PROPUESTA PARA EL LIBRO BLANCO SOBRE LA INTELIGENCIA ARTIFICIAL:

Un enfoque europeo orientado a la excelencia y la confianza

BEATRIZ ALEGRE VILLARROYA

Graduada en el Programa Conjunto
Derecho - Administración y Dirección de Empresas
Universidad de Zaragoza, España (2014-2020)
beatrizalegrevillarroya@gmail.com
https://www.linkedin.com/in/beatriz-alegre-villarroya/

RESUMEN

La adopción de la inteligencia artificial plantea nuevos retos para los que la legislación actual no está preparada. Esta propuesta para el Libro Blanco sobre IA enfatiza algunos de los riesgos de esta tecnología, tratando de aportar desde una perspectiva práctica posibles soluciones para los mismos. Sobre la base de lo sugerido por la Comisión Europea, se recogen recomendaciones para la adaptación y la elaboración de normativa en materia de propiedad intelectual e industrial, transporte terrestre, protección de datos, derecho de consumo, seguros y, finalmente, sobre el régimen de responsabilidad por daños causados por sistemas de IA. Las medidas planteadas deben ser entendidas como una contribución a la futura regulación de la IA, poniendo relevancia en algunos aspectos que pueden ser matizados o desarrollados en mayor profundidad.

ABSTRACT

The adoption of artificial intelligence raises new challenges for which current legislation is not prepared. This proposal for the White Paper on AI emphasizes some of the risks of this technology, trying to provide possible solutions for them from a practical perspective. Based on the suggestions of the European Commission, recommendations are collected for the adaptation and drafting of regulation on intellectual and industrial property, land transportation, data protection, consumer law, insurance, and, finally, on the liability for damage caused by AI systems. The proposed measures must be understood as a contribution to the future regulation of AI, putting relevance in some aspects that can be nuanced or developed in greater depth.

ÍNDICE

1. Introducción 2

2. Concepto de inteligencia artificial 2

3. Adopción de la inteligencia artificial 5

3.1. Riesgos de la IA 6

3.2. Privacidad y la Protección de Datos 9

3.3. Seguridad del sistema de IA 12

3.4. Principio de no discriminación 17

3.5. Responsabilidad por daños y perjuicios causados por la IA 23

4. Regulación de la inteligencia artificial 25

4.1. Posibles adaptaciones del marco legislativo europeo a la IA 25

4.1.1. Derecho de Propiedad Intelectual e Industrial 25

4.1.2. Derecho de Transporte Terrestre 27

4.1.3. Derecho de Protección de Datos 29

4.1.4. Derecho de Consumo y Responsabilidad por Productos Defectuosos 30

4.1.5. Derecho de Seguros 33

4.2. Elaboración de nuevas normas sobre IA 34

4.2.1. Directrices Éticas para una IA fiable 34

4.2.2. Reglamento sobre Responsabilidad por los sistemas de IA 34

5. Conclusión 43

6. Bibliografía 44

7. Documentación 46

1. Introducción

La inteligencia artificial está desarrollándose rápidamente, siendo implementada de forma transversal en diversos campos, como el económico o el biomédico. Su adopción trae consigo grandes oportunidades dada la eficiencia y objetividad característica de los sistemas de IA, pero también presenta riesgos que pueden derivar potencialmente en la vulneración de derechos fundamentales como la protección de datos personales y la privacidad o la no discriminación a través de decisiones parciales.

El doble objetivo planteado por la Comisión Europea es promover la adopción de la IA y abordar los riesgos asociados con ciertos usos de esta nueva tecnología, siendo el propósito del Libro Blanco[1] establecer posibles políticas sobre cómo lograr estas metas. Sin un marco regulatorio claro, la línea de actuación para prevenir, corregir y, en última instancia, determinar la responsabilidad es incierta. Es necesario proporcionar a los Estados Miembros una guía con medidas específicas que, más allá del establecimiento de unos principios éticos básicos para el diseño y utilización de la IA, ayuden a concretar los problemas que pueden surgir y las soluciones a los mismos.

La presente propuesta trata de contribuir a la adaptación y elaboración legislativa sobre inteligencia artificial a nivel de la Unión Europea, profundizando especialmente en la identificación de riesgos de los sistemas de IA y la consideración de aspectos técnicos, éticos y jurídicos para mitigar los mismos.

2. Concepto de inteligencia artificial

El 8 de abril de 2019 fue propuesta por el Grupo de Expertos de Alto Nivel sobre IA una Definición de inteligencia artificial[2], señalando sus principales capacidades y disciplinas científicas. Se parte de un concepto de sistemas de inteligencia artificial como «sistemas de software (y posiblemente también de hardware) diseñados por humanos que, con un objetivo complejo, actúan […] interpretando los datos estructurados o no estructurados

[1] COMISIÓN EUROPEA, *White Paper on Artificial Intelligence: a European approach to excellence and trust*, 19 de febrero de 2020, disponible en <https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en>. Fecha de consulta: 21 de marzo de 2020.

[2] GRUPO DE EXPERTOS DE ALTO NIVEL EN IA DE LA COMISIÓN EUROPEA (a), *A Definition Of AI: Main Capabilities And Disciplines*, 8 de abril de 2019, disponible en <https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines>. Fecha de consulta: 22 de marzo de 2019.

recopilados, razonando sobre el conocimiento, o el procesamiento de la información, derivado de estos datos y la decisión de las mejores acciones para lograr el objetivo dado».

Desde la perspectiva de la disciplina científica, se parte de la idea abstracta de sistemas de IA descrita en el párrafo anterior y se propone una clasificación en tres grupos, con carácter general y sin perjuicio de la existencia de otras técnicas y disciplinas: razonamiento y toma de decisiones, aprendizaje y robótica. El primer grupo comprende los sistemas más complejos que, a través de la aplicación y combinación de diversas técnicas, tienen como objetivo la optimización de una solución ante un problema concreto. Los sistemas de IA referidos al aprendizaje incluyen *machine learning* o *deep learning*, redes neuronales, árboles de decisiones y otras técnicas que permiten al sistema aprender a resolver problemas que no pueden especificarse de forma precisa o cuya solución no puede ser descrita mediante reglas simbólicas de razonamiento. El aprendizaje de estos sistemas puede asimismo clasificarse en términos generales como supervisado, no supervisado o reforzado. Dicha distinción dependerá de si el algoritmo de optimización del problema es modificado o no al recibir nuevos datos del entorno en el que la IA está funcionando, es decir, si el sistema de IA continúa aprendiendo durante su funcionamiento o se mantiene inmutable desde su entrenamiento bajo supervisión. Finalmente, la robótica consiste en la actuación de la IA a través de una máquina físicamente tangible, lo que comúnmente se conoce como robot.

La definición de IA propuesta por la Comisión Europea, centrada en los datos y los algoritmos como sus principales elementos, es a priori amplia y flexible para poder acomodarse al progreso tecnológico, pero a su vez precisa para contribuir a la creación de un ámbito objetivo de las normas que perdure en el tiempo. Es importante que la normativa que sea elaborada o vaya a ser adaptada parta de un concepto de IA con esas características, tratando de prevenir que el ámbito objetivo de la norma quede obsoleto ante los avances tecnológicos futuros.

