades, creando nuevas maneras de ver las cosas y nuevas preocupaciones. Los gobiernos aprovecharán las nuevas herramientas en su favor, por ejemplo, China ya está trabajando en un sistema de crédito social basado en la IA y el *big data* que avale a los dignos de confianza y deniegue a los que no lo merecen (The Economist, 2016). El proyecto suena bien a priori, sin embargo, ¿dónde queda la privacidad del individuo? ¿en base a qué se decide quién es merecedor de confianza? En resumen, ¿dónde está el límite y quién puede imponérselo a los Estados?

Está claro que la inteligencia artificial supone grandes beneficios y avances en muchos sentidos, pero, hay que hacer un balance entre el desarrollo económico y la privacidad de los individuos. No es lícito pasar por encima de los derechos de privacidad de las personas a costa del progreso económico o el poder, puesto que supondría crear sociedades que estarían bajo el control total y absoluto del Estado. Por ello, una solución que afecte exclusivamente a un área no es válida, la cooperación internacional es necesaria en este ámbito (Haenlein & Kaplan, 2019). ¿De qué sirve que Europa regule y proteja la privacidad de las personas si al usar aplicaciones americanas o chinas pasan a estar bajo su control? En el largo plazo una actuación poco integrada supondrá que la privacidad se pierda de todas maneras, dado que debido a la globalización los datos personales de los individuos viajan por el globo sin límite.

Sólo se pueden hacer suposiciones acerca de lo que supondrá realmente la IA en unos años, o sus avances. No se sabe si traerá una gran cooperación internacional y entre máquinas y humanos potenciando la inteligencia y el desarrollo, o, una tercera guerra mundial como dice Elon Musk. Lo que está claro es que retos que actualmente empiezan a sonar pero se ven lejanos, serán muy reales dentro de unos años y el mundo debe estar preparado para esta realidad (Kaplan & Haenlein, 2019).

"¿Cómo podemos regular una tecnología que está en constante evolución y que sólo comprenden unos pocos? ¿Cómo superar el reto que supone ser lo suficientemente amplios y permitir la evolución y ser lo suficientemente precisos para evitar que todo sea considerado como IA?" (Haenlein & Kaplan, 2019). Este es el verdadero reto que llegará en unos años, el ser humano es capaz de normalizar lo que un día le pareció extraordinario, lo que dificultará una reacción a tiempo.

A la vista de un futuro tan incierto y de los potenciales riesgos la IA y su desarrollo se está convirtiendo en una prioridad política cada vez mayor. Está claro que la inteligencia artificial puede ser motor de crecimiento, todos los Estados están de acuerdos en eso, pero la actitud que toman unos y otros es muy distinta (OECD, 2019). Mientras que Estados como China o Estados Unidos intentan eliminar cualquier tipo de barrera al desarrollo y a la implantación de la IA, Europa está estableciendo distintas normativas de regulación como el llamada Reglamento General de Protección de Datos (RGPD) (Haenlein & Kaplan, 2019). Europa es uno de esos ejemplos que intentan educar y abordar los retos vinculados a la IA, mientras que China o Estados Unidos priorizan el desarrollo tecnológico, las ventajas gubernamentales y el crecimiento económico. A nivel internacional, muchos organismos como la OECD, la Comisión Europea o las Naciones Unidas están actuando a nivel regulación. Por ejemplo, en 2019 la OECD adoptó sus Principios sobre Inteligencia Artificial, primer conjunto de normas internacionales acordado por gobiernos para regular de manera responsable la IA (OECD, 2019).

