FACULTAD DE DERECHO

PROBLEMAS ÉTICOS EN TORNO A LA DESCARGA DE DECISIONES EN MÁQUINAS CON ALGORITMOS DE INTELIGENCIA ARTIFICIAL

Autora: Virginia Porrúa Perea

5º E3 Analytics

Tutor: Raúl González Fabre

Madrid

Marzo 2022

RESUMEN

La Era Digital ha supuesto el auge y la proliferación de los algoritmos de Inteligencia Artificial, permitiendo el desarrollo de máquinas y sistemas inteligentes capaces de optimizar la respuesta ante problemas concretos de la sociedad. No obstante, el uso generalizado de esta tecnología plantea problemas de imparcialidad, privacidad, explicabilidad y rendición de cuentas, lo que pone de manifiesto la necesidad de determinar el componente ético exigido en la toma de decisiones por las máquinas que emplean algoritmos de IA. El presente trabajo de investigación ha demostrado que la forma más eficaz de garantizar la seguridad de los sistemas inteligentes es incorporar la tolerancia, la supervisión humana, la responsabilidad y el respeto por el bienestar de la humanidad en su programación y configuración. La importancia de este análisis puede apreciarse en los esfuerzos de múltiples organismos internacionales y gigantes tecnológicos por desarrollar una ética algorítmica capaz de gobernar la IA, el Big Data y el uso de máquinas automatizadas.

Como contribución a este necesario debate, se ofrece un análisis regresivo de bases de datos centrado en el caso particular de los vehículos autónomos. Este estudio ha demostrado que la determinación del componente ético exigido en la toma de decisiones por estos vehículos es extremadamente compleja, ya que implica buscar un equilibrio entre intereses que a primera vista pueden resultar conflictivos, como la seguridad del ser humano y la moralidad del algoritmo.

Palabras clave: inteligencia artificial, ética, vehículos autónomos, algoritmo, supervisión humana, automatización.

ABSTRACT

The Digital Era has led to the rise and proliferation of Artificial Intelligence algorithms, enabling the development of intelligent systems capable of optimizing the response to specific human problems. Nonetheless, the widespread use of this technology raises issues of fairness, privacy, explainability and accountability, highlighting the need to determine the ethical component required in decision making by machines using AI algorithms. This thesis has demonstrated that the most effective way to ensure the safety of intelligent systems is to incorporate tolerance, human oversight, accountability and respect for the welfare of humanity in their programming and configuration. The importance of this analysis can be seen in the efforts of international institutions and technological organisations to develop algorithmic ethics capable of governing AI, Big Data and the use of automated machines.

As a contribution to this necessary debate, a regression analysis of databases focusing on the particular case of autonomous vehicles is offered. This study has shown that the determination of the ethical component required in decision making by these vehicles is extremely complex, due to the fact that it involves seeking a balance between interests that at first sight may be conflicting, such as the safety of the human being and the morality of the algorithm.

Key words: artificial intelligence, ethics, autonomous vehicles, algorithm, human supervision, automation.

ÍNDICE

ÍNDICE DE FIGURAS Y TABLAS

LISTADO DE ABREVIATURAS

EEUU	Estados Unidos
IA	Inteligencia Artificial
ICRAC	Comité Internacional para el Control de Armas Robóticas
IoE	Internet of Everything o Internet de Todo
IoT	Internet of Things o Internet de las Cosas
NLP	Procesamiento del Lenguaje Natural
OCDE	Organización para la Cooperación y el Desarrollo
SAE	Sociedad de Ingenieros de Automoción Económicos
UE	Unión Europea
UNESCO	Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

1. INTRODUCCIÓN

1.1. Estado de la cuestión

El S.XXI ha supuesto el auge de la Era Digital, caracterizada por la velocidad de la información y la proliferación de sistemas inteligentes que se integran en la vida cotidiana de las personas y grandes organizaciones. Ello ha dado lugar a lo que muchos conocen como la Cuarta Revolución Industrial o Industria 4.0, un concepto introducido por el economista y empresario alemán Klaus Schwab, fundador del Foro Económico Mundial. En su obra *The Fourth Industrial Revolution*, Schwab define la revolución industrial como "*el surgimiento* de *nuevas tecnologías y nuevas maneras de percibir el mundo que impulsan un cambio profundo en la economía y la estructura de la sociedad*" (Schwab, 2016). En la Industria 4.0, estas nuevas tecnologías son, entre otras, el Internet de las Cosas, la robótica, la analítica, el Big Data, la nanotecnología y, esencialmente, la Inteligencia Artificial, IA en adelante.

Tal y como ocurrió con las tres primeras revoluciones industriales, la Industria 4.0 no solo afecta a los procesos de fabricación y producción, sino que tendrá – y ya está teniendo – un impacto fundamental en la sociedad. Por un lado, la digitalización implica una mejora sin precedentes en la productividad, la calidad y en la eficiencia de los sistemas de fabricación, una consecuencia directa del ahorro en tiempo y costes que supone la automatización. Sin embargo, la revolución digital va mucho más allá, de modo que ha logrado infiltrarse en los diferentes ámbitos de nuestra sociedad (entre otros, la educación, empleo, prestación de servicios o sanidad). A tal efecto, la obtención masiva de datos por parte de máquinas que emplean algoritmos de IA ha permitido a las empresas anticiparse a las necesidades de sus clientes, proporcionándoles no solo una asistencia instantánea, sino también productos y servicios personalizados. Asimismo, el avance tecnológico ha permitido desarrollar prácticas educativas innovadoras, orientadas a potenciar el aprendizaje y a acelerar el proceso en la consecución del cuarto Objetivo del Desarrollo Sostenible[1] (UNESCO, 2021). Por su parte, el aprendizaje profundo, comúnmente conocido como Deep Learning, está transformando la sanidad, de modo que las grandes organizaciones

[1] El 25 de septiembre de 2015, los líderes mundiales aprobaron 17 objetivos orientados a proteger el planeta, acabar con la pobreza y garantizar el desarrollo sostenible. Concretamente, el cuarto objetivo está orientado a asegurar una educación de calidad, equitativa e inclusiva capaz de promover una oportunidad de aprendizaje a nivel mundial.
A tal efecto, la UNESCO aboga por la implementación de tecnologías de IA en la educación, con el fin de promover el aprendizaje mediante la colaboración entre los educadores y las máquinas que emplean algoritmos inteligentes.

sanitarias emplean algoritmos en el diagnóstico por imagen para la detección temprana de enfermedades, el análisis en tiempo real de ciertas anomalías o para facilitar la toma de decisiones clínicas (IBM, 2021).

Sin perjuicio del optimismo derivado de las múltiples ventajas que conlleva el uso de máquinas que emplean algoritmos de IA, el desarrollo de esta tecnología ha suscitado desde el principio una preocupación fundamental: la determinación del componente ético exigido en la toma de decisiones por las mismas.

Históricamente se ha demostrado que las nuevas tecnologías generan numerosos beneficios para la sociedad, si bien también pueden ser empleadas de tal forma que resulten nocivas para el ser humano. Véase el caso de la fabricación de armas de fuego, para lo que se empleó la pólvora, o de la confección de la bomba atómica, derivada del descubrimiento de la energía nuclear (Marín, 2019). La IA no será una excepción e incluso hay quienes temen que la automatización de las máquinas pueda constituir el inicio del fin de la era del control humano sobre la Tierra. Si bien es cierto que esta afirmación puede resultar desproporcionada, Stephen Hawking advirtió a finales del año 2014 que "*el desarrollo de una completa inteligencia artificial podría traducirse en el fin de la raza humana*" (BBC News Mundo, 2014). Por ello, para expertos en la materia como Douglas Rushkoff, no debemos preguntarnos si la revolución tecnológica resulta perjudicial o no para la humanidad, sino si queremos dirigir la tecnología o que ésta nos dirija a nosotros. En palabras del teórico de nuevos medios, se trata de "*programar o ser programados*" (Rushkoff, 2010). Por su parte, como se analizará en detalle más adelante, el uso de algoritmos puede derivar en problemas de imparcialidad, privacidad, explicabilidad y rendición de cuentas. Por consiguiente, el principal reto al que se enfrenta el ser humano como consecuencia de la proliferación de esta tecnología es de tipo ético, lo que ha puesto de manifiesto la necesidad de incorporar la tolerancia y el respeto por el bienestar de la humanidad en la programación y configuración de los agentes de IA. Es lo que Eliezer Yudkowsky ha denominado como "*Inteligencia Artificial amigable*" (Monasterio Astobiza, 2017), un término respaldado por Peter Norvig, director de investigación de Google, quien señala que, si bien es cierto que la IA presenta múltiples beneficios para el hombre, es importante asegurarse de que todo el mundo se beneficia de esta tecnología (Lufkin, 2017).

1.2. Objetivos

Teniendo todo ello en consideración, el presente trabajo de investigación tiene por objeto analizar las implicaciones éticas derivadas de la toma de decisiones por máquinas que emplean algoritmos de IA. En concreto, a través de estas páginas se intentará determinar el componente ético exigido en el uso de esta tecnología, ya que ello se configura como un asunto de vital importancia tanto para organismos e instituciones internacionales como para los gigantes tecnológicos, tales como Google, Microsoft, Apple, Facebook, Amazon, IBM y Deep Mind, que abogan por el desarrollo de una ética algorítmica que gobierne la IA, el Big Data y el uso de máquinas automatizadas.

Sin perjuicio de ello, con carácter previo al estudio de las implicaciones éticas derivadas del uso de sistemas inteligentes, resulta fundamental realizar un desarrollo conceptual de la IA, analizando sus clases, su origen y sus beneficios y riesgos potenciales.

1.3. Metodología

Para conseguir este objetivo, se recurrirá a la revisión bibliográfica, que permite analizar en profundidad el concepto y desarrollo de la IA, los dilemas éticos derivados del uso de esta tecnología y las propuestas nacionales e internacionales orientadas a la consecución de la *Inteligencia Artificial amigable*. Dado que se trata de examinar las implicaciones éticas derivadas de máquinas inteligentes con el fin de hallar la forma de lograr su uso responsable, transparente y seguro, esta metodología perece la más adecuada, toda vez que permite al investigador recopilar la información necesaria para determinar los riesgos y beneficios potenciales de la IA.

Por su parte, esta revisión bibliográfica se complementa con un análisis regresivo de bases de datos centrado en el caso particular de los vehículos autónomos, para lo que se recurrirá al lenguaje de programación R. Ello permitirá determinar el componente ético exigido en la toma de decisiones por estas máquinas autónomas, así como evaluar la aceptación por la sociedad de una regulación gubernamental.

1.4. Desarrollo

Para concluir, este trabajo de investigación académica se estructura en cuatro partes. En la primera se presenta la justificación del tema, centrada en el análisis del concepto y el desarrollo de la IA, así como en sus beneficios y riesgos potenciales. En un segundo apartado se describen las implicaciones éticas derivadas de la toma de decisiones por máquinas que emplean algoritmos de IA. El tercer bloque repasa las prácticas éticas

definidas a nivel nacional e internacional por instituciones, organismos y grandes empresas. Finalmente, la cuarta y última parte se configura como una revisión en profundidad del caso particular de los vehículos autónomos.

2. LA INTELIGENCIA ARTIFICIAL

2.1. Concepto y tipos de Inteligencia Artificial

Con el objeto de comprender los beneficios e implicaciones éticas de la IA, resulta fundamental realizar una delimitación conceptual de este término. No obstante, cabe destacar que actualmente no existe una definición globalmente aceptada o consensuada de esta tecnología. Sin perjuicio de las referencias previas realizadas por Alan Turing, el concepto de Inteligencia Artificial fue acuñado en 1956 por John McCarthy, quien, en su obra "*What is artificial intelligence?*" la define como "*la ciencia e ingenio de hacer máquinas inteligentes y, especialmente, programas de ordenador inteligentes. Se aproxima a la tarea de emplear ordenadores para comprender la inteligencia humana, si bien la IA no se restringe a métodos biológicamente observables*" (McCarthy, 2004).

Por su parte, Peter Norvig y Stuart Russell, en su obra *Artificial Intelligence: A Modern Approach*, tratan de construir el concepto de IA partiendo del objetivo perseguido por ésta. En función del mismo, Norvig y Russell clasifican las definiciones en cuatro categorías diferentes (2016):

i. **Sistemas que piensan como humanos**: tratan de desarrollar máquinas capaces de razonar como un ser humano, lo que permitiría la automatización de funciones mentales como la toma de decisiones, el aprendizaje o la resolución de problemas (Bellman, 1978). Este es el caso de las redes neuronales artificiales, que buscan imitar el pensamiento humano. Bajo esta concepción, la inteligencia se expresa en el pensamiento.

ii. **Sistemas que actúan como humanos**: la IA permite desarrollar máquinas capaces de llevar a cabo tareas que requieren capacidades analíticas humanas de manera más eficiente. Este es el caso de la robótica o de los androides, que tratan de imitar el comportamiento humano. Para Elaine Rich y Kevin Knight, la IA se define como "*el estudio de cómo lograr que los ordenadores realicen tareas que, por el momento, las personas hacen mejor*". Por consiguiente, la inteligencia se expresa en la acción. Esta es la concepción defendida por Alan

Turing en su experimento el *Turing Test*, el cual será objeto de estudio más adelante.

iii. **Sistemas que piensan racionalmente**: estudian y analizan la lógica del ser humano para razonar, percibir y actuar como tal (Winston, 1992). Este es el caso de los sistemas expertos, que imitan el proceso de razonamiento humano. Para Norvig y Russell, esta concepción de la IA parte de las leyes del pensamiento propias de la Antigua Grecia, en virtud de las cuales una conclusión debe derivarse de dos o más premisas. Ello es lo que se conoce como silogismo, un argumento válido derivado del razonamiento lógico. Por ejemplo, Aristóteles afirmó que "*Todos los hombres son mortales, Sócrates es un hombre, por lo tanto, Sócrates es mortal*" (Norvig y Russell, 2016, p. 4). Como consecuencia de todo ello, los defensores de esta concepción de la IA, como es el caso de Charniak o McDermott, abogan por el desarrollo de sistemas capaces de emular este proceso de razonamiento humano. Por ende, la inteligencia se expresa en el pensamiento.

iv. **Sistemas que actúan racionalmente**: son aquellos que permiten diseñar agentes inteligentes, es decir, programas que actúan racionalmente para la consecución del mejor resultado posible en cada situación concreta (Poole, 1998). Ello implica que la IA trata de imitar el comportamiento humano racional, es decir, actuar de forma lógica para tomar buenas decisiones. Por consiguiente, bajo esta concepción, la inteligencia se expresa en la acción.

Teniendo estas cuatro categorías en consideración, podemos afirmar que existen fundamentalmente dos tipos de IA: la general, que se corresponde con los sistemas que piensan o actúan como humanos, y la específica, referente a aquellos que piensan o actúan racionalmente. Esta distinción se atribuye al filósofo John Searle, quien en un artículo de 1980 distinguió entre IA fuerte, basada en una inteligencia general, e IA débil, basada en una inteligencia específica (López de Mántaras, 2015).