Si bien es cierto que el concepto de inteligencia artificial del que se parte en las Directrices Éticas para una IA fiable[3] es consistente con el anteriormente referido, podría generarse cierta confusión con respecto al carácter determinista de los algoritmos. Que un algoritmo sea determinista implica que, para una misma entrada de datos, la IA proporcionará en

[3] GRUPO DE EXPERTOS DE ALTO NIVEL EN IA DE LA COMISIÓN EUROPEA (b), *Ethics Guidelines for Trustworthy AI*, 8 de abril de 2019, disponible en <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>. Fecha de consulta: 17 de noviembre de 2019.

todos los casos la misma salida. A pesar de no estar dotada de cognoscencia propia, la IA se basa en el aprendizaje autónomo, lo que genera semejanza con la inteligencia humana[4]. En este punto reside el principal problema de distorsión en cuanto a su entendimiento, puesto que, si bien es asimilable su forma de procesar información a la de una persona, esto no significa que exista la posibilidad de que desarrolle ideas por sí misma. La inteligencia artificial trabaja dentro de las relaciones lógicas que establece el algoritmo fruto de su entrenamiento, por lo que no debe ser entendida como un ente pensante, sino como una herramienta matemática como lo es una ecuación, sólo que con un funcionamiento mucho más sofisticado.

Una variación en el *output* o resultado obtenido por el sistema de IA responde a una modificación del *input*. La opacidad de los sistemas de inteligencia artificial puede dificultar el entendimiento de esa relación en sistemas de aprendizaje continuo no supervisado, ya que como previamente se ha expuesto, estos sistemas modifican su algoritmo en función del éxito de sus resultados. Ello no quiere decir, en ningún caso, que la naturaleza del sistema de IA devenga no determinista, puesto que no existe arbitrio en la toma de decisiones, sino un dinamismo que se materializa en la aparición de diferentes decisiones para problemas idénticos, pero siempre respaldado por un cambio de circunstancias reflejado en los datos proporcionados para el entrenamiento de la IA[5].

Si se parte de un concepto de IA que refleje una naturaleza no determinista, podría entenderse que la generación de resultados distintos se debe a la arbitrariedad del sistema. Por el contrario, si se parte de un aprendizaje dinámico, toda decisión está basada en datos y criterios que, con mayor o menor dificultad, podrían ser explicables. Este extremo es relevante para la determinación de un régimen de responsabilidad, puesto que, presuponiendo una imposibilidad de control del sistema por esa supuesta arbitrariedad, podría llegar a plantearse la exoneración de la responsabilidad en el caso de decisiones o actuaciones del sistema no explicables. Sin embargo, el carácter determinista del algoritmo excluye esa arbitrariedad, debiendo orientarse el régimen de responsabilidad

[4] VAN GERVEN, M. y BOHTE. S., *Artificial Neural Networks as Models of Neural Information Processing*, Frontiers in Computational Neuroscience 11:114, 2017, pp. 1-2. DOI: 10.3389/fncom.2017.00114.

[5] ELDRED, C., ZYSMAN, J. y NITZBERG, M., *AI and Domain Knowledge: Implications of the Limits of Statistical Inference*, BRIE / WITS Technology Briefing, Berkeley, 2019, pp. 1-11, disponible en <https://ssrn.com/abstract=3479479>. Fecha de consulta: 18 de mayo de 2020.

hacia la observación de las relaciones establecidas por el algoritmo, velando por su transparencia y explicabilidad y forzando a quienes diseñen los sistemas de IA a reducir en la medida de lo posible la opacidad.

Por todo ello, la definición de IA de la cual debe partir la elaboración o adaptación de la normativa debe ser suficientemente amplia, para procurar que perdure en un futuro a corto y medio plazo a pesar de los avances tecnológicos, y al mismo tiempo precisa, atendiendo a las características de esta tecnología que singularizan su comportamiento. La seguridad de los sistemas de IA y la salvaguarda de los derechos fundamentales que pueden verse comprometidos por sus decisiones y actuaciones sólo puede conseguirse a través de una comprensión íntegra de la inteligencia artificial y de la continua revisión de los progresos y evoluciones de la misma.

3. Adopción de la inteligencia artificial

El desarrollo y aplicación de la inteligencia artificial está siendo impulsado a gran velocidad. En este contexto, es difícil que la legislación y la aplicación e interpretación de las normas evolucionen conforme lo hace la tecnología. Esta disparidad de ritmos puede generar incertidumbre, temor y, desde una perspectiva más práctica, lagunas jurídicas que antes no existían. Es por ello por lo que debe adoptarse una postura de conocimiento de la situación actual y anticipación a los nuevos retos. Solamente conociendo el potencial de la IA, desde el punto de vista positivo y negativo, podrán establecerse normas que verdaderamente se adapten a esta nueva realidad y solucionen los conflictos jurídicos que la misma trae consigo sin poner en peligro el desarrollo y la innovación.

Los sistemas de IA presentan grandes ventajas por la propia naturaleza de los algoritmos, siendo el alto nivel de precisión de los resultados obtenidos una de las principales. El elevado volumen de datos analizados permite que a través de la IA pueda optimizarse la seguridad de los productos, la eficiencia en la toma de decisiones o el acierto de los modelos de predicción. Además, dichos resultados no están influidos por el punto de vista del decisor, lo que permite que los mismos sean más objetivos. Por otra parte, el proceso de clasificación y predicción con IA es considerablemente más rápido en comparación con el que lleva a cabo un ser humano y, además, exige un menor nivel de recursos en cuanto a tiempo y personal invertidos.

El aprendizaje automático y, en particular, su aplicación a sistemas autónomos de toma de decisiones se ha extendido a campos tan diversos como el diagnóstico de enfermedades, la predicción de delitos y la evaluación de seguros[6]. Estos sistemas de toma de decisiones estarían englobados en el primer grupo de sistemas de IA que se describen en la Definición que aporta el Grupo de Expertos de Alto Nivel[7], correspondiendo a los sistemas de razonamiento a través de los cuales trata de optimizarse un problema concreto.

Las decisiones en estas áreas pueden tener implicaciones éticas o legales, por lo que es necesario que el sistema sea utilizado bajo una perspectiva que vaya más allá del objetivo de maximizar la precisión de la predicción, debiendo considerar y ponderar, en su caso, el impacto que pueden tener las decisiones generadas para la sociedad[8]. En este sentido, debe priorizarse la seguridad de los modelos de IA y la salvaguarda de los derechos fundamentales que pueden verse afectados por los mismos.

3.1. Riesgos de la IA

Los riesgos potenciales inherentes a los sistemas de IA pueden tener distinta magnitud, en función de diversos aspectos o criterios. Todo parece indicar que, con el objetivo de atender y mitigar estos riesgos, va a optarse por una clasificación dual: sistemas de alto riesgo y sistemas de bajo riesgo. Una diferenciación en dos grupos limita las posibilidades de crear medidas adecuadas y suficientemente adaptadas para los sistemas de IA, puesto que, al no existir categorías más específicas, habrá una amplia diversidad de sistemas dentro de cada uno de los conjuntos. La división en más grupos o, al menos, en subgrupos, posibilitaría una mayor concreción de los criterios de inclusión o pertenencia y de las estrategias de mitigación. No obstante, dejando a un lado el debate sobre si el planteamiento de una clasificación dual es el más adecuado, debe atenderse a la aplicación práctica de esta distinción.

Precisar qué sistemas de IA serán considerados de alto riesgo o bajo riesgo supone una clasificación muy importante, dado que las medidas de prevención de daños, control del

[6] TAN, Z., YEOM, S., FREDRIKSON, M. y TALWALKAR, A., *Learning Fair Representations for Kernel Models,*ArXiv preprint, arXiv:1906.11813, 2019, pp. 1-15.