Sobre la regulación existente, debería haber evolución, no nuevos marcos normativos. Pero, para que las políticas sean útiles estas deben ser desarrolladas desde la consideración también de las ventajas de la IA, no sólo lo negativo. Siempre desde el punto de vista de los derechos individuales y la ética, que no debe perderse de vista, un buen marco optimizado puede llevar a un crecimiento sostenible y rápido (Bryson, 2019). La optimización irá llegando según se vaya avanzando en este campo, pero ya están surgiendo debates sobre los derechos y las responsabilidades que un sistema dotado de inteligencia artificial podría tener. Como se mencionaba previamente, los límites de la responsabilidad cuando el sistema cuenta con aprendizaje profundo son difusos. El debate es: si una máquina puede ser responsable de una acción, también debería contar con ciertos derechos "humanos". Bajo esta premisa, no debería permitirse la violación de un humanoide, o comportamientos violentos sobre los mismos (Turner, 2019). Estas

cuestiones serán más relevantes en unos años cuando la tecnología haya avanzado, al igual que la regulación. Pero, da una pequeña visión de los debates que habrá en el futuro en este campo.

Más allá de los retos regulatorios, también existen retos a nivel tecnológico. Se debe mejorar el modelado del comportamiento, ya que la IA convivirá cada vez más con los individuos a nivel cercano. Relacionado, está el reto de mejorar la interacción de la máquina con el hombre, para que estas y sus procesos sean más comprensibles a los individuos. Por último, se debe mejorar la verificabilidad del software, para que los comportamientos de las máquinas se ajusten a los requerimientos (CSIC, 2019). Estas condiciones son fundamentales también para que la convivencia entre humanos y la IA sea apacible y natural.

Tras este último apartado sobre la situación actual y la dirección futura, se cierra el punto de actualidad de la IA y la ética. Para terminar se plantean unas conclusiones sobre el escrito al completo.

8. Conclusiones

Este apartado de conclusiones reflejará los principales puntos que se han ido analizando a lo largo de este escrito. Los hipotéticos retos éticos que exigen tomar una decisión ficticia sobre realizar una acción u otra que suponen la muerte de cientos o la muerte de uno sólo ya están cerca de dejar de ser ficticios. En un mundo en el que los coches autónomos toman las carreteras y las máquinas tienen capacidad de pensar por si mismas, estas decisiones formarán parte del día a día, y, aquellos que están detrás de esa tecnología deberán estar preparados para estos retos (Jarvis Thomson, 1976) (Awad, y otros, 2018). Por este motivo, la ética es fundamental en el desarrollo futuro de la inteligencia artificial.

A lo largo de este trabajo se ha ido viendo la importancia de la relación entre la ética y la inteligencia artificial. Cualquier tecnología desarrollada por seres humanos nunca es neutral, puesto que hereda los sesgos de sus creadores. Esto hace necesario que la inteligencia artificial se desarrolle bajo unos principios éticos básicos, para minimizar

los riesgos que puede suponer un mal uso de la IA y poder aprovechar sus ventajas y beneficios.

Otro punto de conclusión es que las sociedades tendrán que adaptarse y cambiar la manera de vivir, por la introducción de la IA en las vidas cotidianas. El entorno económico, los mercados, el ámbito laboral, las relaciones sociales, las campañas políticas, la medicina, todo cambia, para bien o para mal, con la implicación de la IA. Los seres humanos necesitan ser educados en el buen uso de estas tecnologías, puesto que pueden suponer el aumento de desigualdad y discriminaciones. Por ello, es fundamental que a nivel internacional los códigos sobre IA sean consensuados y todos los Estados actúen en la misma dirección.

La tecnología influye en ámbitos que tienen como uno de sus máximos componentes la moralidad y la ética, la tecnología debe adaptarse al ritmo de los mismos, para no exacerbar la discriminación y la desigualdad. Hoy la revolución es sesgada, por el rápido ritmo que lleva, por quienes la controlan y por el uso que se le da. Reproduce estereotipos que excluyen a minorías, y, sino se adaptan estas tecnologías a los nuevos valores, solo se fomentará la fracción de la sociedad, una sociedad menos humana, menos democrática, y, en definitiva, en retroceso.