2.1.1. Inteligencia Artificial general o fuerte

La IA fuerte es aquella que tiene por objeto "*replicar la inteligencia humana mediante máquinas*" (López de Mántaras, 2015). Concretamente, se busca replicar una inteligencia de tipo general, de modo que esta tecnología permitiría al ordenador pensar igual que un ser humano. Por consiguiente, la IA fuerte implicaría que la máquina no busca emular el comportamiento humano, sino que va más allá, pasando de simular una

mente a convertirse propiamente en una. Esta inteligencia general es lo que permite hablar de máquinas totalmente autónomas, es decir, aquellas capaces de actuar por sí mismas.

No obstante, son muchos los autores que defienden que, a día de hoy, la IA fuerte queda reservada a la ciencia ficción, es decir, no existe ninguna máquina capaz de tener una inteligencia igual o superior a la del ser humano (López de Mántaras, 2015). A tal efecto, Searle intenta demostrar que esta tecnología es imposible partiendo de la base de que, si bien es cierto que tanto la mente humana como los programas computacionales presentan una estructura formal, las mentes son semánticas, es decir, tienen contenido e intencionalidad, mientras que las máquinas carecen de sentido común (Searle, 1986). Esta postura se enfrenta a la concepción de IA sostenida por Alan Turing, para quien el hecho de que las respuestas proporcionadas por un ordenador no sean distinguibles de las del ser humano, demuestra que las máquinas son inteligentes y poseen una mente (Russell y Norvig, 2016).

2.1.2. Inteligencia Artificial específica o débil

Por el contrario, la IA específica o débil es aquella que, lejos de replicar la inteligencia humana, trata de optimizar la respuesta ante situaciones concretas. En palabras de Searle, consistiría en "*construir programas que ayudan al ser humano en sus actividades mentales en lugar de duplicarlas*" (López de Mántaras, 2015). Por consiguiente, la IA débil dota a las máquinas computacionales de una inteligencia específica, es decir, aptitudes concretas requeridas en la realización de una tarea determinada. Ello permite a ordenadores y otros dispositivos resolver problemas específicos mejor que el ser humano, de modo que la IA débil es aquella cuyas aplicaciones son empleadas en nuestra vida cotidiana – como Siri, Alexa o Google Assistant.

Por todo ello, lo que interesa en el presente trabajo de investigación es la IA específica dado que, a mi parecer, en línea con López de Mántaras, John Searle y muchos otros autores, es la única que el ser humano ha sido capaz de programar.

2.2. Origen de la Inteligencia Artificial

La historia de la IA se remonta a 1950, año en el que Alan Turing, uno de los padres de la computación y pioneros de la informática moderna, introdujo tímidamente el concepto en su obra *Computing Machinery and Intelligence*. En ésta, Turing sugirió que

si el ser humano era capaz de emplear la razón y la información del exterior para la resolución de problemas, ¿qué impedía que las máquinas pudieran imitar este comportamiento? (Anyoha, 2017). Para demostrar esta premisa, el informático desarrolló el denominado *Turing Test*, un experimento en virtud del cual un investigador, mediante la formulación de preguntas, debía tratar de determinar si las respuestas de texto procedían del ser humano o de un ordenador. Así, Turing había introducido el concepto de máquinas inteligentes capaces de facilitar la toma de decisiones por la sociedad y de emular el comportamiento humano.

Sin perjuicio de ello, el nacimiento de la IA tal y como la conocemos hoy en día tuvo lugar el verano de 1956, año en el que se celebró la Conferencia de Dartmouth a instancias de John McCarthy, informático de las universidades de Princeton y Stanford. McCarthy convenció a los mejores informáticos estadounidenses para participar en un seminario de dos meses consistente en el análisis de la denominada *artificial intelligence*. El estudio se basaba en la premisa fundamental de que, en principio, todas las características del aprendizaje y de la inteligencia humana pueden describirse de forma tan precisa que las máquinas deberían ser capaces de emularlas (Russell y Norvig, 2016). Esta conferencia derivó en la necesidad de crear un campo para la IA independiente de la informática, la lógica o la matemática, algo que se había puesto de manifiesto dado que ninguna de estas disciplinas disponía de la metodología necesaria para alcanzar el objetivo deseado: crear máquinas autónomas capaces de aprender, comunicarse o razonar como el ser humano.

Las décadas posteriores a la Conferencia de Dartmouth fueron marcadas por un excesivo optimismo respecto a los avances de la IA, de modo que, si bien las expectativas y predicciones de los científicos fueron alcanzándose, muchas llegaron con más de 40 años de retraso. Entre 1957 y 1974 tuvieron lugar importantes avances en esta disciplina, como es el caso de la aparición del *General Problem Solver*, de Allen Newell y Herbert Simon, o del *ELIZA*, de Joseph Weizenbaum[2].

Sin perjuicio de ello, durante la década de los 70, la IA se vio frenada por dos obstáculos: la limitada capacidad de almacenamiento y procesamiento de los

[2] Por un lado, el General Problem Solver se configuró como un programa de ordenador que permitía resolver problemas sencillos del ser humano, como probar teoremas, jugar al ajedrez, trabajar con la lógica o dar respuesta a cuestiones matemáticas o geométricas.
Por su parte, ELIZA fue uno de los primeros programas de procesamiento del lenguaje natural, NLP, de modo que empleaba palabras clave para entablar conversaciones de texto con el usuario.

ordenadores de la época y la escasez de fondos para hacer frente a los elevados costes derivados de la investigación científica. Es lo que se conoce como el *Invierno de la IA* (del inglés *AI Winter*), un periodo caracterizado por la escasez de inversión e investigación en esta disciplina. Esta problemática se resolvió en los años 80 gracias a las inversiones masivas procedentes esencialmente del Gobierno japonés, que invirtió aproximadamente 400 millones de dólares en IA entre 1982 y 1990 (Anyoha, 2017). Ello derivó en la aparición de los *sistemas expertos*, que simulaban el procedimiento de toma de decisiones por el ser humano mediante la formulación de reglas lógicas.

No obstante, entre los años 1987 y 1993 el mundo de la IA experimentó su segundo invierno como consecuencia del fracaso de los sistemas expertos, derivado de la imposibilidad de cumplir con los objetivos fijados. A partir de entonces, como ocurre tras cada periodo de *AI Winter*, el campo de la IA experimentó un resurgimiento marcado por el logro de muchos de los objetivos fijados tras la Conferencia de Dartmouth. A tal efecto, en 1997, *Deep Blue*, un programa de ajedrez desarrollado por IBM, logró vencer a Gary Kasparov, experto y campeón del mundo en esta disciplina. Ese mismo año, Windows implementó el *Dragon System*, el primer software de reconocimiento de voz. Dos décadas después de la derrota de Kasparov, *AlphaGo*, un programa de IA desarrollado por Google, logró vencer a Lee Sedol en el Go, un juego estratégico caracterizado por la exigencia de pensamiento crítico e intuición. Esta victoria demostró que el ser humano no puede anticiparse a los movimientos del programa, de modo que la IA puede aprender por sí misma (Coeckelbergh, 2020).

Por su parte, el S.XXI ha venido marcado por la implementación de algoritmos de IA en la vida cotidiana de la sociedad, con gran utilidad en el tratamiento de datos, el diagnóstico médico y, esencialmente, en las cadenas de producción y fabricación. Así, el desarrollo de la IA ha permitido que a día de hoy contemos con vehículos autónomos, juegos de realidad virtual, sistemas de planificación logística, robótica, traductores y programas de reconocimiento de voz, entre muchas otras aplicaciones (Russell y Norvig, 2016). Como se puede apreciar, prácticamente todo se encuentra afectado por algoritmos de IA, lo que nos permite hablar del Internet de Todo o IoE, un concepto desarrollado en el apartado siguiente.

2.3. Implicaciones derivadas de una sociedad gobernada por la IA

Atendiendo al concepto de IA previamente analizado, esta tecnología permite el procesamiento de datos y la automatización de las máquinas, lo que ha derivado en la

omnipresencia del Internet y otras tecnologías similares. En la última década hemos asistido al desarrollo del Internet de las Cosas o IoT, que permite que los dispositivos estén conectados entre sí y simultáneamente a Internet sin que ello requiera la intervención humana, lo que deriva en una interacción de máquina a máquina.

No obstante, el IoT ha quedado desfasado dado que el avance tecnológico ha supuesto la aparición del IoE, que permite establecer conexiones a cualquier nivel, ya sea máquina-máquina o máquina-persona (Monasterio, 2017). IoE supone que hoy en día no hay nada que no esté afectado por la IA, de modo que los vehículos han sido automatizados, las casas domotizadas y los procesos digitalizados. Esta nueva visión del mundo se traduce en un aumento de la comodidad, una reducción de costes y una mayor productividad, si bien IoE ha incrementado la vulnerabilidad del ser humano frente a la calidad e integridad de la información. A tal efecto, los dispositivos electrónicos dotados de IA recopilan cantidades ingentes de datos a tiempo real, lo que supone una amenaza para la privacidad y seguridad de los usuarios. Por ello, sin entrar, por el momento, en sus implicaciones éticas, resulta crucial analizar en detalle los beneficios y riesgos potenciales derivados del uso de IA.

2.3.1. Beneficios potenciales de la Inteligencia Artificial

De acuerdo con la comunicación de la Comisión Europea de 25 de abril de 2018, la IA resulta beneficiosa para las personas, las empresas y las administraciones, contribuyendo a la mejora de la asistencia sanitaria, el transporte, la comunicación, los procesos de producción y logística, la fabricación de energía más barata y sostenible, etc.

En primer lugar, la IA ha revolucionado el mundo de la asistencia sanitaria, prestando soporte a las decisiones clínicas, facilitando las imágenes médicas y favoreciendo la equidad sanitaria (IBM, s.f.). A tal efecto, esta tecnología ha permitido desarrollar algoritmos de Procesamiento del Lenguaje Natural (NLP), que identifican y aíslan datos significativos o anomalías en las radiografías de los pacientes, lo que supone un diagnóstico temprano de enfermedades y una mayor facilidad a la hora de tomar decisiones médicas. Prueba de ello es el *Watson Health*, un agente virtual desarrollado por IBM que, entre sus múltiples aplicaciones, ha sido empleado en la determinación del tratamiento más apropiado para más de 1,000 pacientes de oncología. Por su parte, en mayo de 2019, el Laboratorio de Ciencias de la Computación e Inteligencia Artificial del Instituto de Tecnología de Massachusetts, también conocido como CSAIL,

desarrolló un algoritmo capaz de detectar el cáncer de mama con hasta 5 años de antelación (Conner Simons & Gordon, 2019). Para ello, el equipo entrenó un modelo de Deep Learning basado en mamografías y datos procedentes de más de 60,000 pacientes tratados en el Massachusetts General Hospital, lo que permitió al algoritmo identificar patrones en el tejido mamario tendentes a la malignidad imperceptibles al ojo humano.

Asimismo, la IA ha supuesto el auge del reconocimiento de voz, permitiendo a las máquinas y dispositivos procesar el lenguaje humano e interactuar con él. Apple se convirtió en la primera empresa en desarrollar esta tecnología mediante la creación de Siri, un asistente digital alimentado por algoritmos de IA. A éste le siguieron infinidad de asistentes virtuales, como es el caso de Cortana por Microsoft o Alexa por Amazon. El comando de voz ha permitido a las empresas desarrollar novedosas estrategias de marketing y abrir nuevos canales de comunicación con los clientes, proporcionándoles asistencia personalizada y remota en tiempo real. A tal efecto, el banco suizo SEB ha empleado la IA para la creación de Aida, un agente virtual que proporciona a la empresa un canal de comunicación tanto interno como externo, capaz de atender consultas las 24 horas del día y proporcionar soluciones personalizadas. Consecuentemente, esta tecnología no solo presta a los empleados del banco soporte informático, sino que además puede gestionar peticiones de sus clientes, como efectuar transferencias o abrir una nueva cuenta bancaria (Wilson y Daugherty, 2018). Más aún, el reconocimiento de voz no solo presenta una aplicación fundamental en el mundo empresarial, sino que además puede resultar muy ventajoso a la hora de realizar tareas domésticas de la vida cotidiana. Prueba de ello es el desarrollo de Google Home o Amazon Echo, dispositivos que emplean IA para la automatización de labores tan habituales como reproducir música, crear listas de la compra, consultar el tiempo y las últimas noticias o controlar el hogar en remoto. Asimismo, el reconocimiento de voz es también empleado en la asistencia sanitaria. A tal efecto, los servicios de emergencias daneses utilizan la IA para detectar paradas cardíacas u otras afecciones médicas mediante el análisis de la voz de la persona que ha contactado con ellos (Comisión Europea, 2018).

Por su parte, la IA, como elemento fundamental de la Industria 4.0, ha revolucionado el mundo de la logística, permitiendo la automatización de procesos, la minimización de imprevistos y la optimización de las rutas de reparto. De acuerdo con el informe sobre IA elaborado por McKinsey & Company en 2018, "*The promise and challenge of the age of artificial intelligence*", la industria del transporte y de la logística será el segundo

sector más beneficiado por esta tecnología en el corto plazo, realizando una estimación de crecimiento potencial del 89%[3] (Manyika & Bughin, 2018). Asimismo, la robótica está implementándose en los procesos logísticos de empresas como Amazon, Alibaba o Fizyr, una startup que ha empleado algoritmos de Deep Learning para la creación de robots capaces de identificar, recoger y manipular cualquier tipo de mercancía, con independencia de su volumen, color, material o forma.

Junto con éstas, existen muchas otras aplicaciones de la IA, como es el caso del creciente uso de los vehículos autónomos, que será analizado en profundidad en el apartado 5 del presente trabajo. Todas ellas presentan unos beneficios fundamentales tanto a nivel social como empresarial, lo que ha incentivado la inversión masiva en esta disciplina por parte de las grandes organizaciones e instituciones internacionales. A tal efecto, una encuesta realizada en septiembre de 2021 por la consultora estadounidense Gartner, demostró que el 33% de los participantes (todos ellos proveedores tecnológicos) asegura para el 2022 una inversión mínima en el campo de la IA de 1,000,000 de dólares. Asimismo, de los encuestados, un 87% dota a esta disciplina de prioridad en la recepción de fondos (Rimol, 2021).

En definitiva, las aplicaciones de la IA no solo han generado un notable incremento en la calidad de vida de millones de personas, sino que también han revolucionado el mundo empresarial.

2.3.2. Riesgos potenciales del uso de la Inteligencia Artificial

Sin perjuicio de las múltiples aplicaciones de la IA que permiten una mejora en la calidad de vida del ser humano, resulta necesario advertir de los riesgos potenciales que pueden derivarse de su uso. Éstos son, entre otros, los siguientes:

i. **Destrucción del empleo**: la automatización de los procesos productivos conlleva profundos cambios sociales, entre los que podemos destacar, esencialmente, la destrucción de múltiples puestos de trabajo. Un estudio llevado a cabo en 2018 por PriceWaterhouseCoopers demostró que el 30% de los trabajos existentes podrían desaparecer para mediados del 2030 a causa de la digitalización, de los cuales el 44% corresponde a aquellos que requieren empleados con una baja cualificación (PWC, 2018). Ello se debe a que las máquinas que emplean algoritmos de IA son capaces de desempeñar funciones

[3] Superado exclusivamente por el sector turístico, con un crecimiento potencial estimado del 128%.

reiterativas y de escaso valor añadido en menos tiempo y a un menor precio, lo que supone un notable incremento en la productividad empresarial.