[7] GRUPO DE EXPERTOS DE ALTO NIVEL EN IA DE LA COMISIÓN EUROPEA (a), *A Definition Of AI…cit.*

[8] KLEINBERG, J., LUDWIG, J., MULLAINATHAN, S. y SUNSTEIN, C.S., «Discrimination in the Age of Algorithms», *Journal of Legal Analysis*, Vol. 10, 2018, pp. 113-174. DOI: 10.1093/jla/laz001.

sistema, corrección de deficiencias y el régimen de responsabilidad serán distintas para cada grupo. Además, es importante tener presente que esa diferenciación debe hacerse atendiendo a criterios claros y concretos, pero suficientemente amplios para dar cabida a los avances que puedan darse en los sistemas de IA y las nuevas tecnologías que puedan surgir. La indeterminación o la ausencia de unas reglas de clasificación concisas puede ocasionar que determinados sistemas de IA se encuentren entre los límites de estos dos grupos.

En el Proyecto de Informe que lleva a cabo la Comisión de Asuntos Jurídicos del Parlamento Europeo sobre el régimen de responsabilidad civil en materia de IA[9], que se examinará en mayor profundidad más adelante, se aporta una definición de sistema de IA de alto riesgo. Se entiende como tal aquel cuyo funcionamiento autónomo conlleva «un potencial y significativo riesgo de causar un perjuicio a una o más personas, de forma aleatoria o imposible de predecir; dependiendo la magnitud del riesgo de la relación entre la gravedad del posible perjuicio, la probabilidad de que el riesgo se materialice y el modo en que se utiliza el sistema de IA».

Esta definición resulta poco precisa, lo que conlleva que existan algunos inconvenientes a la hora de su aplicación práctica. En primer lugar, parece extraerse de la misma que únicamente es un sistema de IA de alto riesgo aquel que puede causar un perjuicio de forma aleatoria o imposible de predecir, cuando en realidad los sistemas de IA son deterministas, si bien pueden producirse resultados no esperados o difíciles de explicar por la opacidad o complejidad del sistema. Además, los sistemas que no sean opacos y que funcionen correctamente pueden presentar también alto riesgo en determinadas circunstancias, cuando por ejemplo afectan a derechos fundamentales como la vida o la salud. En el caso de una máquina que a través de IA puede llevar a cabo operaciones quirúrgicas, a pesar de que el sistema puede estar bien calibrado y tener un margen de error mínimo, debe considerarse la conveniencia de que sea calificado como sistema de alto riesgo dada la función que realiza, para que el nivel de prevención, control, corrección y responsabilidad sea el máximo posible.

[9] COMISIÓN DE ASUNTOS JURÍDICOS DEL PARLAMENTO EUROPEO, *Proyecto de Informe con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial* (2020/2014 (INL)), disponible en <https://www.europarl.europa.eu/doceo/document/JURI-PR-650556_ES.pdf>.

La gravedad del posible perjuicio, la probabilidad de materialización del riesgo y el modo de utilización de la IA son criterios que, por el contrario, pueden ser más adecuados para hacer esta clasificación. Se dice en el Informe que «el grado de gravedad debe determinarse sobre la base de la magnitud del daño potencial resultante del funcionamiento, el número de personas afectadas, el valor total del posible perjuicio y el daño a la sociedad en su conjunto». Parece entenderse de lo anterior que se valorará este criterio en función de cómo y por qué cuantía puedan afectar los potenciales daños a la esfera del perjudicado, incluyendo menoscabos materiales y personales, a cuántas personas puedan afectar dichos daños y la relevancia de estos para la sociedad.

Sobre la probabilidad de materialización del riesgo, uno de los factores más importantes a tener en cuenta, se dice que «debe determinarse sobre la base del papel de los cálculos algorítmicos en el proceso de toma de decisiones, la complejidad de la decisión y la reversibilidad de los efectos». La abstracción de esta afirmación, especialmente cuando se refiere al «papel de los cálculos algorítmicos en el proceso de toma de decisiones», requiere de una aclaración o concreción. Lo anterior, sumado a la subjetividad que implica medir la complejidad de la decisión o la dificultad que puede suponer determinar si los daños causados son reversibles y hasta qué punto lo son, hace que resulte cuestionable la utilización de estas características para medir el nivel riesgo de la IA. Un criterio que permite determinar esa probabilidad de materialización del riesgo de forma más objetiva es el margen de error de un sistema de IA, en el sentido de permitir el funcionamiento de los sistemas que tengan un porcentaje de éxito mínimo en sus resultados.

Establecer un límite máximo en el margen de error puede ser la solución para evitar que sean utilizados sistemas de IA deficientes que arrojen resultados incorrectos. Sin embargo, no es suficiente con determinar el margen de error que es permisible, puesto que para que esta medida sea útil debe acompañarse del menor nivel de incertidumbre posible. La incertidumbre de una IA mide la previsibilidad del comportamiento de la IA, de manera que, a mayor nivel de incertidumbre, menor es la certeza sobre el margen de error. Un sistema de IA que no esté bien calibrado y, por tanto, tenga un nivel de incertidumbre elevado, puede presentar un margen de error inexacto, de manera que la probabilidad de comportamientos fuera de lo esperado sea superior a la que el sistema indica.

Es por ello por lo que, para asegurar que la probabilidad de materialización del riesgo se minimiza, primero debe requerirse un nivel de incertidumbre de los sistemas de IA que

permita aproximar con suficiente exactitud el margen de error. El segundo paso, una vez se ha calculado de forma precisa dicho margen, es concretar a partir de qué nivel el sistema de IA no es aceptable por tener un porcentaje insuficiente de resultados satisfactorios.

Por último, el denominado modo de utilización de la inteligencia artificial se refiere «al sector en el que opera el sistema de IA, si puede tener efectos jurídicos o reales sobre derechos importantes de la persona afectada protegidos desde el punto de vista jurídico y si los efectos pueden evitarse razonablemente». Cabe suponer, siguiendo lo anterior, que el nivel de riesgo será mayor si las acciones de la IA afectan a personas especialmente vulnerables o actúan en áreas que requieren especial protección, como podrían ser la sanidad, el medio ambiente o la educación.

Una valoración y ponderación de estos tres factores puede ayudar a identificar el nivel de riesgo para cada sistema de IA. Por otra parte, debe estudiarse si puede ser interesante introducir algunos ejemplos concretos que sirvan como referencia para los sistemas de difícil identificación. En este sentido, el Parlamento Europeo recomienda en su Informe que todos los sistemas de IA de alto riesgo figuren en un anexo de la propuesta de Reglamento que sea sometida a una revisión periódica cada seis meses. Sin embargo, debe ampliarse la explicación con respecto a esta lista de sistemas de alto riesgo, concretando si es una lista orientativa o una lista cerrada, qué criterios de inclusión o permanencia se seguirían o si dentro de la misma hay distintos niveles o todos se acogerán al mismo tipo de régimen de responsabilidad y medidas preventivas.

Para que la clasificación de sistemas de IA en función de su riesgo sea adecuada, debe atenderse a los principios de no discriminación y de privacidad y protección de datos, dado que son dos de los aspectos que más protección requieren dada la situación y el estado actual de avance de esta tecnología. A continuación, se exponen algunos de los riesgos concretos que existen respecto a estos derechos, aportándose posteriormente una línea de actuación para mitigar los mismos.