La ciencia representa la posibilidad de reconstruir los límites de la vida cotidiana, puede ser tanto algo muy beneficioso como algo que amenace la existencia humana. La tecnología redefine las relaciones y la identidad humana ¿se perderá la humanidad para ser otra cosa más influenciada por la tecnología?

La tecnología debe ser sostenible y respetar los límites morales y éticos. Debe unir, no fragmentar. Debe crear, no destruir. Pero, eso está en la mano del ser humano y cómo lo gestione. Querer cambios rápidos e instantáneos para mantener el ritmo de desarrollo exacerbado que llevamos, no lleva a ningún buen puerto, pues, habrá algún momento que ese desarrollo sea incluso más rápido que nosotros y no podamos controlarlo ni pararlo. Hay que cambiar de modelo y no dejar atrás a nadie.

"El mundo necesita nuevas ideas, ideas que fomenten las bases éticas de la cuarta revolución industrial" (Férnandez de la Vega, 2019).

9. Bibliografía

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (23 de Mayo de 2016). *Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks.* Obtenido de ProPublica: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Argandoña, A. (25 de Marzo de 2019). *Ética e inteligencia artificial (I)* . Obtenido de IESE : https://blog.iese.edu/antonioargandona/2019/03/25/etica-e-inteligencia-artificial-i/

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., . . . Rahwan, I. (2018). The Moral Machine experiment. *Nature*.

Banafa, A. (7 de Agosto de 2016). *¿Qué es el aprendizaje profundo?* Obtenido de Open Mind, BBVA: https://www.bbvaopenmind.com/tecnologia/mundo-digital/que-es-el-aprendijaze-profundo/

Banko, M., & Brill, E. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. En ACL, *ACL Proceedings: Association for Computational Linguistics.* Association for Computational Linguistics.

BBC Mundo. (2 de Julio de 2015). *Google pide perdón por confundir a una pareja negra con gorilas*. Obtenido de BBC: https://www.bbc.com/mundo/noticias/2015/07/150702_tecnologia_google_perdon_confundir_afroamericanos_gorilas_lv

BBC Mundo. (21 de Marzo de 2018). *5 claves para entender el escándalo de Cambridge Analytica que hizo que Facebook perdiera US$37.000 millones en un día*. Obtenido de BBC: https://www.bbc.com/mundo/noticias-43472797

Bellman, R. (1978). *Artificial Intelligence: Can Computers Think?* Thomson Course Technology.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*, 1573-1576.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence . En K. Frankish, & W. M. Ramsey, *Cambridge Handbook of Artificial Intelligence*. Cambridge University Press.

Bryson, J. J. (2019). La última década y el futuro del impacto de la IA en la sociedad. En B. OpenMind, *¿Hacia una nueva ilustración? Una década trascendente.*

Buchholz, R. A., & Rosenthal, S. B. (2002). Technology and Business: Rethinking the Moral Dilemma. *Journal of Business Ethics* .

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*.

Campbell, M., Hoane, A. J., & Hsu, F.-h. (2002). Deep Blue. *Artificial Intelligence*, 57-83.

Ceballos, J. Á. (2020). *Ética y Responsabilidad Social de la Empresa.*

Charniak, E., & Mcdermott, D. (1985). *Introduction to Artificial Intelligence.* Addison-Wesley.

Cohen, P. R. (1995). *Empirical Methods for Artificial Intelligence.* Bradford Books.

Comisión Europea. (2019). *Directrices éticas para una IA fiable.* Obtenido de European Union Newsroom: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60423

Crevier, D. (1993). *AI: The Tumultuous History of the Search for Artificial Intelligence.* Basic Books.