Sin perjuicio de ello, hay expertos que se muestran optimistas respecto de estos datos. Dado el concepto de IA analizado, las máquinas simplemente pretenden optimizar la respuesta ante un problema determinado, de modo que no pueden reemplazar al ser humano. A tal efecto, David Plaza, director de Información e Innovación Tecnológica del grupo Adecco, considera que no se trata de una cuestión de creación o destrucción de puestos de trabajo, sino de una transformación del empleo (Munera, s.f.). Por ello, gobiernos, trabajadores y empresas deben trabajar conjuntamente para adaptarse a los cambios que conlleva la proliferación de la robótica. Este proceso de transformación por las sociedades no es sencillo, si bien resulta fundamental a la hora de garantizar que las empresas más tradicionales se adecúan al marco laboral resultante tras las innovaciones tecnológicas derivadas de la Cuarta Revolución Industrial[4]. Por su parte, muchos autores defienden que el desarrollo tecnológico derivará en el nacimiento de nuevas necesidades a satisfacer que requerirán la intervención humana, fomentando así la aparición de nuevos puestos de trabajo (Bessen, 2018).

Los grandes organismos internacionales son conscientes de la necesidad de hacer frente a este reto, de modo que la Comisión Europea, en la citada comunicación de 25 de abril de 2018, presentó una serie de medidas orientadas a convertir la digitalización en una oportunidad laboral, acercando la IA a pequeñas, medianas y grandes empresas. Entre ellas podemos destacar extensos programas de formación sobre competencias digitales, informes sobre el impacto de la IA en el mundo empresarial acompañados de recomendaciones a seguir por los Estados miembros de la UE, programas para atraer y retener el talento y la colaboración en IA, etc. (Comisión Europea, 2018).

ii. **Seguridad y vulnerabilidad digital**: el impulso de la digitalización y la IA ha derivado en la intensificación de los ciberataques, acciones dirigidas contra

[4] En esta línea de pensamiento, la experiencia demuestra que las revoluciones tecnológicas suelen percibirse como una amenaza para el mundo laboral, si bien finalmente acaban generando una gran cantidad de empleo. Véase el caso de Cabify frente a los taxistas, de Airbnb frente a las agencias hoteleras o del sector de las ventas online frente a las compras presenciales en tienda. En estos tres supuestos, la innovación tecnológica inicialmente se percibió como una amenaza para múltiples puestos de trabajo, si bien el tiempo ha demostrado que la coexistencia entre la digitalización y la prestación tradicional de servicios es posible.

bases de datos y otros sistemas de información que pretenden perjudicar a empresas, instituciones y ciudadanos. Iberdrola distingue fundamentalmente cuatro clases de ciberataques; phishing, malware, inyección de SQL y ataque de denegación de servicio (Iberdrola, 2021). Todos ellos tienen un objetivo común; acceder al sistema informático de la víctima, ya sea persona física o jurídica, con el fin de robar datos personales e información confidencial. Entre las brechas de ciberseguridad más importantes se encuentran los hitos de *Wannacry*, que derivó en la paralización de miles de empresas europeas, o *Conficker*, un gusano que infectó a más de 10 millones de equipos Windows en todo el mundo, afectando a miles de departamentos de Seguridad de Estado, hospitales y Fuerzas Armadas.

En esta línea de pensamiento, cada vez son más los expertos en política tecnológica y ciberseguridad que advierten que las guerras tradicionales se verán sustituidas por ataques cibernéticos, configurándose el desarrollo de la IA como una herramienta crucial. A tal efecto, el experto en desarrollo digital Alec Ross apuntó en el *Shapes* de marzo de 2021 que la nube se convertirá en el campo de batalla de los principales conflictos que puedan surgir en los años venideros (Iberdrola, 2021).

iii. **Manipulación de la información**: en gran medida, la proliferación de la IA se debe a la ingente acumulación de datos a través del Big Data. En enero de 2022 Facebook registró más de 3,000 millones de usuarios activos, Twitter unos 383 millones y Google más de 1,000 millones. Con más de 5,000 millones de usuarios conectados a Internet en todo el mundo, en un solo segundo se realizan más de 98,000 búsquedas en Google, se publican casi 10,000 tweets y se envían más de 3,000,000 de correos electrónicos[5]. A todo ello se le suman las millones de transacciones económicas que se realizan a diario en todo el mundo, lo que supone un gran intercambio de información. Asimismo, IoE implica que la IA empleada en los dispositivos electrónicos que nos acompañan en nuestro día a día (el GPS, el Bluetooth o el WIFI) se encuentra constantemente captando y transmitiendo información.

A tal efecto, el principal riesgo deriva del tratamiento que estos datos reciben, ya que la información captada por las máquinas de IA es habitualmente empleada para la generación de patrones dinámicos que identifican tendencias futuras. Si

[5] Toda la información en tiempo real se encuentra disponible en https://www.internetlivestats.com/ (consultada el 18 de enero de 2022).

bien es cierto que ello resulta de gran utilidad a la hora de desarrollar estrategias de marketing, personalizar productos y servicios, agilizar la toma de decisiones o mejorar la visión de negocio, la captación de información no siempre se realiza en beneficio de la sociedad. En esta línea de pensamiento, el *Future of Humanity Institute* de Oxford explica cómo la información obtenida mediante IA proporciona una ventaja en el análisis del comportamiento humano, sus ideologías, preferencias o sentimientos, lo que puede ser empleado, esencialmente por Estados autoritarios, para la manipulación social (Brundage & Avin, 2018).

Prueba de ello es la divulgación de las denominadas *fake news*, término empleado para conceptualizar la manipulación de la información transmitida por los medios de comunicación. La fuerte dependencia actual de las redes sociales ha derivado en una gran vulnerabilidad frente a estas noticias falsas, una peligrosa herramienta de desinformación ocasionalmente empleada por las autoridades para controlar el mensaje que se pretende transmitir. China ejemplifica perfectamente cómo el gobierno filtra y manipula la información antes de trasmitirla a la población. De acuerdo con un artículo publicado por la Universidad de Cambridge y elaborado por tres profesores de las universidades de Harvard, Stanford y San Diego, respectivamente, el régimen chino no solo selecciona minuciosamente las publicaciones que pueden aparecer en la web, sino que además lleva a cabo lo que se conoce como *astroturfing*, un fenómeno que permite al gobierno fabricar opiniones afines a su actual líder, Xi Jinping, para su posterior divulgación en redes sociales (King, Pan & Roberts, 2017)[6]. Sin perjuicio de ello, también los ciudadanos de países democráticos se ven afectados por las *fake news*, como es el caso de los bots políticos empleados en las elecciones generales del Reino Unido y Francia en 2017 o en las de EEUU en 2016 para la difusión en redes sociales de mensajes con una carga claramente ideológica (Polonski, 2017).

[6] Más aun, el Gobierno chino no solo manipula la información disponible, sino que además realiza un meticuloso filtrado de los medios de comunicación que pueden transmitirla. A ello se le conoce como el "*Gran Cortafuegos*" o la "*Gran Muralla de Fuego*", un proyecto llevado a cabo por el Partido Comunista del país que pretende bloquear el acceso a ciertas páginas de Internet para evitar la divulgación de temas especialmente sensibles, tales como la brutalidad policial o la censura de información.
Si bien el régimen chino afirma que este proyecto trata de garantizar la seguridad de Internet para evitar la propagación de las *fake news* o los ciberataques, la realidad es que el Gran Cortafuegos no es más que un perfecto ejemplo de cómo las autoridades pueden servirse del IoE y la IA para manipular la información.

Así, es posible apreciar cómo la IA puede ser empleada para dar forma a discursos políticos, propiciar la desinformación de los ciudadanos, controlar los mensajes transmitidos por las autoridades y, en general, para manipular la opinión pública.

iv. **Impacto social**: por un lado, la revolución tecnológica conlleva una transformación en las relaciones sociales, lo que podría derivar en una pérdida inminente de habilidades personales (Groth, Nitzberg y Esposito, 2018). Asimismo, la apertura de nuevos canales de comunicación implica que el ser humano se encuentra a un solo clic de conectar con sus seres queridos, de modo que la mayor parte de las relaciones interpersonales han sido virtualizadas o trasladadas al mundo digital (Velarde Hermida & Casas-Mas, 2018).

Por su parte, la IA genera lo que habitualmente se conoce como síndrome de exceso de información (*Information Fatigue Syndorme*), un concepto acuñado por el psicólogo David Lewis para explicar cómo el bombardeo de información afecta a la estabilidad emocional e incluso a la salud física del ser humano. Así, esta sobreexposición a la información derivada del IoE puede interferir con nuestras horas de sueño, impedir la concentración o contribuir enormemente al estrés (Thomas, 1998).

3. DILEMAS ÉTICOS DERIVADOS DEL USO DE MÁQUINAS CON ALGORITMOS DE INTELIGENCIA ARTIFICIAL

Junto con los riesgos analizados, el uso generalizado de la IA presenta múltiples implicaciones éticas, que pueden ser clasificadas en tres grandes categorías: rendición de cuentas, imparcialidad y privacidad (Marín, 2019).

3.1. Rendición de cuentas

En un mundo cada vez más gobernado por las interacciones entre el ser humano y los sistemas dotados de IA, resulta fundamental contar con procesos de rendición de cuentas capaces de garantizar que las decisiones adoptadas por las máquinas inteligentes son afines a los valores que consideramos fundamentales (Comisión Europea, 2019). Este reto ético presenta una doble dimensión: la responsabilidad de los sistemas inteligentes y la explicabilidad o trazabilidad de las decisiones adoptadas por los mismos.

3.1.1. Responsabilidad

La automatización de procesos mediante la IA puede derivar en un resultado negativo o perjudicial para el ser humano, ya sea por errores de programación, diagnóstico, diseño o actualización del software o por la existencia de sesgos en el algoritmo. Ello nos lleva a plantearnos cuál debe ser la responsabilidad exigida a las máquinas y dispositivos inteligentes por los daños causados en su proceso de toma de decisiones, un reto que preocupa especialmente en el campo de los vehículos autónomos o de la robótica.

El fallo de estos sistemas inteligentes ha derivado en la muerte de múltiples personas en los últimos años (Marín García, 2019). A tal efecto, en el 2015, un equipo de cirujanos del Centro Médico Rush de Chicago, liderado por Jai Raman, descubrió que, entre los años 2000 y 2013, los robots quirúrgicos habían sido empleados en un total de 10,624 operaciones en EEUU, existiendo algún tipo de funcionamiento defectuoso en el equipo en más de 8,061 (75.9%). Asimismo, este estudio, basado en los datos proporcionados por la Agencia de Alimentos y Medicamentos estadounidense, encontró que los principales efectos adversos de estas máquinas automatizadas podían clasificarse en cuatro categorías: (i) los errores en la retransmisión de vídeo o imágenes durante la operación constituyeron el 7.4% de los perjuicios médicos en estas intervenciones; (ii) el desprendimiento de piezas robóticas que se quedan dentro del cuerpo del paciente constituyeron el 14.7%; (iii) la emisión de chispas por el equipo causó quemaduras a más de 193 pacientes, constituyendo así el 10.5% de los efectos adversos; y (iv) los movimientos involuntarios de las máquinas causaron 52 lesiones graves y 2 muertes, constituyendo así un 10.1% de los perjuicios. El resto de los perjuicios causados por el funcionamiento defectuoso de los robots quirúrgicos eran imputables a causas específicas que no pueden encuadrarse en ninguna de las categorías observadas, como problemas en la fuente de alimentación, rotura de cables, etc. (Alemzadeh, Iyer, Kalbarczyk, Leveson y Raman, 2015).

Asimismo, el 18 de marzo de 2018, Elaine Herzberg se encontraba cruzando con su bicicleta por el medio de una carretera de Arizona cuando fue atropellada por un vehículo autónomo de Uber (Shaw, 2019). Tras llevar a cabo una extensa investigación, la Junta Nacional de Seguridad en el Transporte de EEUU, NTSB, concluyó que el accidente fue consecuencia de un error en la programación del software y no de un fallo en el sistema. A tal efecto, el algoritmo de IA empleado por el vehículo estaba programado para reconocer a las personas que cruzaban por pasos de cebra, si bien era

incapaz de detectar a quienes lo hacían por zonas no autorizadas. Como consecuencia de ello, el vehículo no fue capaz de identificar a Herzberg como un riesgo inminente de colisión. Sin perjuicio de ello, Uber no se enfrentó a cargos legales dado que el informe de la NTSB concluyó que el perjuicio podía haber sido evitado si la persona que se encontraba en el interior del vehículo en el momento del impacto, Rafaela Vásquez, hubiese estado prestando atención a la carretera en vez de atender a su teléfono móvil (BBC, 2020).

Como consecuencia de las preocupaciones derivadas de este reto ético al que se enfrenta la IA, el Parlamento Europeo, en su Resolución de 16 de febrero de 2017 sobre robótica, propone dos alternativas: el enfoque de responsabilidad objetiva o el de gestión de riesgos (Parlamento Europeo, 2017). El primero permite imputar responsabilidad a los sistemas automáticos siempre que se pueda demostrar que el daño se ha producido efectivamente y, esencialmente, que existe un nexo causal entre el perjuicio sufrido por el ser humano y el funcionamiento defectuoso de la máquina inteligente. Por el contrario, el enfoque de gestión de riesgos sostiene que la responsabilidad deberá imputarse en proporción al nivel de autonomía del sistema, es decir, cuanto mayor sea su capacidad de aprendizaje, mayor será la responsabilidad de su programador. En consecuencia, la principal diferencia entre ambos enfoques radica en que el primero se centra únicamente en la actuación negligente, imputando la responsabilidad esencialmente al robot, mientras que el segundo aboga por la necesidad de que la misma recaiga sobre el ser humano, siendo éste capaz de minimizar los riesgos y el impacto del perjuicio.

Por su parte, la responsabilidad de las máquinas que emplean algoritmos de IA ha abierto el debate de la posibilidad de dotar a las mismas de una personalidad jurídica específica, pasando a considerar a los robots como *personas electrónicas* (Parlamento Europeo, 2017). Ello permitiría a los sistemas inteligentes tener una responsabilidad civil similar a la de las sociedades. A favor de esta postura se muestra el jurista Moisés Barrio Andrés, quien concibe la personalidad jurídica de los robots como una alternativa plausible al reto jurídico derivado de su imputabilidad y responsabilidad (Barrio, 2018). Por el contrario, cientos de integrantes de la Comisión Europea, entre los que destacan investigadores de IA, expertos en bioética o líderes políticos procedentes de más de catorce países, se muestran radicalmente en contra de esta posibilidad, fundando su pretensión en múltiples razones éticas y legales. Entre éstas, destaca el hecho de que un

sistema inteligente no puede gozar de derechos inherentes al ser humano, como la dignidad o integridad, por lo que su estatus legal no puede derivarse del modelo de una persona natural. Asimismo, las máquinas automáticas no se encuentran representadas o dirigidas por seres humanos, por lo que su estatus legal tampoco puede derivarse del modelo de una persona jurídica (Fernández y Cortés, 2018).

Teniendo todo ello en consideración, no existe una respuesta unánime sobre la forma de proceder ante los perjuicios que derivan de la toma de decisiones por máquinas que emplean algoritmos de IA, lo que, a mi parecer, constituye uno de los principales retos éticos de la Cuarta Revolución Industrial.