3.2. Privacidad y la Protección de Datos

La inteligencia artificial tiene un gran potencial para procesar y analizar datos a gran escala. Si bien la intención que se persigue con su adopción es el buen uso de dicha capacidad para asistir y ayudar a la sociedad, una incorrecta utilización de esta tecnología puede suponer una amenaza hacia la privacidad e intimidad de las personas. Los datos

procesados, la forma en que se diseñan las aplicaciones y el alcance de la intervención humana pueden afectar los derechos de libertad de expresión, la protección de datos personales, la privacidad y las libertades políticas.

La privacidad puede verse afectada no sólo por aquellos datos sensibles que manifiestamente se contengan en el *big data*. Dada la capacidad de inferencia estadística de la IA, existe riesgo de extracción de datos sensibles a través del algoritmo. La inferencia de datos sensibles consiste en la inducción de características concretas del individuo, como pueden ser la orientación sexual, edad, opiniones políticas o religiosas, género o raza. Esto es posible a través del proceso inductivo por el cual el algoritmo, durante su entrenamiento a base de prueba y error, lleva a cabo simplificaciones de las variables recogidas en la base de datos que tienen relación entre sí e infiere una nueva variable[10]. Esta nueva variable es la que puede representar una característica del individuo que la propia persona no haya querido revelar. Por ejemplo, si una persona proporciona datos como los estudios cursados, el tipo de productos o servicios consumidos o sus aficiones o intereses, es posible que pueda inferirse estadísticamente su género.

El uso potencial de la información necesaria para usar los sistemas de IA o generada por estos mismos puede causar cierta inquietud con respecto a la utilización de la información. En ocasiones, las características que pueden inferirse a través de la IA son datos que pertenecen a la esfera privada de la persona. De esta forma, sin necesidad de preguntar de forma directa al individuo, que en su caso mostraría su conformidad o desacuerdo con proporcionar ciertos datos, puede llegar a conocerse información personal.

Otra preocupación importante con respecto al impacto de la IA en la privacidad de la información es si debe haber límites a las sugerencias personalizadas gracias a los sistemas de IA. Las recomendaciones que llegan al individuo, desde un anuncio publicitario a una lista de reproducción automática, se crean basándose en una construcción de la propia concepción de la identidad de la persona. Si bien en principio la personalización de contenido en función de las preferencias puede tener un impacto positivo, existen ciertos riesgos asociados a esta actividad.

[10] COLE, G.W. y WILLIAMSON, S.A., *Avoiding Resentment Via Monotonic Fairness*, ArXiv preprint, arXiv:1909.01251v1, 2019, pp. 1-16.

Cuando se muestra contenido personalizado a un individuo, se está dejando de mostrar el contenido que no queda registrado entre las preferencias del usuario. Si pensamos en publicidad u ofertas de bienes y servicios, esto puede tener sentido y beneficiar a comprador y vendedor, ya que el primero encuentra más rápido cosas que le gustan y el segundo maximiza el beneficio de su inversión en publicidad al dirigirla al público objetivo. Sin embargo, en el caso de las sugerencias de contenido en las noticias o en prensa, puede ocurrir que una persona comience a recibir únicamente el contenido que es afín a su opinión. En este caso, el derecho a la información se vería puesto en peligro, ya que el lector no tiene acceso por igual a toda la información y su criterio puede verse parcializado por las recomendaciones personalizadas que retroalimentan sus preferencias iniciales. Para evitar problemas de esta índole, podría establecerse un sistema de activación o desactivación de la personalización de contenido por parte del individuo, de manera que pudiese dar su consentimiento expreso a esta función al activarla, siendo consciente de que está recibiendo información acorde a sus preferencias. Si, de forma voluntaria, prefiere omitirse la personalización, podría desactivarse la función en cualquier momento.

En relación con lo anterior, la elección de prestar o no consentimiento para el tratamiento de datos puede causar otros problemas. No todos los usuarios prestan consentimiento, de manera que la información de una facción importante de población no se incluye en los datos que entrenan la IA. En determinadas circunstancias puede suceder que esta situación se extreme, existiendo un riesgo potencial de que la IA no pueda generalizarse bien. Si esto ocurre, probablemente el margen de error de la IA sea más elevado, puesto que durante su entrenamiento no ha tenido acceso a todos los datos posibles para optimizar su funcionamiento en el mundo real. Una posible solución técnica para este inconveniente es la eliminación de la línea de aprendizaje que incorpora los datos no consensuales o el reentrenamiento de los modelos de IA utilizando conjuntos de datos modificados.

La garantía y salvaguarda de la protección de la privacidad individual podría estar amenazada por la adopción de la IA[11], infringiéndose el respeto de la confidencialidad en el uso de los datos, la protección de su integridad y el acceso a los mismos. Es por ello imprescindible que las prácticas con IA se supervisen desde esta perspectiva, evitando la

[11] NELSON, G.S., «Bias in artificial intelligence», *North Carolina Medical Journal*, Vol. 80(4), 2019, pp. 220-222. DOI: 10.18043/ncm.80.4.220.

vulneración de estos derechos y cuidando que se cumpla lo dispuesto en la normativa de protección de datos[12].

3.3. Seguridad del sistema de IA

La utilización de sistemas de IA para determinadas funciones puede ocasionar la presencia de nuevos riesgos para la seguridad, como un accidente de tráfico causado por un error de un vehículo autónomo o una dosis farmacológica inadecuada para un paciente calculada o administrada por un robot médico. Estos problemas pueden derivar de un fallo en el diseño de la IA, la insuficiencia de cantidad o cualidad de datos utilizados en su entrenamiento u otras razones propias del aprendizaje automático. En cualquier caso, los posibles riesgos que puedan aparecer deben ser mitigados para evitar potenciales daños o desconfianza en la adopción de la IA, lo que podría causar finalmente un rechazo hacia esta tecnología y por consiguiente una pérdida de competitividad de las empresas residentes en la UE.

Con el objetivo de reducir el impacto de los riesgos en seguridad de sistemas de IA, puede resultar conveniente establecer una línea de actuación que contenga medidas en tres direcciones: robustez técnica, transparencia y explicabilidad del sistema de IA y dirección y supervisión humana.

Para asegurar la robustez técnica de un sistema de IA es imprescindible que los datos utilizados para su entrenamiento sean suficientes y de calidad y que se minimice el nivel de incertidumbre y el margen de error. En cuanto a los aspectos que deben revisarse relativos a los datos, en primer lugar, debe controlarse que su obtención y finalidad es conforme a la normativa de privacidad y protección de datos. La información recogida debe ser amplia, en número y casuística, para cubrir todos los escenarios relevantes y evitar que haya situaciones peligrosas para las que no se haya entrenado la IA. Finalmente, es importante que los datos sean representativos en cuanto a género, raza o cualquier otro motivo que pueda resultar discriminatorio.

El nivel de incertidumbre representa con qué probabilidad no existe certeza sobre un resultado, de manera que no se sabe con exactitud qué margen de error tiene el sistema de IA. La determinación del nivel de incertidumbre que hay en estos sistemas se

[12] Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE.

determina a través de su calibración. Un sistema bien calibrado será capaz de expresar con gran exactitud la incertidumbre en cada decisión, lo que permite conocer en qué medida la misma es o no confiable, en función de si existe o no certeza sobre que dicha decisión sea correcta[13].