CSIC. (Junio de 2019). *El impacto de la inteligencia artificial en nuestra sociedad. Retos y oportunidades.* . Obtenido de Somos Consejo Superior de Investigaciones Científicas: https://newsletter.corp.csic.es/el-impacto-de-la-inteligencia-artificial-en-nuestra-sociedad-retos-y-oportunidades/

European Parliamentary Research Service, Scientific Foresight Unit (STOA). (Junio de 2016). *Ethical Aspects of Cyber-Physical Systems*. Obtenido de European Parliament : https://www.europarl.europa.eu/RegData/etudes/STUD/2016/563501/EPRS_STU%282016%29563501_EN.pdf

Eyal, N. (26 de Junio de 2017). *Here's How Amazon's Alexa Hooks You. A four-step model explains the psychology behind what makes the technology so habit-forming.* Obtenido de Inc. : https://www.inc.com/nir-eyal/heres-how-amazons-alexa-hooks-you.html

Férnandez de la Vega, M. T. (21-22 de Noviembre de 2019). The fourht industrial revolution, again without women? *The 4th Industrial Revolution & Ethics Conference*. Madrid.

Fussell, S. (4 de Junio de 2019). *Consumer Surveillance Enters Its Bargaining Phase. Amazon and Google are happy to give users the option to pause tracking. Why can’t we stop it entirely?* Obtenido de The Atlantic :

https://www.theatlantic.com/technology/archive/2019/06/alexa-google-incognito-mode-not-real-privacy/590734/

Goldhill, O. (13 de Abril de 2015). *Why smartphones are making you ill*. Obtenido de The Telegraph: https://www.telegraph.co.uk/technology/news/11532428/Why-smartphones-are-making-you-ill.html

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. *California Management Review*.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., . . . Uhlmann, L. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea.* MIT Press.

Hawksworth, J., Berriman, R., & Goel, S. (2017). *Will robots really steal our jobs? An international analysis of the potential long term impact of automation.* Obtenido de PWC: https://www.pwc.com/hu/hu/kiadvanyok/assets/pdf/impact_of_automation_on_jobs.pdf

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid.*

Huang, M.-H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*.

Hutson, M. (24 de Mayo de 2018). *How researchers are teaching AI to learn like a child*. Obtenido de Science: https://www.sciencemag.org/news/2018/05/how-researchers-are-teaching-ai-learn-child

IAT. (2020). *Inteligencia artificial: qué es, tipos, técnicas, ventajas*. Obtenido de IAT: https://iat.es/tecnologias/inteligencia-artificial/

Iberdrola. (2019). *¿Somos conscientes de los retos y principales aplicaciones de la Inteligencia Artificial?* Obtenido de ¿Qué es la inteligencia artificial?: https://www.iberdrola.com/innovacion/que-es-inteligencia-artificial

Iberdrola. (2020). *Industria 4.0: ¿qué tecnologías marcarán la Cuarta Revolución Industrial?* Obtenido de Cuarta Revolución Tecnológica, Iberdrola: https://www.iberdrola.com/innovacion/cuarta-revolucion-industrial

Jarvis Thomson, J. (Abril de 1976). Killing, letting die, and the trolley problem . *The Monist*.

Kaplan, A. M., & Haenlein, M. (2019). Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence. *Business Horizons* .

Kaplan, A., & Haenlein, M. (2019). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons* .

Koo, C. W., Anand, V., Girvin, F., Wickstrom, M. L., Fantauzzi, J. P., Bogoni, L., . . . Ko, J. P. (2012). Improved Ef ciency of CT Interpretation Using an Automated Lung Nodule Matching Program. *American Journal of Roentgenology*.

Kurzweil, R. (1992). *The Age of Intelligent Machines.* MIT Press.

Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. *Artificial Intelligence* .

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

Marín García, S. (2019). Ética e inteligencia artificial . *Cuadernos de la Cátedra Caixabank de Responsabilidad Social Corporativa*.

Martin, K. E., & Freeman, R. E. (2004). The Separation of Technology and Ethics in Business Ethics. *Journal of Business Ethics, 53*, 353-364.

McCarthy, J. (12 de Noviembre de 2007). *What is artificial intelligence?* . Obtenido de Stanford University: http://jmc.stanford.edu/articles/whatisai.html

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine* .

McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence.* Routledge.

McDermott, J. (1980). *R1: A rule-based configurer of computer systems.* Carnegie-Mellon University .

McKinsey&Company. (13 de Noviembre de 2018). *AI adoption advances, but foundational barriers remain*. Obtenido de McKinsey&Co: https://www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain

Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry.* MIT Press.

Mira, J., & Delgado, A. E. (1995). *Aspectos básicos de la inteligencia artificial.* Sanz y Torres.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*.

Nilsson, N. J. (1998). *Artificial Intelligence: A New Synthesis.* Morgan Kaufmann Publishers Inc.

OECD. (2019). *Artificial Intelligence in Society.* Paris: OECD Publishing.

Palma, J., & Marín, R. (2008). *Inteligencia artificial. Técnicas, métodos y aplicaciones.* McGraw-Hill Interamericana de España S.L.

Piketty, T. (2015). About Capital in the Twenty-First Century . *American Economic Review* .

Poole, D., & Goebel, R. (1998). *Computational Intelligence: A Logical Approach.* OUP USA.

Quinn, M. (2016). *Ethics for the Information Age.* Pearson.

RAE. (2020). *Ética*. Obtenido de Diccionario de la Lengua Española: https://dle.rae.es/ético

Raffino, M. E. (24 de Junio de 2020). *Ética*. Obtenido de Concepto.de: https://concepto.de/etica/

Real Academia Española . (2019). *Inteligencia Artificial* . Obtenido de Diccionario de la Lengua Española : https://dle.rae.es/inteligencia

Real Academia Española. (2019). *Inteligencia* . Obtenido de Diccionario de la Lengua Española : https://dle.rae.es/inteligencia

Redacción APD. (8 de Marzo de 2019). *Los cuatro tipos de inteligencia artificial que debes conocer*. Obtenido de Asociación para el Progreso de la Dirección (APD) : https://www.apd.es/tipos-de-inteligencia-artificial/

Rich, E., & Knight, K. (1991). *Artificial Intelligence.* McGraw-Hill Education.

Russel, S., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach (3.ª edición)* . New Jersey: Pearson Education, Inc.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., . . . Kalchbrenner, N. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*.

Simon, H. (1965). *The shape of Automation for Men and Management.* Harper.

Skilton, M., & Hovsepian, F. (2018). *The 4th industrial revolution. Responding to the impact of the artificial intelligence on business* . Palgrave macmillan.

Tanguay, G. (21-22 de Noviembre de 2019). Addressing the ethical and social challenges of emerging technologies - Creating the conditions to play a leadership role in the 4IR. *The 4th Industrial Revolution & Ethics Conference* . Madrid.

The Economist. (17 de Diciembre de 2016). *Big data, meet Big Brother: China invents the digital totalitarian state. The worrying implications of its social-credit project.* Obtenido de The Economist: https://www.economist.com/briefing/2016/12/17/china-invents-the-digital-totalitarian-state

Turner, J. (21-22 de Noviembre de 2019). Artificial intelligence - we have to get the foundation pillars right for our future. *The 4th Industrial Revolution & Ethics Conference*. Madrid.

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

Wilson, B., Hoffman, J., & Morgenstern, J. (21 de Febrero de 2019). *Predictive Inequity in Object Detection.* Obtenido de Cornell University: https://arxiv.org/pdf/1902.11097.pdf

Wilson, H. J., & Daugherty, P. R. (Julio de 2018). *Collaborative Intelligence: Humans and AI Are Joining Forces*. Obtenido de Harvard Business Review : https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces

Winston, P. H. (1992). *Artificial Intelligence.* Pearson.

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. En N. Bostrom, & M. M. Ćirković, *Global Catastrophic Risks.* Oxford University Press.

Ziegler, K.-C. (21-22 de Noviembre de 2019). The requirement of freedom of thoughts in a digital future. *The 4th Industrial Revolution & Ethics Conference*. Madrid.