3.1.2. Explicabilidad y trazabilidad

Estrechamente ligados a la responsabilidad de las máquinas inteligentes se encuentran los principios de explicabilidad y trazabilidad. El primero, cuyo fundamento se encuentra en el artículo 22 del Reglamento General de Protección de Datos[7], atiende a la necesidad ética o legal de explicar cómo los dispositivos dotados de IA adoptaron una determinada decisión. Por el contrario, el principio de trazabilidad busca que todas las fases del proceso de toma de decisiones por sistemas inteligentes se encuentren claramente especificadas y documentadas.

En esta línea de pensamiento, la trazabilidad de los algoritmos no es una tarea sencilla dado que muchos de ellos se configuran como "cajas negras", es decir, carecen de capacidad explicativa (López de Mántaras, 2015). Es precisamente esta opacidad en los sistemas inteligentes lo que impide el acceso del ser humano al funcionamiento interno de un algoritmo. A la hora de emplear IA para evaluar a los candidatos a un determinado puesto de trabajo, decidir sobre la concesión de un préstamo hipotecario o realizar un diagnóstico clínico, la transparencia del algoritmo es una característica fundamental exigida por el usuario. Por consiguiente, el dilema ético surge cuando la complejidad del sistema inteligente impide determinar los motivos que llevaron a la adopción de una decisión frente a otra. Tal y como sostiene López de Mántaras, este es el caso del *Deep Learning* o aprendizaje profundo, una de las ramas del Machine Learning cuya complejidad impide la explicabilidad o seguimiento de su proceso de toma de decisiones (Monasterio Astobiza, 2017).

[7] Artículo 22.1 del Reglamento General de Protección de Datos: "*Todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar*". Disponible en: https://www.boe.es/doue/2016/119/L00001-00088.pdf

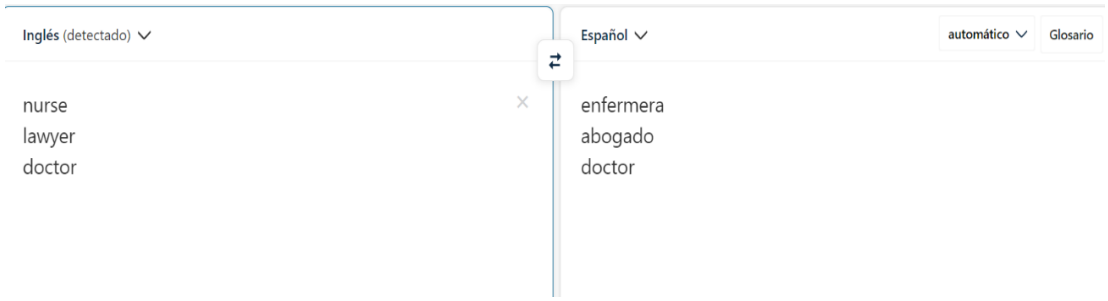
A título de ejemplo, en el año 2007, Adrian Fenty, alcalde de Washington D.C., impulsó a Michelle Rhee, rectora en centros educativos del Estado, a que desarrollara una herramienta para la evaluación del personal docente con el fin de corregir los deficiencias educativas de la ciudad. Para ello, Rhee encargó a la consultora Mathematica la elaboración de un algoritmo, denominado IMPACT, capaz de realizar valoraciones objetivas a nivel individual de los profesores, despojándose de sesgos humanos tales como la amistad u otros elementos subjetivos. Aquellos docentes que no llegaban a un nivel mínimo eran despedidos y ello es lo que le ocurrió a Sarah Wysocki, una maestra muy apreciada por el rector del centro, padres y alumnos, que recibió del algoritmo una baja puntuación en sus competencias lingüísticas y matemáticas. Cuando quiso conocer el porqué de su despido, los programadores no pudieron conceder a Sarah una explicación dado que el modelo era una caja negra (Cortina, 2019).

3.2. Imparcialidad

La discriminación algorítmica se caracteriza por la existencia de sesgos en las máquinas que emplean algoritmos de IA, privando así de objetividad al proceso de toma de decisiones por las mismas. El experto en ética de datos Kevin Macnish sostiene que "*los valores del programador, intencionadamente o no, se congelan en el código, institucionalizando efectivamente dichos valores*" (Mittelstadt, Allo, Taddeo, Watcher y Floridi, 2016, p. 7). Por consiguiente, para Macnish, la discriminación algorítmica tiene su origen en la enorme cantidad de datos empleada para entrenar los modelos de Machine Learning, de modo que la justicia de un algoritmo dependerá de lo justos que sean los datos de entrenamiento. En línea con Macnish, Cynthia Dwork, científica de la Universidad de Harvard, sostiene que, con independencia de lo objetivo y justo que sea un algoritmo, la imparcialidad de los datos determinará el sesgo existente en las decisiones adoptadas por los sistemas inteligentes, ya que "*los algoritmos no tienen acceso a los datos reales sobre el terreno*" (Shaw, 2019). Por su parte, Friedman y Nissenbaum realizan una clasificación de la discriminación algorítmica en atención a su origen, lo que les permite distinguir entre sesgos: (i) preexistentes, que derivan de los valores o influencias sociales y culturales presentes en las instituciones de las que surge la IA; (ii) técnicos, derivados de las limitaciones técnicas en su diseño y desarrollo; o (iii) emergentes, cuando su origen radica en cambios culturales o sociales no detectados por el algoritmo (Mittelstadt, Allo, Taddeo, Watcher y Floridi, 2016).

En cualquier caso, los sistemas dotados de IA habitualmente presentan prejuicios que derivan en conclusiones parciales y discriminatorias. Prueba de ello son los algoritmos de traducción automática, cuyo sesgo por razón de género queda demostrado al traducir palabras o expresiones a idiomas que carecen de género gramatical, como el inglés. Este es el caso del traductor gratuito de DeepL, especialmente sesgado en el ámbito laboral.

Figura 1. Sesgo algorítmico por razón de género

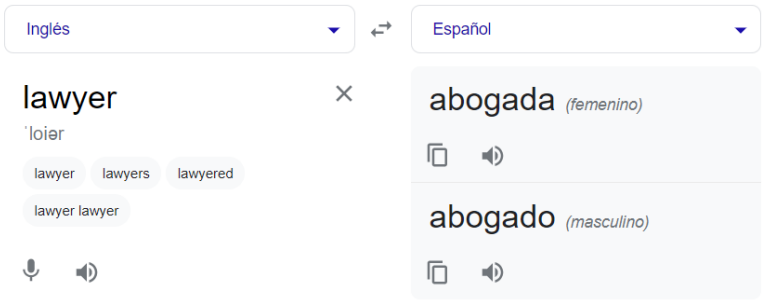


Fuente: Captura de pantalla tomada de DeepL, con fecha 6 de febrero de 2022

Como se puede apreciar en la Figura 1, la palabra "nurse" se traduce como "enfermera", mientras que "lawyer" se traduce como "abogado" y "doctor" como "doctor". Al carecer el inglés de género gramatical, el algoritmo debería haber empleado indistintamente el masculino y femenino, si bien queda demostrado que, por el contrario, asocia la profesión de enfermería con el género femenino y las de abogacía y medicina con el masculino.

A tal efecto, Google Translate adolecía del mismo sesgo, por lo que fue objeto de múltiples críticas hasta que, finalmente, en 2018 adoptó un nuevo paradigma orientado a suprimir las discriminaciones por razón de género en la traducción automática. Por ello, como puede apreciarse en la Figura 2, al introducir términos aislados, Google ofrece ambas posibilidades, indicando que la traducción varía en función del género.

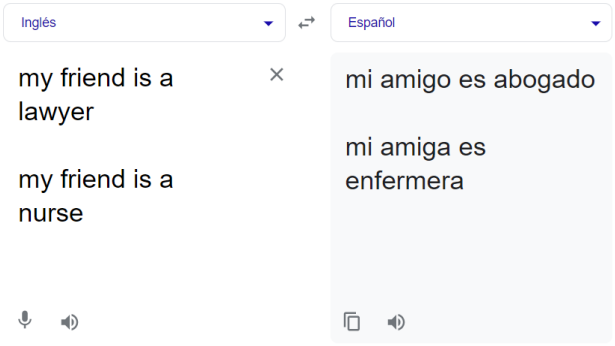
Figura 2. Nuevo paradigma de Google Translate



Fuente: Captura de pantalla tomada de Google Translate, con fecha 6 de febrero de 2022

No obstante, sin perjuicio de los esfuerzos de Google por suprimir el sesgo algorítmico padecido por su sistema de traducción, esta aplicación de IA sigue discriminando por razón de género en el campo laboral, esencialmente a la hora de incluir expresiones no contextualizadas. Ello puede apreciarse en la Figura 3, en la cual el algoritmo traduce "*my friend is a lawyer*" como "mi amigo es abogado", mientras que "*my friend is a nurse*" queda traducido como "mi amiga es enfermera".

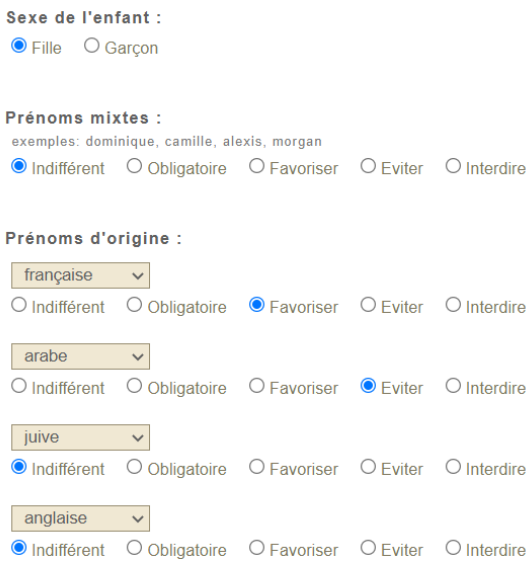
Figura 3. Sesgo por razón de género de Google Translate



Fuente: Captura de pantalla tomada de Google Translate, con fecha 6 de febrero de 2022

Otra muestra de la parcialidad algorítmica viene dada por la página web francesa Ton Prenom (http://tonprenom.com/bebe), que emplea IA para ayudar al usuario a encontrar un nombre para su futuro bebé. Concretamente, la página se encuentra sesgada por razón de etnia, ya que, tal y como se desprende de la Figura 4, el algoritmo evita por defecto la selección de nombres de origen árabe (Monasterio Astobiza, 2017).

Figura 4. Discriminación algorítmica por razón de etnia

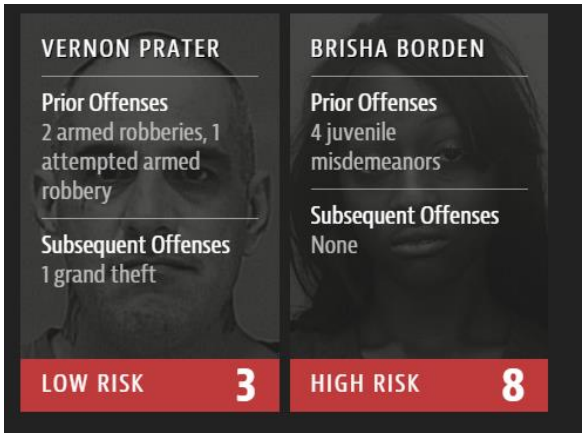


Fuente: Captura de pantalla de http://tonprenom.com/bebe tomada el 6 de febrero de 2022

Asimismo, en 2017 Amazon se vio obligada a abandonar un algoritmo que había comenzado a emplear en el año 2014 para captar desarrolladores de software, al percatarse de que esta herramienta discriminaba por razón de género. El sistema penalizaba aquellos currículums que incluían la palabra "mujer", ya que el conjunto de entrenamiento estaba integrado por los datos de los principales postulantes al empleo durante la primera década de los 2000, siendo hombres la mayoría de ellos (Shaw, 2019). En consecuencia, en línea con el argumento sostenido por Macnish, al estar los datos de entrada sesgados, las decisiones adoptadas por el sistema de IA inevitablemente iban a resultar discriminatorias.

Sin perjuicio de ello, de los múltiples casos que ejemplifican la discriminación algorítmica, quizás el más grave sea el relativo al proceso automático empleado en múltiples Estados de Norteamérica - como Virginia, Arizona, Washington o Colorado, entre otros – para la determinación del riesgo de reincidencia por los presos estadounidenses (Monasterio, 2017). Un estudio llevado a cabo en 2016 por ProPublica demostró que este algoritmo, desarrollado por Northpointe Inc., se encontraba fuertemente sesgado por razón del color de piel del preso. Para ello, la agencia analizó los factores de riesgo asignados por el sistema inteligente a más de 7,000 personas arrestadas entre 2013 y 2014 en un condado de Florida, para posteriormente comprobar cuántas de estas predicciones se cumplieron en los dos años siguientes. El estudio de ProPublica proporcionó dos conclusiones fundamentales: (i) el algoritmo asignaba a los acusados blancos un menor riesgo de reincidencia que a los negros; y (ii) el sistema etiquetaba erróneamente como futuros delincuentes a los detenidos de raza negra en una proporción casi dos veces mayor que a los de raza blanca (ProPublica, 2020).

Figura 5. Discriminación racial en el algoritmo que mide el riesgo de reincidencia criminal



Fuente: captura de pantalla tomada de ProPublica el 7 de febrero de 2022

La Figura 5 muestra como el algoritmo de Northpointe Inc. asignó un bajo riesgo de reincidencia a Vernon Prater, varón blanco que había cometido dos robos a mano armada y un intento de robo con arma de fuego, frente al alto riesgo de comisión de nuevos delitos por Brisha Borden, mujer negra que había sido condenada por cuatro faltas de pintar las paredes de una vía pública (ProPublica, 2020). Al comparar las ofensas cometidas por ambos acusados, resulta evidente que las de Prater son objetivamente más graves que las de Borden, si bien el algoritmo asignó un riesgo de reincidencia dos veces mayor a la mujer de raza negra, lo que denota la evidente falta de imparcialidad del mismo.

3.3. Privacidad

La mayor parte de las máquinas que emplean algoritmos de IA recopilan ingentes cantidades de datos para facilitar el proceso automático de toma de decisiones, lo que ha derivado en una creciente preocupación por el uso y gestión que pueda hacerse de esta información. La Oficina del Comisionado de Información (ICO) del Reino Unido sostiene que los datos recopilados por el Big Data pueden ser de cuatro tipos: (i) proporcionados, aquellos que el usuario facilita voluntariamente; (ii) observados, aquellos que son registrados automáticamente, como los biométricos; (iii) derivados, que se obtienen de información conocida; e (iv) inferidos, aquellos que derivan de procesos analíticos complejos orientados a encontrar correlaciones entre bases de datos (ICO, 2017). El problema surge cuando toda esta información es empleada de forma malévola por empresas, partidos políticos, servicios de inteligencia, hackers u organizaciones, difuminando así los límites entre vigilancia legítima e ilegítima y erosionando la privacidad de los usuarios (Monasterio Astobiza, 2017).