Por tanto, antes de estudiar las medidas de reducción del margen de error de un sistema de IA, debe comprobarse que su nivel de incertidumbre es mínimo. De lo contrario, podría pensarse que el margen de error es elevado cuando en realidad no lo es o viceversa, haciendo que las medidas impuestas traten de dar solución a un problema que no ha sido bien concretado. Cuando un sistema de IA presenta un nivel de incertidumbre superior al aceptable, debe recalibrarse el sistema hasta que se alcance el nivel deseado. Una vez ha sido recalibrado, el margen de error que arroje será más preciso, de manera que puede entonces pasarse a una segunda etapa donde se examine si ese margen de error es adecuado.

El margen de error es un indicador que muestra el número aproximado de veces que la IA no consigue los resultados esperados. El funcionamiento de la IA se basa en el uso de datos para la optimización de un problema, de manera que es prácticamente imposible que no exista margen de error. Lo anterior implicaría que todos los datos posibles han sido utilizados para su entrenamiento, de manera que no exista ningún elemento o escenario nuevo al que pueda enfrentarse el sistema. A pesar de que esto no puede lograrse en su totalidad, sí debe intentarse que ese margen de error que representa las situaciones para las que la IA no ha sido entrenada sea el menor posible. Para minimizarlo, debe procurarse que los datos de entrenamiento sean suficientes y variados, de manera que se recoja en los mismos la mayor casuística posible.

Este no es un aspecto que perdure en el tiempo en todos los casos, por lo que debe revisarse durante la vida de la IA que la misma funciona correctamente y que su margen de error se mantiene bajo. Cuando no sea así, el sistema deberá volver a ser entrenado para adecuarse a las nuevas circunstancias, y tendrá que repetirse este proceso tantas veces como sea necesario. Por ejemplo, un vehículo puede ser entrenado de forma que su sistema de conducción autónoma tenga un margen de error muy reducido. Sin embargo, si no se actualiza el sistema, en el momento en el que una nueva señal sea introducida

[13] ANTORÁN CABISCOL, J., *Understanding Uncertainty in Bayesian Neural Networks*, Departamento de Ingeniería de la Universidad de Cambridge, 2019, *Pro manuscript*, pp. 1-94.

cabe la posibilidad de que el sistema no reconozca la misma y no pueda interpretarla. En función de la IA y de su entrenamiento –supervisado, no supervisado o reforzado–, el sistema requerirá un mayor o menor control del margen de error y actualización.

En el Libro Blanco de IA[14], cuando se hace referencia a la robustez técnica del sistema, se dice que uno de los elementos a considerar es que los sistemas de IA pueden lidiar con errores o inconsistencias durante todas sus fases de vida. Si bien pueden existir diversas interpretaciones sobre esta afirmación, una de las principales causas de la existencia de defectos funcionales es el alto nivel de incertidumbre y el elevado margen de error. Para cumplir con el requisito de robustez técnica de que la IA actúe de forma exacta durante todas las fases de su ciclo, es necesario limitar el uso de sistemas con estas características. En la futura regulación sobre IA debe concretarse el origen de esos «errores o inconsistencias» y tenerse en cuenta los dos aspectos mencionados en la búsqueda de una solución para la optimización del comportamiento robusto de los sistemas de IA.

Por último, otra consideración importante de la robustez técnica es la seguridad *per se* de los sistemas de IA, en el sentido de que los mismos sean resistentes a ataques o intentos de manipulación de los datos o algoritmos. Para ello deben emplearse técnicas adecuadas de protección, de manera que no se pueda tener acceso al sistema de IA ni extraer los datos de entrenamiento. Además, en caso de que los mismos ocurran, deben haber sido previstas medidas suficientes de mitigación o corrección del daño causado.

Una característica que debe acompañar a la robustez técnica y seguridad del sistema de IA es la transparencia y explicabilidad. La robustez persigue procurar que la IA funcione de forma adecuada, tratando de evitar los errores o imprevistos. No obstante, cuando tienen lugar fallos a pesar de los intentos de hacer seguro el sistema, la transparencia y explicabilidad se hacen necesarias. Estas cualidades son esenciales para que el sistema, además de ser seguro, se perciba como tal.

Conforme a lo expuesto en el Libro Blanco[15], la transparencia implica que se informe de forma clara a quienes apliquen el sistema, a las autoridades competentes e incluso a las partes afectadas sobre las capacidades y limitaciones de la IA, su propósito específico y su probabilidad de éxito al funcionar. La información proporcionada debe ser objetiva,

[14] COMISIÓN EUROPEA, *White Paper on Artificial Intelligence: a European approach to excellence and trust…cit.*

[15] COMISIÓN EUROPEA, *White Paper on Artificial Intelligence: a European approach to excellence and trust…cit.*

concisa y fácilmente entendible para las personas a las que va dirigida. Además, en algunos casos se plantea que no sea necesario advertir del uso de la IA si resulta obvio o manifiesto.

Si bien el planteamiento que hace la Comisión Europea sobre la transparencia parte de ideas concretas, es conveniente introducir algunos matices. En primer lugar, no se establece de forma clara si la información se hará siempre accesible para el usuario final de la IA. De hecho, parece interpretarse que esta posibilidad sólo se dará en algunos casos cuando se habla de informar «incluso a las partes afectadas». Tampoco se precisa hasta qué extremo se entenderá que resulta obvia la utilización de la IA, lo cual puede generar dudas para el implementador y cierto temor o inseguridad por parte del usuario expuesto.

La transparencia posibilita el conocimiento y la comprensión del funcionamiento de la inteligencia artificial, un presupuesto básico para la confianza en esta tecnología. Establecer que el sistema de IA sea transparente como un requisito de carácter obligatorio enfatiza la responsabilidad de quienes emplean inteligencia artificial y, por extensión, de quien diseña el sistema. Estos agentes se ven en la obligación de explicar el conjunto de datos concreto que ha servido de entrenamiento a la red neuronal y el algoritmo resultante, lo que permite inferir qué decisiones fueron tomadas y cuál fue el fundamento de las mismas[16].

El problema de la «caja negra» es un desafío que a menudo plantea problemas para quienes han utilizado sistemas inteligentes asistiendo sus decisiones. Ante el requerimiento de las personas afectadas por las mismas, se encuentran en una situación que exige justificar determinadas decisiones y, por tanto, deben buscar información sobre la actividad de la IA. El conflicto surge con algunas técnicas de aprendizaje automático, que a pesar de tener éxito en lo relativo al nivel de precisión, son opacas en términos de explicación de resultados[17]. La noción de IA de caja negra se refiere a tales escenarios, donde no es posible rastrear la justificación que existe detrás de ciertas decisiones[18], lo que puede incluso poner en duda su imparcialidad.

[16] NELSON, G.S., «Bias in artificial intelligence»…cit.

[17] COECKELBERGH, M., «Ethics of artificial intelligence: Some ethical issues and regulatory challenges», *Technology and Regulation*, 2019, pp. 31-34. DOI: 10.26116/techreg.2019.003.

[18] GRUPO DE EXPERTOS DE ALTO NIVEL EN IA DE LA COMISIÓN EUROPEA (a), *A Definition Of AI…cit.*

La explicabilidad es la solución a este problema de opacidad, puesto que, si se logra entender el mecanismo subyacente del sistema y encontrar soluciones para los errores cometidos, desaparece la incertidumbre en cuanto a su funcionamiento. Es por ello por lo que uno de los campos de investigación más vanguardistas en inteligencia artificial es lo que se conoce como XAI o *Explainable Artificial Intelligence*, que consiste en el desarrollo de técnicas y métodos orientados a la explicación del algoritmo de IA.