A efectos de privacidad, interesan especialmente los datos observados por los sistemas inteligentes. Por un lado, IoE supone que los dispositivos dotados de IA empleados en nuestra vida cotidiana, como Siri, Google Now o Amazon Echo, recogen y procesan miles de conversaciones que tienen lugar a diario en nuestros hogares, incluso cuando el dispositivo se encuentra apagado o inactivo (Marín, 2019). Por su parte, preocupa especialmente el reconocimiento facial, una herramienta biométrica cada vez más empleada en la vía pública para preservar la seguridad de los ciudadanos. Este es el caso del Reino Unido y varios Estados de EEUU, como Tampa, Oakland y San Francisco, donde esta aplicación de la IA se emplea en lugares públicos para la identificación de criminales o personas buscadas por los servicios policiales. No obstante, hay quienes

consideran que las implicaciones éticas derivadas del uso de esta tecnología superan los beneficios aportados por la misma, motivo por el cual la ciudad de San Francisco prohibió el reconocimiento facial en la vía pública (Domingo Jaramillo, 2021).

Sin perjuicio de ello, los datos privados proporcionados voluntariamente por los usuarios también pueden ser empleados por empresas y organizaciones de forma ilegítima, lo que nos hace plantearnos cuestiones como su capacidad de manipulación, la posible creación de hábitos dañinos para la sociedad o la erosión de las instituciones democráticas (Marín, 2019). Un claro ejemplo de ello viene dado por el escándalo de Facebook y Cambridge Analytica, que motivó la introducción de medidas tendentes a garantizar la privacidad y seguridad de los usuarios de la plataforma. Hasta el 2014, Facebook permitía el libre acceso a los datos de sus clientes siempre y cuando se pudiese demostrar la existencia de fines de investigación, es decir, siempre que no mediase una motivación comercial. No obstante, esta situación cambió cuando en marzo de 2018 salió a la luz que Cambridge Analytica, una empresa londinense centrada en el análisis de datos, había comprado al gigante tecnológico la información de unos cincuenta millones de americanos sin su consentimiento, con el fin de crear una herramienta de IA capaz de inferir perfiles psicológicos de los usuarios. Para ello, Aleksandr Kogan, el encargado de recopilar la información, desarrolló una prueba de personalidad por la que miles de usuarios de Facebook proporcionaron voluntariamente sus datos para completar una encuesta remunerada de uso académico. Sin embargo, Mark Zuckerberg, director ejecutivo de la plataforma, permitió que el sistema inteligente accediese también a la información de toda la red de contactos del usuario que había rellenado la encuesta. Posteriormente, esta información fue vendida a Cambridge Analytica y empleada para manipular el comportamiento de los votantes en las elecciones estadounidenses de 2016, que resultaron en el nombramiento de Donald Trump como presidente de EEUU (BBC News Mundo, 2018). Como consecuencia de este escándalo, Mark Zuckerberg reforzó la ciberseguridad y política de privacidad de los usuarios, si bien se demostró el grado de exposición al que se encuentra sometida nuestra información privada.

4. LA ÉTICA EN EL DESARROLLO DE UNA INTELIGENCIA ARTIFICIAL

Los retos derivados del uso de máquinas que emplean algoritmos de IA han puesto de manifiesto la necesidad de definir unas prácticas éticas en el mundo digital. A tal efecto, el primer sistema ético-normativo introducido en esta materia fueron las tres leyes de la

robótica propuestas por el escritor de ciencia ficción Isaac Asimov, autor del libro *Yo, Robot*. Éstas fueron publicadas por primera vez en 1942 en un ensayo titulado *Runaround* y se configuraron como formulaciones matemáticas o leyes éticas a las que cualquier robot debe someterse. Estas tres leyes de la robótica son:

"*1.- Un robot no hará daño a un ser humano o, por su inacción, permitirá que un ser humano sufra daño.*

2.- Un robot debe obedecer las órdenes que le son dadas por los seres humanos, excepto si éstas entrasen en conflicto con la primera Ley.

3.- Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la primera o la segunda Ley" (Monasterio Astobiza, 2017).

Si bien estas normas tan solo deben ser consideradas en abstracto, han servido como base en el desarrollo de múltiples prácticas y principios éticos que buscan garantizar una IA beneficiosa para el ser humano, como es el caso de los 23 Principios de Asilomar, que serán analizados a continuación (Cortina, 2019).

En este contexto, los grandes organismos internacionales abogan por la necesidad de garantizar una IA fiable, de calidad, segura y al servicio del ser humano, para lo que se requiere la aprobación de normas y pautas orientadas a la consecución de este objetivo. A tal efecto, Thierry Breton, profesor de Harvard Business School y comisario del Mercado Interior de la Unión Europea, UE en adelante, sostiene que "*La IA es un medio, no un fin*". En esta misma línea, Margrethe Vestager, responsable de la cartera "Una Europa Adaptada a la Era Digital", defiende que "*En cuanto a la inteligencia artificial, la confianza es una obligación, no un adorno*" (Comisión Europea, 2021). Consecuentemente, desde la publicación de las tres leyes de la robótica de Asimov, múltiples instituciones internacionales y organizaciones tecnológicas han firmado acuerdos y elaborado directrices para el impulso de la excelencia en materia de IA.

4.1. Definición de prácticas éticas por organismos nacionales e internacionales

4.1.1. Recomendaciones a nivel internacional

a. Directrices éticas sobre una Inteligencia confiable de la Comisión Europea

En abril de 2019, la Comisión Europea publicó un informe con las Directrices éticas sobre una Inteligencia confiable, elaborado por el Grupo de expertos de alto nivel sobre IA, un grupo independiente constituido por la Comisión en el año 2018.

El informe tiene por objeto la promoción de una IA fiable, lo que implica que ésta sea lícita, ética y robusta. Por consiguiente, la fiabilidad de esta tecnología se desprende de su respeto por las leyes vigentes, los derechos fundamentales y los valores y principios esenciales, así como de su robustez técnica y social, que pasa por el dominio tecnológico de sus programadores para evitar perjuicios involuntarios. Por su parte, la Comisión sostiene que los profesionales de la IA deben respetar una serie de imperativos éticos en el desarrollo de máquinas inteligentes, tales como la autonomía humana, prevención del daño, equidad, justicia, explicabilidad y transparencia. Más aún, el informe recoge siete requisitos esenciales que debe reunir cualquier sistema inteligente y son: (i) agencia humana y supervisión; (ii) robustez y seguridad; (iii) privacidad y control del tratamiento de los datos; (iv) transparencia; (v) diversidad, equidad y no discriminación; (vi) bienestar social y medioambiental; y (vii) responsabilidad y rendición de cuentas (Comisión Europea, 2018).

En este sentido, el concepto de IA manejado en el presente trabajo de investigación se caracteriza por la supervisión e intervención humana en el proceso de toma de decisiones por las máquinas inteligentes, si bien, como se ha podido apreciar en el apartado anterior, ello no impide que éstas puedan causar un perjuicio social, cultural e incluso político. Consecuentemente, los organismos internacionales inciden en la importancia de que los programadores respeten el principio de proporcionalidad entre fines y medios, permitiendo a los usuarios oponerse a las decisiones adoptadas por los sistemas inteligentes. Según el grupo de expertos de la Comisión, ello es lo que permitirá la explicabilidad, trazabilidad y rendición de cuentas en caso de perjuicios causados por la intervención de máquinas dotadas de IA (Comisión Europea, 2018).

Sin perjuicio de todo ello, mediante un comunicado de prensa de 21 de abril de 2021, la Comisión Europea anunció la incorporación de nuevas normas y medidas orientadas a lograr la excelencia y fiabilidad de los sistemas inteligentes. Esta iniciativa, coordinada por la cartera de “Una Europa Adaptada a la Era Digital”, distingue entre aplicaciones de IA de riesgo inadmisible, alto riesgo, riesgo limitado y riesgo mínimo o nulo (Comisión Europea, 2021), con el fin de determinar las medidas necesarias para garantizar la seguridad de cada una de éstas. La Tabla 1 contiene un resumen de cada una de estas aplicaciones de los sistemas inteligentes.

Tabla 1. Clasificación de los sistemas inteligentes según el riesgo

Riesgo inaceptable	• Sistemas de clasificación social al servicio de los gobiernos • Juguetes que emplean asistencia vocal para la manipulación de los menores
Riesgo alto	• Sistemas inteligentes de aplicación quirúrgica • Sistemas de identificación biométrica en remoto • Software de clasificación de currículums • Sistemas de calificación crediticia • Vehículos autónomos
Riesgo reducido	• Robots conversacionales
Riesgo mínimo	• Videojuegos basados en IA • Sistemas de filtrado de spam o correo no deseado

Fuente: elaboración propia a partir de la Comisión Europea (2021)

El objeto de este nuevo paquete de medidas es prohibir los sistemas de IA de riesgo inadmisible por atentar contra la seguridad, los derechos y los medios de subsistencia de los seres humanos. Asimismo, se pretende garantizar que los sistemas de alto riesgo, con carácter previo a su comercialización, quedan sujetos a estrictos procesos de evaluación, trazabilidad de los resultados y mitigación de riesgos, así como a controles de seguridad y de calidad de la información (Comisión Europea, 2021).

b. Recomendación del Consejo de la OCDE sobre Inteligencia Artificial

En mayo de 2009, los treinta y cuatro integrantes de la Organización para la Cooperación y el Desarrollo Económicos, OCDE en adelante, y otros países no miembros - Argentina, Brasil, Colombia, Costa Rica, Perú y Rumanía - aprobaron una serie de recomendaciones sobre los principios éticos que deben regir los sistemas inteligentes, así como los mecanismos orientados a lograr la cooperación internacional en el desarrollo de una IA confiable.

En línea con las Directrices éticas de la Comisión Europea, el Consejo de la OCDE sostiene que los sistemas inteligentes deberán respetar simultáneamente los siguientes principios: (i) crecimiento inclusivo, bienestar y desarrollo sostenible; (ii) equidad y respeto por los valores humanos; (iii) explicabilidad y transparencia; (iv) seguridad, protección y robustez; y (v) responsabilidad. Más aún, la Recomendación defiende la

necesidad de desarrollar un marco normativo internacional que garantice la ética de los sistemas inteligentes, por lo que propone fomentar la inversión en IA, la creación de ecosistemas digitales accesibles para todo ser humano y la preparación del mercado ante las inminentes trasformaciones laborales que derivan de los avances en este campo (OCDE, 2019).

c. Libro Blanco de la IA

El Libro Blanco de la UE, aprobado por la Comisión Europea en febrero de 2020, recoge diversas alternativas políticas para facilitar el desarrollo de ecosistemas de excelencia y de confianza en el marco de la IA.

Por un lado, para lograr un ecosistema de confianza, el Libro Blanco apuesta por el desarrollo de un nuevo marco normativo antropocéntrico basado en las Directrices éticas para una IA fiable, prestando especial atención a los riesgos de opacidad en la toma de decisiones, discriminación por razón de género, raza u otro tipo, intromisión en la vida privada del usuario o uso de sistemas inteligentes con fines delictivos (Comisión Europea, 2020).

Por su parte, la Comisión sostiene que para garantizar un ecosistema de excelencia se requiere una colaboración activa entre los Estados miembros de la UE, la creación de centros de innovación e investigación, el desarrollo de programas educativos en IA a fin de atraer académicos y científicos, el apoyo a empresas emergentes en el acceso a financiación para la digitalización de sus negocios y la implantación de sistemas inteligentes en sectores públicos prioritarios (Comisión Europea, 2020).

d. Recomendación sobre la Ética de la IA de la UNESCO

El 7 de septiembre de 2020 se publicó el primer borrador de recomendación sobre la Ética de la IA por la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, UNESCO en adelante. Este documento, firmado por los 193 países que integran la organización internacional, presenta una importancia fundamental por considerarse el primer acuerdo mundial sobre la ética de los sistemas que emplean algoritmos de IA.

La recomendación se basa en las directrices éticas ya apuntadas por la OCDE y por el Grupo de expertos de alto nivel sobre IA de la Comisión Europea, reiterando así la importancia de los principios de proporcionalidad, seguridad, equidad, sostenibilidad,

privacidad, supervisión humana, explicabilidad, responsabilidad, trazabilidad y gobernanza de los sistemas inteligentes (UNESCO, 2020).

4.1.2. Recomendaciones a nivel nacional

En marzo de 2017 se aprobó la Declaración de Barcelona sobre el desarrollo y uso apropiado de la IA, cuyo objetivo principal era garantizar que las aplicaciones de esta tecnología eran "*prudentes, confiables, responsables, identificables y con autonomía regida por un conjunto de reglas que velen por el acervo humano*" (Ministerio de Ciencia, Innovación y Universidades, 2019, p. 41). Ello motivó la elaboración de la Estrategia española de I+D+I en Inteligencia Artificial, un proyecto desarrollado en 2019 por el Ministerio de Ciencia, Innovación y Universidades con el objeto de garantizar la implantación de sistemas inteligentes en sectores esenciales de la sociedad y economía españolas.

En aras de alcanzar este objetivo, el Gobierno marcó una serie de prioridades entre las que destaca aquella orientada a la redacción de un Código Ético de la IA que asegure su uso justo, transparente y responsable. Asimismo, la Estrategia busca eliminar de los algoritmos los sesgos no intencionados y demás perjuicios discriminatorios de los que adolece nuestra sociedad.

El desarrollo de esta estrategia derivó en la aprobación de la Orden ETD/670/2020, de 8 de julio, por la que se crea y regula el Consejo Asesor de Inteligencia Artificial, un órgano consultivo de apoyo al Gobierno en materia de IA. Los miembros del Consejo, representantes de la comunidad científica con experiencia en el ámbito de la ética algorítmica, prestan asesoramiento al Gobierno para garantizar un uso responsable y eficiente de los sistemas inteligentes. Por consiguiente, este órgano vela por el desarrollo de una IA que contribuya a la recuperación económica, el crecimiento sostenible, la lucha contra la discriminación y el uso y acceso equitativos a esta tecnología.

4.2. Definición de prácticas éticas a nivel empresarial

Los gigantes tecnológicos de Silicon Valley son conscientes de la necesidad de garantizar un uso transparente, justo y ético de los sistemas inteligentes, dado que ésta es la única forma de controlar quién gobierna y distribuye la información captada por las máquinas que emplean algoritmos de IA.

Como consecuencia de ello, el 28 de septiembre de 2016, Google, Amazon, Microsoft, IBM, Facebook y Deep Mind anunciaron la creación de la Alianza en IA (traducido del inglés *Partnership on AI*), una organización sin ánimo de lucro que trata de dar respuesta a las preocupaciones más complejas e importantes que rodean el futuro de los sistemas inteligentes. Lejos de presionar a los órganos de gobierno para la adopción de un sistema normativo en materia de IA, la entidad se centra en investigar las aplicaciones beneficiosas de esta tecnología, así como en combatir los usos perjudiciales de la misma. Dada la importancia de esta iniciativa, Apple se unió al consorcio en enero de 2017 (Monasterio Astobiza, 2017).

Por su parte, en el año 2015, Elon Musk fundó *Open AI*, una organización sin ánimo de lucro encargada de velar por el desarrollo de una IA amigable que beneficie al conjunto de la humanidad. En reiteradas ocasiones, Musk ha puesto de manifiesto la necesidad de actuar cautelosamente en lo que al desarrollo de sistemas inteligentes respecta, llegando incluso a comparar la aceptación de los mismos con "invocar al demonio" y a defender que el uso de la IA es mucho más peligroso que el de las armas nucleares (Lincoln, 2021). Como consecuencia de estas preocupaciones, Musk aboga por la configuración responsable de máquinas inteligentes, garantizando la supervisión humana en todas las fases del proceso de programación. En este contexto, *Open AI* busca desarrollar una IA segura, beneficiosa y capaz de aprender de la retroalimentación humana (Amodei, 2020).

En esta línea de pensamiento, el Future of Life Institute, en su segunda conferencia sobre IA beneficiosa (del inglés *Beneficial AI*) celebrada en enero de 2017, aprobó los Principios de Asilomar, un proyecto ratificado por un total de 847 investigadores en IA y robótica y apoyado por múltiples científicos y autores relevantes en el mundo de la ética, la filosofía, el derecho o la economía (Monasterio Astobiza, 2017). Los veintitrés principios, orientados al desarrollo de una IA segura, se clasifican en tres grandes categorías: temas de investigación, ética y valores, y precauciones a largo plazo. El primer bloque de principios determina el objeto de la investigación, la importancia de la financiación en materia de IA, la necesidad de un vínculo constructivo entre ciencia y política, la cultura que debe fomentarse entre los expertos en sistemas inteligentes y la cooperación activa entre las organizaciones desarrolladoras de esta tecnología. Por su parte, el segundo bloque está integrado por los principios que determinan el componente ético exigido en la toma de decisiones por máquinas que emplean algoritmos de IA: (i)

seguridad; (ii) transparencia en el funcionamiento; (iii) transparencia judicial; (iv) responsabilidad; (v) alineación de valores; (vi) valores humanos; (vii) privacidad personal; (viii) libertad; (ix) beneficio compartido; (x) prosperidad compartida; (xi) control humano; (xii) no subversión; y (xiii) evasión de una carrera armamentística en IA. Finalmente, el último bloque afronta las principales preocupaciones a largo plazo en materia de sistemas de IA, poniendo de manifiesto la importancia de que la superinteligencia se desarrolle al servicio del bien común de la humanidad (Future of Life Institute, 2021).

Junto a estos principios y alianzas entre organismos tecnológicos, empresas e instituciones, existen otras iniciativas como el Comité Internacional para el Control de Armas Robóticas, ICRAC en adelante, o la fundación *Responsible Robotics*. Por un lado, el ICRAC se configura como una organización sin ánimo de lucro integrada por científicos, abogados y expertos en ética algorítmica, que vela por la prohibición del uso de IA en el ámbito militar. Para el ICRAC, los sistemas armamentísticos autónomos constituyen una amenaza para la paz, el estado de derecho, los civiles y la seguridad internacional, de modo que las máquinas no pueden tener la capacidad de decidir sobre la vida de un ser humano (ICRAC, s.f.). Por su parte, la fundación sin ánimo de lucro *Responsible Robotics* mira por la rendición de cuentas de la innovación humana en materia de robótica, proporcionando una etiqueta de calidad a todas aquellas máquinas inteligentes que han superado un control previo de seguridad, privacidad, protección, transparencia, sostenibilidad, responsabilidad y equidad (*Foundation for Responsible Robotics*, 2020).

5. EL DILEMA SOCIAL DE LOS VEHÍCULOS AUTÓNOMOS

Los vehículos autónomos constituyen una de las aplicaciones de la IA que mayor controversia ha causado a nivel social. De acuerdo con la Sociedad de Ingenieros de la Automoción, SAE en adelante, existen seis niveles de conducción autónoma, siendo 0 el nivel de nula intervención de sistemas inteligentes y 5 la plena automatización. A tal efecto, hay compañías que ya han desarrollado y se encuentran testando vehículos de nivel 4 para fines logísticos o comerciales, como es el caso de Daimler, en colaboración con Google, o General Motors (Marín, 2019). La Tabla 2 recoge una caracterización general de cada uno de estos niveles de automatización (SAE, 2021).

Tabla 2. Características de los niveles de automatización de vehículos según la SAE

Nivel 0	• Nula automatización de la conducción • Las tareas de conducción dinámica son realizadas por el ser humano, si bien se incorporan funciones básicas de apoyo al conductor, como el sistema de frenado automático
Nivel 1	• Asistencia al conductor • Incorporación de funciones de apoyo al conductor, tales como movimiento longitudinal o control crucero adaptativo
Nivel 2	• Automatización parcial de la conducción • Incorporación de funciones de apoyo al conductor, tales como movimiento longitudinal y control crucero adaptativo, simultáneamente
Nivel 3	• Automatización condicionada de la conducción • Incorporación de funciones de automatización condicionada, tales como sistemas de conducción autónoma en atascos. El conductor debe ir preparado para reaccionar adecuadamente e intervenir ante fallos del sistema
Nivel 4	• Automatización elevada de la conducción • El vehículo controla las condiciones del entorno así como la situación de tráfico y cuenta con un sistema de respaldo en caso de fallo del sistema. No obstante, las funciones de automatización siguen estando condicionadas y, por ello, ante determinadas situaciones, el vehículo puede no seguir conduciendo de forma autónoma
Nivel 5	• Automatización plena de la conducción • Las funciones de automatización no quedan sujetas al cumplimiento de ciertas condiciones, de modo que el vehículo podrá conducir de forma autónoma bajo cualquier circunstancia

Fuente: elaboración propia a partir de SAE (2021)

Desde el éxito en el año 2007 de la arquitectura de *Junior*, el primer coche robótico capaz de recorrer entornos urbanos de forma autónoma, los gigantes tecnológicos se han lanzado al desarrollo de vehículos que progresivamente presentan un mayor nivel de automatización, como es el caso de Tesla o Waymo, el coche autónomo de Google. Ello

se debe a que esta aplicación de la IA presenta múltiples beneficios, tales como la disminución de la contaminación, el incremento de la eficiencia en el tráfico y, esencialmente, la drástica reducción en el número de accidentes en carretera, que quedarían eliminados en un 90% (Bonnefon, Shariff & Rahwan, 2016).

Sin perjuicio de ello, el uso de vehículos altamente automatizados implica supeditar la seguridad del ser humano al funcionamiento del sistema de IA del automóvil (Marín, 2019), lo que requerirá la toma de decisiones éticas complejas. A título de ejemplo, en caso de colisión inminente, el vehículo deberá optar entre sacrificar la vida de los pasajeros en aras de salvar a los peatones o viceversa. Ello es lo que habitualmente se conoce como distribución del daño y guarda una estrecha relación con el *dilema del tranvía*, un famoso experimento mental ideado por Philippa Foot que trata de entender cómo los seres humanos reaccionan ante una situación moralmente compleja. El problema planteado es el siguiente: un tranvía que avanza sin frenos va a atropellar a cinco personas que se encuentran sobre el carril, si bien tú tienes el poder de tirar de una palanca para desviar su trayectoria, lo que supondría sacrificar la vida de un peatón que se encuentra al otro lado de las vías. En diez segundos tendrás que decidir si tiras de la palanca, lo que implicaría matar a una persona, o si, por el contrario, no haces nada, lo que les costaría la vida a cinco. La mayoría coincide en que lo correcto sería tirar de la palanca dado que ésta es la opción que causa el menor daño posible; ello es lo que se conoce como doctrina moral utilitaria, en cuya virtud la mejor vía de actuación moral es aquella que permite minimizar el perjuicio causado. Llevando este experimento al caso que aquí concierne, en un supuesto de colisión inminente, programar un vehículo autónomo para que se guíe por la doctrina moral utilitaria podría suponer el sacrificio de los pasajeros si ello permite salvar la vida de un mayor número de peatones. Consecuentemente, este planteamiento podría desalentar a compradores potenciales que creen firmemente que su seguridad debe estar por encima de otras consideraciones (Bonnefon, Shariff & Rahwan, 2016).

Si bien se trata de una situación hipotética que podría no llegar a ocurrir, a la hora de diseñar el sistema informático del vehículo autónomo, los programadores deberán incorporar reglas de decisión que determinen qué hacer ante tales circunstancias. Por consiguiente, los fabricantes de estos automóviles deben tratar de cumplir tres objetivos potencialmente incompatibles: no desalentar al comprador, coherencia en la programación del sistema y evitar la indignación social. Ello conlleva determinar cuál

es el algoritmo moral que estamos dispuestos a aceptar como ciudadanos y al que estamos dispuestos a someternos como conductores (Bonnefon, Shariff & Rahwan, 2016).

Teniendo todo ello en consideración, a fin de determinar el componente ético exigido en la programación de los vehículos autónomos, he llevado a cabo un análisis regresivo en el lenguaje de programación de R basado en la información proporcionada por seis encuestas realizadas por Jean-François Bonnefon, de la Universidad de Toulouse, Azim Shariff, de la Universidad de Oregon, e Iyad Rahwan, profesor del Instituto de Tecnología de Massachusetts (Bonnefon, Shariff & Rahwan, 2016). El script se adjunta al presente trabajo como Anexo I. Los participantes, todos ellos ciudadanos estadounidenses de entre 18 y 72 años, fueron sometidos a diversos escenarios relativos a la distribución del daño en la programación de vehículos autónomos. La información proporcionada por los mismos, distribuida en seis bases de datos, me ha permitido llevar a cabo tres estudios, correspondientes con las encuestas 1, 3 y 6, respectivamente:

a. Estudio 1

Para la realización del primer estudio, se recogió una muestra de 92 participantes (n = 92) y se les planteó el siguiente escenario: eres el único pasajero de un vehículo autónomo que va circulando por carretera. De repente, 10 peatones aparecen en la trayectoria del mismo, por lo que tienes dos opciones: (i) mantener la ruta, lo que implica atropellar a los peatones, si bien tú saldrías ileso; o (ii) desviarte de la carretera, en cuyo caso fallecerías en el acto, pero habrías salvado la vida de 10 personas.

Teniendo todo ello en consideración, se preguntó a los participantes si el curso moral de acción era desviarse o mantener la trayectoria, así como si esperaban que los vehículos autónomos del futuro fuesen programados para obedecer a la doctrina moral utilitaria.

Este estudio demostró que el 79.35% de los participantes pensaba que el curso moral de acción del algoritmo era desviarse, es decir, sacrificar al pasajero para salvar la vida de los 10 peatones. Más aún, puede apreciarse una preferencia generalizada hacia los vehículos autónomos programados para minimizar el número de bajas, con una mediana de 88 en una escala de 0 (‘proteger al pasajero a toda costa’) a 100 (‘minimizar el daño causado’). Sin embargo, los participantes pensaron que, en la práctica, los fabricantes no iban a ser tan propensos a introducir un modelo moral utilitario en la programación del algoritmo, ya que la mediana se redujo de 88 a 63.50. Ello puede deberse al hecho de

que la programación del vehículo para minimizar los daños en vez de proteger a su pasajero presenta un impacto claramente negativo en las expectativas de venta. Estas conclusiones pueden apreciarse en las Figuras 6 y 7.

Figura 6. Preferencia por la doctrina moral utilitaria

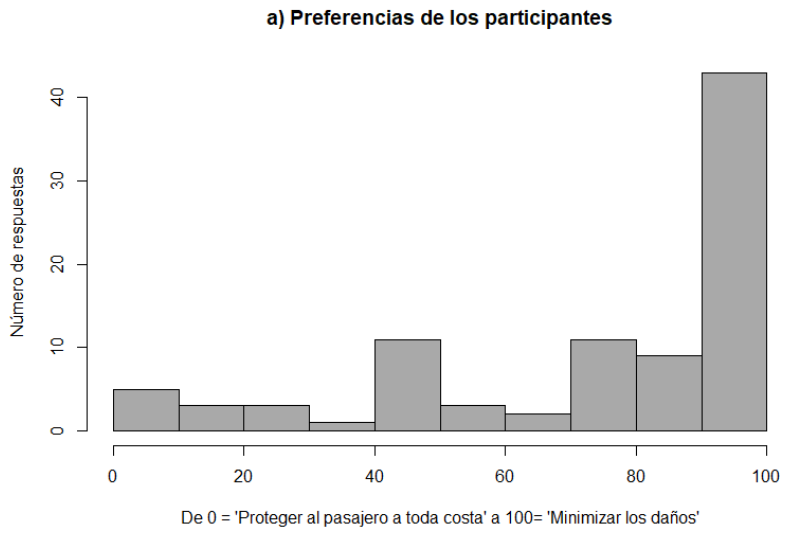
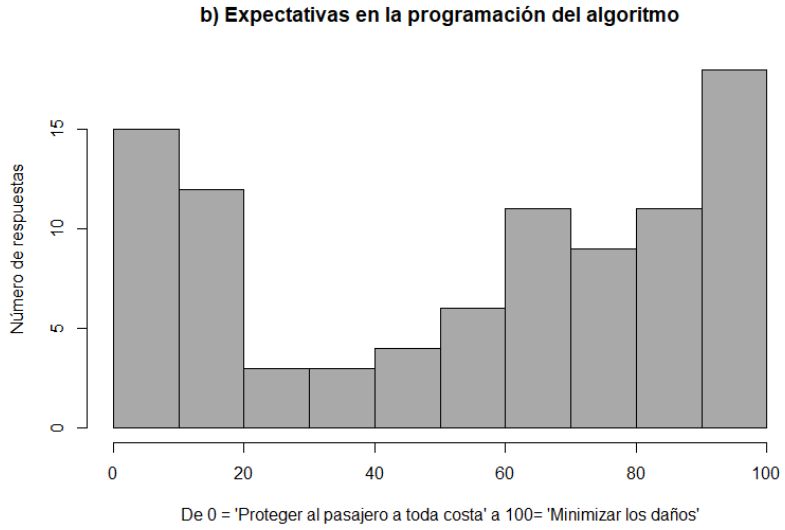


Figura 7. Expectativas en la programación de vehículos autónomos



Fuente: elaboración propia a partir de Bonnefon, Shariff & Rahwan (2016)

Por su parte, cabe destacar que los participantes no se mostraron propensos a comprar un vehículo autónomo, con una mediana de 2.78 en una escala de 0 (‘nunca lo compraría’) a 7 (‘con total seguridad lo compraría’). A tal efecto, a partir de los resultados obtenidos se puede concluir, con un nivel de confianza del 95%, que la

voluntad promedio de los participantes de comprar un vehículo autónomo se encuentra entre 2.34 y 3.23.