Una de las medidas propuestas que podría contribuir a mejorar la explicabilidad del sistema de IA es registrar la base de datos de entrenamiento incluyendo una descripción de razones de elección de los datos utilizados, la información sobre el algoritmo y los objetivos establecidos para el sistema. Asimismo, otra de las sugerencias es conseguir que los resultados sean reproducibles, de manera que repitiendo la situación objeto de estudio se observe el funcionamiento de la IA al obtener un resultado. Sin embargo, debe tenerse en cuenta que esto puede ser válido en el caso de sistemas que actúan bajo supervisión, pero no necesariamente en sistemas no supervisados o reforzados. Estos sistemas de IA se caracterizan por estar sujetos a un entrenamiento continuo, por lo que, si el algoritmo se ha modificado como consecuencia de haber obtenido nuevos datos, el resultado que se obtiene para la misma situación podría ser distinto.

La dirección y supervisión humana deviene necesaria en aquellos casos en los que el sistema de IA no es suficientemente robusto técnicamente, transparente o explicable. De esta forma, las deficiencias que pueda presentar la IA pueden ser corregidas por la intervención de un ser humano. Existen diferentes niveles de actuación por parte de la persona, proponiéndose cuatro rangos en el Libro Blanco[19]. El primer nivel supone que los resultados no sean válidos sin la revisión de un ser humano. El segundo, que los resultados sean válidos sin esa revisión, pero que la persona pueda intervenir para modificarlos. En tercer lugar, que exista la posibilidad de que la persona intervenga en tiempo real y monotorice el sistema de IA. Y, por último, que sean impuestas en la fase de diseño de la IA ciertas restricciones que determinen su funcionamiento desde un principio. Podría incluso plantearse para los sistemas de más alto riesgo la opción de impedir su utilización si no se presta conformidad con anterioridad a su lanzamiento al

[19] COMISIÓN EUROPEA, *White Paper on Artificial Intelligence: a European approach to excellence and trust*...cit.

público, tras haberse comprobado el algoritmo resultante o incluso la base de datos con la que ha sido entrenado.

Finalmente, diseñar sistemas de normalización y certificación, entendidos como métodos de acreditación que valoren la calidad del sistema de IA, puede ser una buena medida que, aunque no incrementa la seguridad del sistema, sí ayuda a la transparencia del mismo. Esta opción permitiría a los usuarios tener conocimiento certificado de que el sistema funciona correctamente, en el sentido de esté bien calibrado. A pesar de no poder ofrecer una solución en todo caso para la seguridad, de esta forma se consigue que las personas afectadas por la utilización de la inteligencia artificial sean conscientes de su nivel de incertidumbre y margen de error. La persona puede entonces, con una mayor información, optar o no por utilizar el sistema de IA, recibir productos o servicios que lo integren o incluso decidir qué grado de intervención humana quiere aplicarse.

El Libro Blanco sugiere el etiquetado voluntario para aplicaciones de IA que no son de alto riesgo. Sin embargo, considerando que esta medida no es demasiado gravosa y contribuye a hacer los sistemas de IA más transparentes en cuanto a su seguridad, sería conveniente que no únicamente los sistemas de bajo riesgo sigan un régimen de certificación. Etiquetar los sistemas de alto riesgo para visibilizar el nivel de seguridad o robustez técnica es incluso más necesario que en los de riesgo medio o bajo. Por tanto, el etiquetado debería ser incluso obligatorio en estos casos, quizá dando la opción de etiquetar voluntariamente para los sistemas de bajo riesgo.

Para conseguir la seguridad de un sistema de IA deben aplicarse medidas que aseguren su robustez técnica, transparencia y explicabilidad. De forma preventiva, la dirección y supervisión humana con distintos grados de intervención permite reducir los riesgos inherentes a esta tecnología. Debe adoptarse una postura que combine estos aspectos para priorizar el objetivo de crear sistemas seguros de forma coherente y responsable.

3.4. Principio de no discriminación

Los datos empleados en el entrenamiento de la IA pueden presentar ciertos sesgos por prejuicios históricos u otros factores fuera del control del proveedor de los datos o el desarrollador, de manera que, si los datos no son suficientemente equilibrados o inclusivos, la inteligencia artificial entrenada con los mismos no puede generalizarse bien. En el caso de los sistemas de IA dedicados a la toma de decisiones, estas podrían resultar parciales, favoreciendo a algunos grupos sobre otros por razón de raza, género u

orientación sexual, entre otras [20]. Ante esta situación, es necesario que los modelos de aprendizaje automático encaminados a la toma de decisiones sean diseñados previniendo la interpretación o la práctica discriminatoria[21].

Si bien a priori no parece plantear inconvenientes la eliminación de este tipo de sesgos en las bases de datos, la solución no resulta sencilla por la capacidad de análisis de la IA. La complejidad de este asunto reside en el doble origen de esta parcialidad, que finalmente da lugar a una decisión sesgada. Por un lado, es evidente que, si en el conjunto estructurado de datos se incluye una variable discriminatoria, la decisión resultante estará basada en las correlaciones que se hayan creado entre dicha variable y otras que no lo son, por lo que la casuística puede dar lugar a resultados discriminatorios. Sin embargo, dicha parcialidad puede aparecer incluso en situaciones en las que esa variable, a pesar de no haberse incluido de forma expresa, pueda inferirse por métodos estadísticos.

Mediante inferencia estadística la IA es capaz de inducir características concretas del individuo susceptibles de causar decisiones discriminatorias que afecten al mismo, como la orientación sexual, edad, opiniones políticas o religiosas, género o raza[22]. Cuando debido a estas correlaciones se acaba reconociendo una característica que cumple con la función de optimización y funciona correctamente en términos probabilísticos, la red neuronal se basará en la misma para tomar sus decisiones. Ese rasgo que se infiere del resto de variables y que deviene determinante en el proceso de toma de decisiones, a pesar de no llevar el nombre de la característica a la que representa, causa el mismo resultado que si dicha cualidad existiese desde un principio. Es por ello por lo que, en estos casos, a pesar de no estar presentes dichas variables sensibles en el conjunto de datos de entrenamiento, a través de otras características que están altamente correlacionadas con las que sí se contienen, las decisiones pueden tornarse sesgadas[23].

---

[20] MAYSON, S.G., «Bias in, Bias out», *Yale Law Journal* 128, 2019, pp. 1-84, ¿disponible en <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3257004>. Fecha de consulta: 3 de enero de 2020; DOSHI-VELEZ, F., KORTZ, M., BUDISH, R., BAVITZ, C., GERSHMAN, S., O'BRIEN, D., SCHIEBER, S., WALDO, J., WEINBERGER, D. y WOOD. A., *Accountability of AI under the law: The role of explanation.* ArXiv preprint, arXiv:1711.01134, 2017, pp. 1-15.

[21] KUSNE, M., LOFTUS, J., RUSSELL, C. y SILVA, R., *Counterfactual Fairness*, ArXiv preprint, arXiv:1703.06856v3, 8 de marzo de 2018, pp. 1-18; KLEINBERG, J., LUDWIG, J., MULLAINATHAN, S. y SUNSTEIN, C.S., «Discrimination in the Age of Algorithms»…*cit.*

[22] COLE, G.W. y WILLIAMSON, S.A., *Avoiding Resentment Via Monotonic Fairness…cit.*

[23] TAN, Z., YEOM, S., FREDRIKSON, M. y TALWALKAR, A., *Learning Fair Representations… cit.*

Póngase el caso de una red neuronal que clasifica a un grupo de posibles prestatarios como aptos o no aptos, cuya base de entrenamiento contiene únicamente las variables renta, estudios, contrato de trabajo y lugar de residencia. Si la red neuronal infiere por la información contenida en dichas características la existencia de una nueva variable que permite simplificar en el mayor de los casos el proceso de decisión, la incluirá en sus cálculos para obtener los resultados. Así, si en la mayor parte de los casos la persona con menos solvencia está situada en un rango de 25 a 30 años, tiene un nivel de estudios bajo, trabaja en el sector del textil y vive en un gueto camboyano, dichas características pueden ser abstraídas por la red en una sola que las hace coincidir con base en la estadística: una persona media en España procedente de Camboya. A pesar de las generalizaciones –en todo caso irreales y únicamente al efecto de poner el ejemplo– y la extrema simplificación del caso ilustrativo, la idea subyacente es que el análisis de la inteligencia artificial va más allá de un simple cálculo matemático, puesto que mediante el uso de métodos estadísticos utiliza razonamientos lógicos complejos para la toma de decisiones.

La consecuencia de lo anterior es que, en determinados casos, existe el riesgo de que las personas puedan aprovecharse del uso de la IA y, por extensión, de la estadística, para tomar decisiones que sean discriminatorias en última instancia. Para ello, bastaría con encontrar cualidades concretas que presenten alta correlación con el rasgo que se pretende excluir de la selección e incluirlas, de manera que de forma encubierta se obtengan decisiones sesgadas a través de un procedimiento que aparentemente no lo es.

Buena parte de la doctrina científica[24] ha criticado duramente esta posibilidad, concluyendo que, si es posible inferir a través de un algoritmo la característica sensible en cuestión, el uso de dicho algoritmo no debería permitirse. Sin embargo, podría optarse por una posición más moderada y que transige el uso de inteligencia artificial sin imponer excesivas restricciones. Si en lugar de excluir todos aquellos algoritmos de los cuales puede inferirse una característica concreta –algo que no resulta extraño teniendo en cuenta que puede darse por probabilidad esta casuística, sin que ello conlleve necesariamente una pretensión discriminatoria encubierta–, se lleva a cabo un control exhaustivo de las

[24] YEOM, S., DATTA, A. y FREDRIKSON, M., «Hunting for discriminatory proxies in linear regression models», *Advances in Neural Information Processing Systems*, 2018, pp. 4568-4578, disponible en <http://papers.nips.cc/paper/7708-hunting-for-discriminatory-proxies-in-linear-regression-models.pdf>. Fecha de consulta: 14 de abril de 2020; TAN, Z., YEOM, S., FREDRIKSON, M. y TALWALKAR, A., *Learning Fair Representations… cit.*

variables que se contienen en la base de datos, la decisión que se obtiene finalmente no tiene por qué ser discriminatoria, a pesar de que por cifras pueda parecerlo.

Establecer restricciones que traten de eliminar la discriminación supone, en primer lugar, decidir qué definición de equidad o justicia es la más adecuada para la tarea en cuestión. Un estudio llevado a cabo por investigadores estadounidenses y británicos[25] demuestra que, dependiendo de la relación entre un atributo protegido y los datos, ciertas definiciones de equidad pueden incluso aumentar la discriminación. Sobre la base de esa contradicción, se plantea una definición de equidad, denominada *counterfactual fairness*. Según dicha definición, una decisión es justa hacia un individuo si es la misma en dos escenarios, tanto en el mundo real, como en un mundo contrafactual, que es aquel en el que el individuo pertenecería a un grupo demográfico diferente. Esta definición de equidad señala la necesaria causalidad de las diferentes actuaciones por parte del algoritmo, conocidas como *local explanations,* de manera que analiza si cada una de esas decisiones locales ha sido justa éticamente[26]. Si la decisión final está basada en características cuya justificación de pertenencia en la base de datos no es discriminatoria, la misma no debe entenderse como tal[27].

En la línea del ejemplo anterior, si la persona que selecciona qué variables deben incorporarse en la base de datos es capaz de justificar por qué es conveniente que sean tenidas en cuenta las mismas, el algoritmo no debe entenderse como discriminatorio. La renta, el nivel estudios, el tipo de contrato de trabajo e incluso el lugar de residencia son variables que pueden tener relación con la solvencia de una persona. En consecuencia, no son variables discriminatorias y el algoritmo creado con base en las mismas tampoco debe ser considerado como tal.

Aunque el resultado de la selección de prestatarios pueda dar lugar a la calificación de aptos del 90 por ciento de los españoles y no aptos del 70 por ciento de los camboyanos, ello no significa *per se* que la decisión haya sido discriminatoria. Con base en los criterios que se tienen en cuenta, si el año siguiente se incrementa la población inmigrante camboyana en España altamente cualificada, con un contrato de trabajo con una remuneración elevada que les permita vivir en el centro de la ciudad, las decisiones

[25] KUSNE, M., LOFTUS, J., RUSSELL, C. y SILVA, R., *Counterfactual Fairness… cit.* pp. 1-2.

[26] DOSHI-VELEZ, F., KORTZ, M., BUDISH, R. *et al., Accountability of AI under the law… cit.*

[27] KLEINBERG, J., LUDWIG, J., MULLAINATHAN, S. y SUNSTEIN, C.S., «Discrimination in the Age of Algorithms»...*cit.*

adoptadas cambiarán drásticamente, pudiendo ocurrir que se consideren como aptos el 80 por ciento de los potenciales prestatarios camboyanos.

En conclusión, la existencia de discriminación debe tenerse en cuenta desde un punto de vista objetivo, de manera que únicamente se consideren sesgadas aquellas decisiones que, efectivamente, se basen en criterios que no sean imparciales[28]. Asimismo, los datos que sirven de entrenamiento a la IA deben revisarse y actualizarse, de manera que el algoritmo se modifique conforme lo hace la información que contienen las variables no discriminatorias. De esa manera, la variable que ha surgido por inferencia estadística desaparecerá o cambiará en función de las nuevas circunstancias. Con ello se asegura el principio de igualdad entre las personas, puesto que las decisiones no estarán basadas en variables discriminatorias, sino en razonamiento lógico y estadística.

Especial relevancia adquiere dicha distinción en Derecho de Consumo, ya que en muchas ocasiones tiene lugar la selección de clientes en cuanto a la proposición u oferta de productos o servicios[29]. El consumidor, en los casos en los que crea vulnerado el principio de igualdad o no discriminación, tiene el derecho de conocer la justificación del rechazo de una empresa a prestarle servicios o venderle productos. Ante la falta de una explicación pertinente, podrá accionar contra el mismo por considerar que existe una infracción del principio de no discriminación, recogido no sólo en la legislación del Estado en el que se encuentre, sino también de forma general en el artículo 21 de la Carta Europea de los Derechos Fundamentales[30].