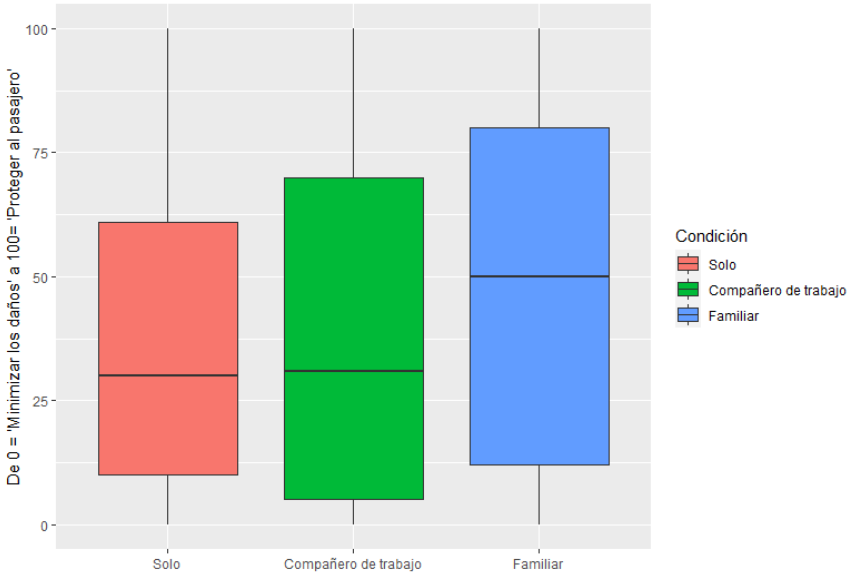
Para concluir, el conjunto de datos analizado reveló que el sexo del participante no era una variable significativa a la hora de determinar la probabilidad de compra de un vehículo autónomo. De hecho, tan solo nueve de los encuestados (n = 9) votaron la probabilidad máxima de compra, siendo 5 de ellos hombres frente a 4 mujeres. Por el contrario, la edad sí que resultó ser una variable altamente significativa, de modo que la mediana en la máxima probabilidad de compra se situaba en 31, con un mínimo de 21 y un máximo de 48 años. Como consecuencia de ello, se puede concluir que los participantes más jóvenes se mostraron más propensos a adquirir un vehículo autónomo.

b. Estudio 2

El segundo estudio parte de las mismas condiciones que el primero, si bien en este caso se pide a los participantes (n = 259) que imaginen tres escenarios posibles: (i) que van solos en el vehículo; (ii) que van acompañados de un familiar; y (iii) que van con un compañero de trabajo. La presencia de un miembro de la familia en el vehículo presenta un impacto moral claramente negativo a la hora de optar por la doctrina moral utilitaria, especialmente en comparación con imaginarse a uno mismo solo en el vehículo (p = 0.003). No obstante, incluso en la situación más adversa (ir acompañado de un familiar), los participantes optaron por minimizar el perjuicio causado, con un intervalo de confianza del 95% entre 53.7 y 65.79.

Sin perjuicio de ello, el estudio comienza a mostrar el dilema moral que plantea esta aplicación de la IA. En una escala de 0 a 100, se preguntó a los participantes sobre la probabilidad de comprar un vehículo autónomo programado para minimizar el daño (0 en la escala), frente a otro programado para proteger a sus pasajeros a toda costa (100 en la escala). Si bien el estudio demostró que la probabilidad de comprar un vehículo autónomo era baja incluso estando el algoritmo programado para proteger a sus pasajeros, ésta era mucho menor cuando los participantes imaginaban que iban acompañados de un familiar. La Figura 8 ilustra cómo los participantes claramente optan por la doctrina moral utilitaria cuando van solos o con un compañero de trabajo, mientras que, en compañía de un familiar, comienzan a abogar por vehículos programados para proteger al pasajero a toda costa (con una mediana superior a 50 en dicha escala).

Figura 8. Preferencia por vehículos autónomos programados para proteger a los pasajeros



Fuente: elaboración propia a partir de Bonnefon, Shariff & Rahwan (2016)

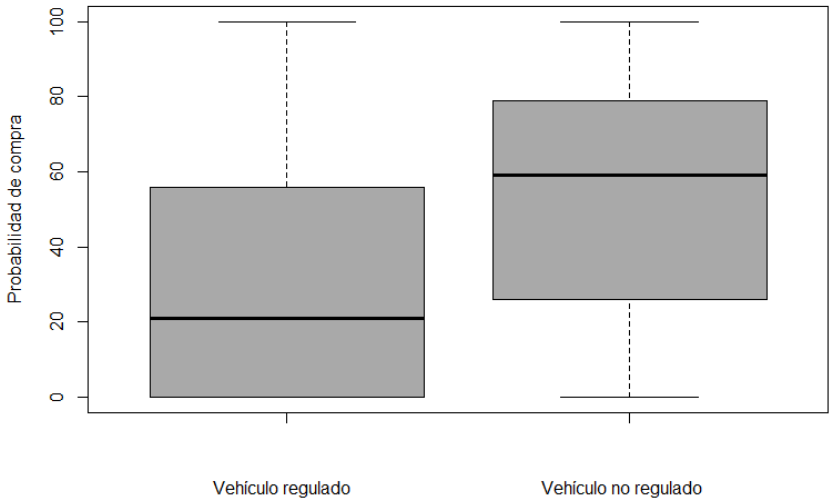
Teniendo todo ello en consideración, el estudio nos permite concluir que, si bien los participantes siguen percibiendo la doctrina moral utilitaria como la opción éticamente correcta, lo cierto es que, para sí mismos y los miembros de su familia, comienza a haber una cierta preferencia por vehículos programados para proteger a sus pasajeros.

c. Estudio 3

Finalmente, partiendo de las condiciones del primer estudio, se plantean tres escenarios alternativos: (i) ir solos en el vehículo; (ii) ir con un familiar cualquiera; o (iii) ir acompañados de un hijo. En este contexto, en una escala de 0 (‘algoritmos no regulados por el gobierno’) a 100 (‘regulación plena’), se pregunta a los participantes (n = 393) sobre la probabilidad de comprar un vehículo autónomo cuyo algoritmo haya sido previamente regulado por el gobierno para seguir la doctrina moral utilitaria. Los resultados demuestran que incluso en el caso más favorable, ir solos en el vehículo, los participantes no se muestran a favor de una regulación gubernamental, con un intervalo de confianza del 95% entre 36.39 y 47.66.

Por su parte, en una escala de 0 (‘probabilidad mínima de compra’) a 100 (‘máxima probabilidad de compra’), la mediana que representa la probabilidad de comprar un vehículo autónomo no regulado por el gobierno es de 59, mientras que la de comprar uno cuyo algoritmo ha sido programado en base a la normativa gubernamental es de tan solo 21 (Figura 9).

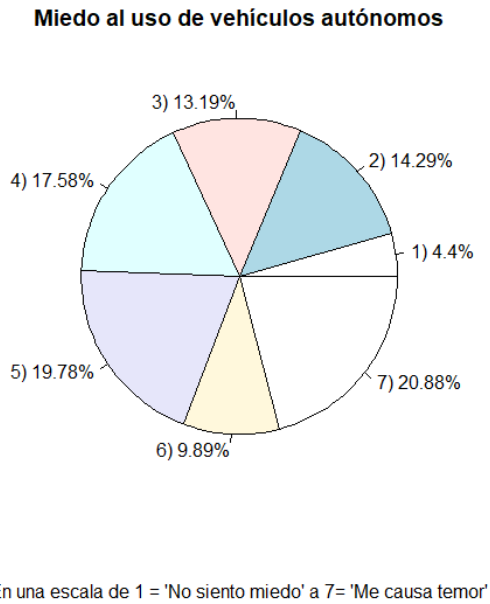
Figura 9. Probabilidad de compra de un vehículo autónomo



Fuente: elaboración propia a partir de Bonnefon, Shariff & Rahwan (2016)

Teniendo todo ello en consideración, a la luz de los resultados obtenidos, se puede concluir que no existe un gran entusiasmo por la popularización de los vehículos autónomos en un futuro, lo cual puede deberse, en gran parte, al temor que su uso genera entre los participantes (Figura 10). Por su parte, dicho entusiasmo es aún menor si éstos se encuentran sometidos a una normativa gubernamental. Por consiguiente, si bien es cierto que la regulación de los algoritmos de IA es necesaria, ésta puede resultar contraproducente, retrasando el uso generalizado de los vehículos autónomos.

Figura 10. Temor causado por el uso de vehículos autónomos



Fuente: elaboración propia a partir de Bonnefon, Shariff & Rahwan (2016)

Más aún, aunque los participantes coinciden en que la minimización del perjuicio causado es la opción éticamente correcta y, por ende, preferirían que otros viajasen en vehículos utilitarios, lo cierto es que todos ellos se sentirían más seguros en un vehículo programado para proteger a sus pasajeros a toda costa. Por ello, los algoritmos morales causan un dilema social (Bonnefon, Shariff & Rahwan, 2016) y, si bien es cierto que la regulación gubernamental puede contribuir a solucionar este problema, el análisis regresivo llevado a cabo demuestra la existencia de dos retos fundamentales: (i) la mayoría de los ciudadanos estadounidenses no se muestra a favor de algoritmos sometidos a la doctrina utilitaria cuando ellos mismos o sus familiares son los pasajeros del vehículo; y (ii) la aprobación de una normativa gubernamental podría retrasar la adopción de los vehículos autónomos.

En conclusión, determinar el componente ético exigido en la toma de decisiones por máquinas autónomas constituye uno de los principales retos de la IA, ya que ello implica buscar un equilibrio entre intereses que a primera vista pueden resultar antitéticos, como la seguridad del ser humano y la moralidad del algoritmo.

6. CONCLUSIONES

La delimitación conceptual efectuada en la primera parte del presente trabajo de investigación ha permitido fijar una caracterización débil o específica de la IA. Ello se debe a que, en línea con John Searle y frente a la postura de Alan Turin, el ser humano no ha sido capaz de desarrollar un algoritmo que permita a los sistemas inteligentes pensar o actuar como humanos. Por el contrario, la IA busca optimizar la respuesta ante problemas o situaciones concretas, de modo que carece de sentido común. A tal efecto, el desarrollo de *AlphaGo* por Google ha demostrado que los algoritmos de IA pueden aprender por sí mismos, si bien ello no implica que los sistemas inteligentes presenten una estructura semántica dotada de contenido e intencionalidad. Consecuentemente, una primera conclusión que puede extraerse de este artículo es que la IA, a día de hoy, no puede replicar la inteligencia humana, sino que dota a las máquinas computacionales de aptitudes concretas requeridas en la realización de tareas determinadas para obtener mejores resultados que el ser humano.

Por su parte, como ya afirmó Peter Norvig, es fundamental garantizar que todo el mundo se beneficia de esta tecnología, de modo que las decisiones adoptadas por los sistemas inteligentes deben ser afines a los valores fundamentales proclamados por el

ser humano. Ello se concreta en la necesidad de crear una IA robusta, ética y legal basada en los 23 Principios de Asilomar, los cuales, como se ha podido apreciar en este artículo, son defendidos por la totalidad de la comunidad científica, las instituciones internacionales y las grandes empresas tecnológicas. Consecuentemente, es posible concluir que el componente ético exigido en la toma de decisiones por máquinas inteligentes viene dado por la transparencia, trazabilidad, seguridad, responsabilidad y, esencialmente, por la intervención y supervisión humana en todas las fases de programación del algoritmo. Ello es lo que permitirá configurar una IA amigable, creando así sistemas inteligentes sometidos al principio de proporcionalidad ente fines y medios y permitiendo al usuario oponerse a las decisiones adoptadas por las máquinas.

En esta línea de pensamiento, en lo que a los perjuicios causados por sistemas inteligentes respecta, considero que el concepto de IA defendido en este artículo determina la necesidad de imputar la responsabilidad de acuerdo con el enfoque de gestión de riesgos, es decir, en proporción al nivel de autonomía del sistema. En consecuencia, cuanto mayor sea la autonomía del algoritmo, mayor será la responsabilidad imputable a su programador. Por el contrario, los perjuicios causados por sistemas con una autonomía reducida o limitada deberían imputarse al ser humano encargado de su control. Ello se debe a que, si partimos de la premisa de que la IA se encuentra ligada a la supervisión humana, no podemos eximir al hombre de responsabilidad por los daños derivados de su aplicación. Por consiguiente, se puede concluir que no es posible atribuir personalidad jurídica a los robots o sistemas inteligentes, ya que, a mi parecer, ello supondría aproximar el concepto de IA a aquel de una inteligencia general, es decir, totalmente autónoma y capaz de actuar por sí misma.

Por todo ello, urge aprobar un marco normativo internacional que, en línea con los Principios de Asilomar y en cooperación con las empresas y organizaciones desarrolladoras de IA, garantice la inversión en esta tecnología, así como la investigación de sus aplicaciones beneficiosas, combatiendo a su vez los usos perjudiciales de la misma. Se trata, por tanto, de continuar la labor iniciada por la UNESCO y de desarrollar un régimen antropocéntrico de IA al servicio del ser humano.

Finalmente, el análisis regresivo centrado en el caso particular de los vehículos autónomos ha permitido extraer conclusiones de gran importancia a la hora de determinar el componente ético exigido en la programación de los mismos. Sin perjuicio de ello, con carácter previo al estudio de éstas es preciso destacar que, tal y

como se indica en el apartado correspondiente, el análisis se basa en una encuesta completada por ciudadanos estadounidenses de entre 18 y 72 años. Por consiguiente, dado que las opiniones éticas o políticas pueden depender fuertemente del entorno sociocultural, debe realizarse una interpretación amplia de las consecuencias derivadas del estudio, que han de entenderse en el contexto de países democráticos, laicos y capitalistas como es EEUU.

Teniendo todo ello en consideración, no cabe duda de que los vehículos autónomos presentan múltiples beneficios para el ser humano, si bien el análisis regresivo llevado a cabo denota una desconfianza generalizada hacia el uso de esta tecnología. En gran medida, ello se debe a los dilemas morales y sociales que plantea el hecho de someter las decisiones de la conducción a un sistema inteligente. En esta línea de pensamiento, el estudio revela que, como ciudadanos, el algoritmo que la sociedad está dispuesta a aceptar es aquel que realiza una ponderación objetiva de los intereses en juego, garantizando así el menor perjuicio posible en caso de peligro inminente en la carretera. No obstante, la carga moral del algoritmo al que el ser humano está dispuesto a someterse como conductor es mucho menor, ya que, en este caso, los intereses en juego son personales. En mi opinión, este conflicto constituye el principal impedimento para el éxito y la penetración en el mercado de los vehículos autónomos.

A tal efecto, el desarrollo de un marco normativo que regule el algoritmo de IA de los vehículos autónomos se plantea como una posible solución a este problema. No obstante, el estudio llevado a cabo en el presente trabajo de investigación revela una tendencia generalizada hacia el rechazo de la intervención del gobierno en esta materia, lo que podría retrasar la adopción de esta aplicación de la IA. La aprobación pública es crucial para la popularización de cualquier tecnología, si bien la percepción de la sociedad varía fuertemente en función de la cultura o incluso del tiempo, de modo que la opinión del coche autónomo que tenemos hoy muy probablemente sea distinta en un futuro próximo. Como consecuencia de ello, a mi juicio, desarrollar un marco normativo internacional al que deban someterse los algoritmos de estos vehículos no solo es extremadamente complejo, sino que además puede resultar contraproducente. Por el contrario, considero que debería optarse por una configuración de la IA que, respetando las exigencias éticas fijadas por los Principios de Asilomar, surja del diálogo y de la cooperación entre los consumidores de esta tecnología, los programadores de su algoritmo y los órganos de gobierno.

7. BIBLIOGRAFÍA

Alemzadeh, H., Raman, J., Leveson, N., Kalbarczyk, Z. & Iyer, R (2016). Adverse Events in Robotic Surgery: A Retrospective Study of 14 Years of FDA Data. *Plos One*. Disponible en: https://arxiv.org/abs/1507.03518

Amodei, D. (2020, 5 octubre). Learning from Human Preferences. *OpenAI*. Recuperado 28 de febrero de 2022, de https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/

Anyoha, R. (2017, 28 Agosto). The History of Artificial Intelligence. *Science in the News*. https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Barrio Andrés, M. (2018). Derecho de los robots. *La Ley-Wolters Kluwer*.

Bassen, J. (2018). *Scarce skills, not scarce jobs*. En Marseguerra, G. y Tarantola, A., Catholic social teaching in action: facing the challenges of the digital age, 237-247. Ciudad del Vaticano, Libreria Editrice Vaticana. https://www.centesimusannus.org/en/publications/inclusive-growth-and-financial-reforms-global-emeregencies-and-the-search-of-the-common-good-2/

BBC News (2020, 16 septiembre). Uber's self-driving operator charged over fatal crash. Recuperado 12 de febrero de 2022, de https://www.bbc.com/news/technology-54175359

BBC News Mundo (2014, 2 diciembre). Stephen Hawking: «La inteligencia artificial augura el fin de la raza humana». https://www.bbc.com/mundo/ultimas_noticias/2014/12/141202_ultnot_hawking_inteligencia_artificial_riesgo_humanidad_egn

BBC News Mundo (2018, 21 marzo). 5 claves para entender el escándalo de Cambridge Analytica que hizo que Facebook perdiera US$37.000 millones en un día. Recuperado el 13 de febrero de 2022, de https://www.bbc.com/mundo/noticias-43472797

Bonnefon, J.F., Shariff, A. & Rahwan, I. (2016, 24 junio). The Social Dilemma of Autonomous Vehicles. *Science*, pp. 1573-1576. https://www.researchgate.net/publication/301293464_The_Social_Dilemma_of_Autonomous_Vehicles

Brundage, M. & Avin, S. (2018, febrero). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *Oxford University*. https://docs.google.com/document/d/e/2PACX-

1vQzbSybtXtYzORLqGhdRYXUqiFsaEOvftMSnhVgJ-jRh6plwkzzJXoQ-sKtej3HW_0pzWTFY7-1eoGf/pub

Coeckelbergh, M. (2020). AI Ethics (The MIT Press Essential Knowledge series). *The MIT Press*.

Comisión Europea (2018). *IA para Europa. Comunicación de la Comisión al Parlamento europeo, al Consejo Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones*. COM (2018) 237 final. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=COM%3A2018%3A237%3AFIN

Comisión Europea (2020, febrero). *LIBRO BLANCO sobre la inteligencia artificial - un enfoque europeo orientado a la excelencia y la confianza*. Oficina de Publicaciones de la Unión Europea. https://op.europa.eu/es/publication-detail/-/publication/ac957f13-53c6-11ea-aece-01aa75ed71a1

Comisión Europea (2021, 21 abril). *Una Europa Adaptada a la Era Digital: la Comisión propone nuevas normas y medidas para favorecer la excelencia y la confianza en la inteligencia artificial* [Comunicado de prensa]. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age_es

Comisión Europea, Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial (2019). *Directrices éticas para una IA fiable.* https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-es/format-PDF

Cortina, A. (2019). Ética de la inteligencia artificial. *Anales de la Real Academia de Ciencias Morales y Políticas*, pp. 379–394. https://dialnet.unirioja.es/servlet/articulo?codigo=7426666

Daugherty, P. R. & Wilson, J. H. (2018). Human + Machine: Reimagining Work in the Age of AI. *Harvard Business Review Press*.

Domingo Jaramillo, C. (2021). *Utilización del sistema de reconocimiento facial para preservar la seguridad ciudadana*. El criminalista digital, 20–37. https://revistaseug.ugr.es/index.php/cridi/article/view/20899

Fernández, C. y Cortés, I. (2018). Dos centenares de expertos europeos piden que no se reconozca personalidad jurídica a los robots. *Wolters Kluwer*. https://www.wolterskluwer.es/sobrewolters-kluwer/wolters-kluwer-espana/sala-de-prensa/noticias-de-prensa/noticias/Dos-centenares-deexpertos-europeos-piden-que-no-se-reconozca-personalidad-jurid.html

Foundation for Responsible Robotics (2020, 18 marzo). Home | Foundation for Responsible Robotics (FRR). *Responsible Robotics*. Recuperado 3 de marzo de 2022, de https://responsiblerobotics.org/

Future of Life Institute (2021, 15 diciembre). *AI Principles*. Recuperado 3 de marzo de 2022, de https://futureoflife.org/2017/08/11/ai-principles/?cn-reloaded=1

Groth, O., Nitzberg, M. & Esposito, M. (2018, 15 octubre). *AI & Global Governance: A New Charter of Rights for the Global AI Revolution - United Nations University Centre for Policy Research*. United Nations University. Recuperado 21 de enero de 2022, de https://cpr.unu.edu/publications/articles/ai-global-governance-a-new-charter-of-rights-for-the-global-ai-revolution.html

Iberdrola (2021, 30 junio). Ataques cibernéticos: ¿Cuáles son los principales y cómo protegerse de ellos? *Iberdrola*. Recuperado el 19 de enero de 2022, de https://www.iberdrola.com/innovacion/ciberataques

IBM (2021) Qué es la inteligencia artificial en la asistencia sanitaria. España | IBM. https://www.ibm.com/es-es/topics/artificial-intelligence-healthcare

ICRAC (s. f.). About ICRAC. Recuperado 3 de marzo de 2022, de https://www.icrac.net/about-icrac/

Information Commissioner's Office – ICO (2017). *Big data, artificial intelligence, machine learning and data protection*. https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

King, G., Pan, J. & Roberts, M. E. (2017). How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument. *American Political Science Review*, 111(3), pp. 484–501. https://doi.org/10.1017/s0003055417000144

Lincoln, S. (2021, 22 abril). Find What Elon Musk Said About Real-World AI. *The European Business Review*. https://www.europeanbusinessreview.com/find-what-elon-musk-said-about-real-world-ai/

López De Mántaras, R. (2015). Algunas reflexiones sobre el presente y futuro de la Inteligencia Artificial. *Novatica*, pp. 97–101. https://digital.csic.es/handle/10261/136978

Lufkin, B. (2017, 7 marzo). Why the biggest challenge facing AI is an ethical one? *BBC Future*. Recuperado 21 de enero de 2022, de https://www.bbc.com/future/article/20170307-the-ethical-challenge-facing-artificial-intelligence

Manyika, J. & Bughin, J. (2018, 15 octubre). The promise and challenge of the age of artificial intelligence. *McKinsey & Company*. https://www.mckinsey.com/featured-insights/artificial-intelligence/the-promise-and-challenge-of-the-age-of-artificial-intelligence

Marín García, S. (2019). *Ética e inteligencia artificial*. IESE, ST-522. Recuperado de https://dx.doi.org/10.15581/018.ST-522.

McCarthy, J. (2004, 24 noviembre). What is artificial intelligence? *Stanford University, Computer Science Department*. https://borghese.di.unimi.it/Teaching/AdvancedIntelligentSystems/Old/IntelligentSystems_2008_2009/Old/IntelligentSystems_2005_2006/Documents/Symbolic/04_McCarthy_whatisai.pdf

Ministerio de Ciencia, Innovación y Universidades (2019). *Estrategia española de I+D+I en Inteligencia Artificial*. Catálogo general de publicaciones oficiales. https://portal.mineco.gob.es/es-es/ministerio/areas-prioritarias/Paginas/inteligencia-artificial.aspx

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society. In press*. https://journals.sagepub.com/doi/pdf/10.1177/2053951716679679?source=post_page----------

Monasterio Astobiza, A. (2017). Ética algorítmica: Implicaciones éticas de una sociedad cada vez más gobernada por algoritmos. *Dilemata*, nº24, 185–217. https://www.dilemata.net/revista/index.php/dilemata/article/view/412000107

Munera, I. (s.f.). Inteligencia artificial (IA): ¿nos quitan el trabajo los robots? *El Mundo Lab*. https://lab.elmundo.es/inteligencia-artificial/trabajo-robots.html

Muñoz Corcuera, A. (Ed.). (2009). *Esclavos y superhombres: la ética en los relatos de Isacc Asimov*. Asociación Cultural Xatafi. http://www.congresoliteraturafantastica.com/pdf/EnsayosCFyLF.pdf

OCDE (2019, mayo). *Recommendation of the Council on Artificial Intelligence*. OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Orden ETD/670/2020, de 8 de julio, por la que se crea y regula el Consejo Asesor de Inteligencia Artificial [BOE núm. 199, de 22 de julio de 2020]. https://www.boe.es/diario_boe/txt.php?id=BOE-A-2020-8302

Polonsky, V. (2017, 10 agosto). Artificial intelligence can save democracy, unless it destroys it first. *Oxford Internet Institute*. https://www.oii.ox.ac.uk/news-events/news/artificial-intelligence-can-save-democracy-unless-it-destroys-it-first/

PricewaterhouseCoopers (2018, 6 febrero). How will automation impact jobs? *PwC*. Recuperado 8 de enero de 2022, de https://www.pwc.co.uk/services/economics/insights/the-impact-of-automation-on-jobs.html

ProPublica (2020, 29 febrero). Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2103(INL)). https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_ES.html

Rimol, M. (2021, 29 septiembre). Gartner Finds 33% of Technology Providers Plan to Invest $1 Million. *Gartner*. https://www.gartner.com/en/newsroom/press-releases/2021-09-29-gartner-finds-33-percent-of-technology-providers-plan-to-invest-1-million-or-more-in-ai-within-two-years

Rushkoff, D. (2010). *Program or Be Programmed (English Edition)*. OR Books.

Russell, S. & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Schwab, K. (2017). *The Fourth Industrial Revolution*. Currency.

Searle, J. (1986). Minds, Brains and Science: 1984 (Reprint ed.). *Harvard University Press*. http://www.rooyeshedigar.ir/site/wp-content/uploads/2015/03/John-Searle-Minds-Brains-and-Science.pdf

Shaw, J. (2019, 30 abril). Artificial Intelligence and Ethics. *Harvard Magazine*. https://www.harvardmagazine.com/2019/01/artificial-intelligence-limitations

Society of Automotive Engineers (2021, 3 mayo). SAE Levels of Driving Automation. Refined for Clarity and International Audience. *SAE International*. Recuperado 3 de marzo de 2022, de https://www.sae.org/blog/sae-j3016-update

Thomas, S. P. (1998). Editorial: Information Fatigue Syndrome - Is there an epidemic? *Issues in Mental Health Nursing*, *19*(6), 523–524. https://doi.org/10.1080/016128498248818

UNESCO. (2020, septiembre). *First draft of the Recommendation on the Ethics of Artificial Intelligence*. UNESDOC Biblioteca Digital. https://unesdoc.unesco.org/ark:/48223/pf0000373434_spa

UNESCO. *La Inteligencia Artificial en la Educación*. (2021, 13 octubre). UNESCO. https://es.unesco.org/themes/tic-educacion/inteligencia-artificial

Velarde Hermida, O. & Casas-Mas, B. (2018, 29 mayo) La virtualización de las comunicaciones interpersonales. *Chasqui*. *Revista Latinoamericana de Comunicación* (137). pp. 55-72. https://eprints.ucm.es/id/eprint/64435/1/3%20LaVirtualizacionDeLasComunicacionesInterpersonales-6578597.pdf

8. ANEXOS

8.1. Anexo I. R Script

```
library(dplyr)

library(ggplot2)

##### ESTUDIO 1 #####

datos<-Study_1

str(datos)

datos$Philosophy <- factor(datos$Philosophy, labels=c("STAY", "SWERVE"))

str(datos)

datos<-select(datos,-ID)

datos<-datos[datos$Measure!="ExpectMoral", ]

datos<-datos[datos$Driver!="Other", ]

summary(datos)

hist(datos$SlideMoral,

     main = "a) Preferencias de los participantes",

     xlab = "De 0 = 'Proteger al pasajero a toda costa' a 100= 'Minimizar los daños'",

     ylab = "Número de respuestas",

     border = "black",

     col = c("darkgrey"))

hist(datos$SlideExpect,

   main = "b) Expectativas en la programación del algoritmo",

   xlab = "De 0 = 'Proteger al pasajero a toda costa' a 100= 'Minimizar los daños'",

   ylab = "Número de respuestas",

   border = "black",

   col = c("darkgrey"))
```

```
t.test(x=datos$SlideMoral, conf.level=0.95)$conf.int

t.test(x=datos$BuyCar, conf.level=0.95)$conf.int

ggplot(datos, aes(x=Sex, y=BuyCar, fill=Sex)) +

  geom_boxplot() +

  labs(x="Género", y="Propensión a comprarse un vehículo autónomo", fill="Género")
+

  scale_x_discrete(labels=c("Hombre","Mujer")) +

  scale_fill_discrete(labels=c("Hombre","Mujer"))

porcentajes <- as.numeric(round(((prop.table(table(datos$Fearful)))*100),2))

porcentajes

etiquetas <- c("1)", "2)", "3)", "4)", "5)", "6)","7)")

etiquetas

etiquetas <- paste(etiquetas, porcentajes)

etiquetas

etiquetas <- paste(etiquetas, "%", sep = "")

etiquetas

pie(porcentajes, etiquetas,

    main = "Miedo al uso de vehículos autónomos",

    sub = "En una escala de 1 = 'No siento miedo' a 7= 'Me causa temor'")

datos_compra<-datos[datos$BuyCar==7, ]

summary(datos_compra)


##### ESTUDIO 2 #####

datos2<-Study_3

summary(datos2)
```

```
datos_familiar<-datos2[datos2$Condition!="You", ]

datos_familiar<-datos_familiar[datos_familiar$Condition!="Coworker", ]

datos_solo<-datos2[datos2$Condition!="Family", ]

datos_solo<-datos_solo[datos_solo$Condition!="Coworker", ]

datos_tercero<-datos2[datos2$Condition!="You", ]

datos_tercero<-datos_tercero[datos_tercero$Condition!="Family", ]

mu0<- mean(datos_familiar$Moral, na.rm = TRUE)

t.test(datos_familiar$Moral,mu=mu0,alternative="two.sided",conf.level = 0.95)

median(datos2$BuySelfProtective, na.rm = TRUE)

median(datos_familiar$BuySelfProtective, na.rm = TRUE)

median(datos_solo$BuySelfProtective, na.rm = TRUE)

median(datos_tercero$BuySelfProtective, na.rm = TRUE)

ggplot(Study_3, aes(x=Condition, y=BuySelfProtective, fill=Condition)) +

  geom_boxplot() +

  labs (x="", y="De 0 = 'Minimizar los daños' a 100= 'Proteger al pasajero'", fill="Condición") +

  scale_x_discrete(labels=c("Solo","Compañero de trabajo", "Familiar")) +

  scale_fill_discrete(labels=c("Solo","Compañero de trabajo", "Familiar"))


##### ESTUDIO 3 #####

datos3<-Study_6

summary(datos3)

datos3_hijo<-datos3[datos3$Condition!="You", ]

datos3_hijo<-datos3_hijo[datos3_hijo$Condition!="Family", ]

datos3_solo<-datos3[datos3$Condition!="Family", ]
```

```
datos3_solo<-datos3_solo[datos3_solo$Condition!="Kid", ]

datos3_familiar<-datos3[datos3$Condition!="You", ]

datos3_familiar<-datos3_familiar[datos3_familiar$Condition!="Kid", ]

mu1<- mean(datos3_solo$GovReg, na.rm = TRUE)

t.test(datos3_solo$GovReg,mu=mu1,alternative="two.sided",conf.level = 0.95)

median(datos3$BuyRegulated, na.rm = TRUE)

median(datos3$BuyUnregulated, na.rm = TRUE)

boxplot(datos3$BuyRegulated,

    datos3$BuyUnregulated,

    xlab = "Vehículo regulado                    Vehículo no regulado",

    ylab = "Probabilidad de compra",

    border = "black",

    col = c("darkgrey"),

    outline = TRUE)
```