Comienzan a ser numerosos los casos que plantean dudas acerca de la legalidad de la utilización de la IA. Uno de los más conocidos, relacionado con el riesgo de parcialidad y discriminación expuesto es el de la tarjeta de crédito Apple, lanzada en colaboración con Goldman Sachs y Mastercard para los usuarios estadounidenses en agosto de 2019[31]. La tarjeta Apple es una tarjeta de crédito que proporciona una línea de crédito a los usuarios y utiliza *machine learning* para ayudarles en la gestión y control de gastos. En

[28] MAYSON, S.G., «Bias in, Bias out»…*cit.*

[29] PERC, M., OZER, M. y HOJNIK, J., «Social and juristic challenges of artificial intelligence», *Palgrave Communications*, 5:61, 2019, pp. 1-7. DOI: 10.1057/s41599-019-0278-x.

[30] Carta de los Derechos Fundamentales de la Unión Europea, 2000, C 364/01, disponible en <https://www.europarl.europa.eu/charter/pdf/text_es.pdf>. Fecha de consulta: 18 de mayo de 2020.

[31] APPLE, «Apple Card launches today for all US customers», *Apple Newsroom*, 2019, disponible en <https://www.apple.com/newsroom/2019/08/apple-card-launches-today-for-all-us-customers/>. Fecha de consulta: 6 de abril de 2020.

noviembre de 2019 un usuario manifestó públicamente a través de Twitter que el límite de crédito para su mujer era veinte veces inferior al suyo, estando ambos casados y habiendo proporcionado la misma información con respecto a sus activos, e incluso presentando la mujer una mejor puntuación crediticia. Al ser compartidas situaciones similares por otros clientes, el Departamento de Servicios Financieros del Estado de Nueva York inició una investigación todavía en curso para determinar si los criterios utilizados por la tarjeta Apple podían ser considerados discriminatorios[32].

En un comunicado a CNN Business[33], Goldman Sachs explicó que los clientes de la tarjeta Apple no comparten una línea de crédito bajo la cuenta de un familiar u otra persona al obtener una tarjeta complementaria. Las solicitudes se evalúan de forma independiente, teniendo en cuenta los ingresos del individuo y el nivel de solvencia, considerando para ello la puntuación crediticia, el volumen de deuda y cómo se ha gestionado esa deuda. Esto podría justificar que se asignen a dos miembros de la misma familia límites de crédito significativamente diferentes.

Con independencia de los resultados que arroje la investigación, cabe la posibilidad de que éste sea un ejemplo ilustrativo de parcialidad con respecto a la variable género por inferencia estadística. Sin un fundamento sólido que justifique las decisiones del sistema autónomo de toma de decisiones que determina el límite de crédito de los usuarios, podría considerarse que este sistema de IA actúa de forma parcial o discriminatoria.

Por otro lado, el derecho de los consumidores a la igualdad de trato podría verse cuestionado por la nueva estrategia de precios de Uber, que está estudiando la posibilidad de fijar el precio del trayecto en función de la predisposición al gasto del usuario o de su comportamiento[34]. En un principio, Uber establecía sus precios a través de un mecanismo de oferta y demanda, siendo el precio igual para todos los usuarios del servicio que solicitaban el mismo trayecto. Implementando este mecanismo alternativo, los precios podrían incrementarse para los consumidores que tengan una capacidad económica

[32] VIGDOR, N., «Apple Card Investigated After Gender Discrimination Complaints», *The New York Times Business*, 2019, disponible en <https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html>. Fecha de consulta: 6 de abril de 2020.

[33] NEDLUND, E., «Apple Card is accused of gender bias. Here's how that can happen», *CNN Business*, 2019, disponible en <https://edition.cnn.com/2019/11/12/business/apple-card-gender-bias/index.html>. Fecha de consulta: 6 de abril de 2020.

[34] NEWCOMER, E., *Uber Yield Management: Uber Starts Charging What It Thinks You're Willing to Pay*, Bloomberg, 2017, disponible en <https://www.bloomberg.com/news/articles/2017-05-19/uber-s-future-may-rely-on-predicting-how-much-you-re-willing-to-pay>. Fecha de consulta: 20 de mayo de 2020.

superior, por ejemplo, haciendo que los trayectos reiterados desde las zonas residenciales más ricas sean más caros. Otro de los aspectos que pueden considerarse en la determinación del precio es la vida restante de la batería, de manera que el precio sea más elevado si el teléfono va a apagarse en poco tiempo[35].

Ambos casos han tenido por el momento una mayor incidencia en Estados Unidos, pero se trata de multinacionales que operan en todo el mundo y que también ponen en riesgo los derechos de los ciudadanos de la Unión Europea. Es necesario un control constante, si bien no excesivo, para evitar que se produzcan situaciones de amenaza a los principios éticos básicos que están establecidos por la UE. La transparencia y explicabilidad del algoritmo deviene en este tipo de supuestos necesaria a fin de proteger a los consumidores y salvaguardar sus derechos.

3.5. Responsabilidad por daños y perjuicios causados por la IA

El régimen de responsabilidad aplicable a la inteligencia artificial es uno de los puntos más controvertidos. Debe reflexionarse acerca de quiénes serán los sujetos sobre los que puede recaer la responsabilidad, qué tipo de responsabilidad en cada caso y en qué medida existe responsabilidad subsidiaria. Además, una de las novedades que deriva de las características particulares de la IA es la inversión de la carga de la prueba.

El Grupo Experto en Responsabilidad y Nuevas Tecnologías de la Comisión Europea publicó en 2019 un informe sobre Responsabilidad para la IA y otras tecnologías Emergentes[36], en el que se recoge la distinción entre dos tipos de operadores a la hora de asignar la responsabilidad: el operador frontal y el operador de *backend*. El operador frontal representa a aquellos que «principalmente» deciden y se benefician del uso de la tecnología, mientras que el operador de *backend* se refiere a quienes «continuamente» definen las características de «la tecnología relevante» y proporcionan soporte de *backend* esencial y continuo. Esta última categoría podría incluir a los fabricantes que ofrecen actualizaciones continuas de software y servicios de *backend,* entendiendo los mismos como la instrumentación de recursos informáticos entre el marco de *deep learning* y la infraestructura de la nube con la que trabaja la IA. Este informe apunta, además, que

[35] MARTIN, N., *Uber Charges More If They Think You're Willing To Pay More*, Forbes, 2019, disponible en <https://www.forbes.com/sites/nicolemartin1/2019/03/30/uber-charges-more-if-they-think-youre-willing-to-pay-more/#1825e1747365>. Fecha de consulta: 21 de mayo de 2020.

[36] GRUPO EXPERTO EN RESPONSABILIDAD Y NUEVAS TECNOLOGÍAS DE LA COMISIÓN EUROPEA, *Liability for Artificial Intelligence and other emerging digital technologies*, 2019, pp. 1-70. DOI:10.2838/573689.

cuando dos operadores coexisten, la responsabilidad estricta «debe recaer en el que tiene más control sobre los riesgos de operación».

En cuanto a la responsabilidad del fabricante o productor, se sugiere en este informe se impute estrictamente la responsabilidad por los defectos en productos o contenido digital que incorporen tecnología digital emergente –entre los cuales se incluirían aquellos que funcionen con IA–, incluso si «el defecto apareció después de que el producto se puso en circulación». Se considera que el riesgo soportado en la fase de desarrollo no debería recaer sobre los fabricantes en los casos en que fuese predecible que pudieran ocurrir desarrollos o comportamientos imprevistos.

Debido a la dificultad para las víctimas a la hora de demostrar la existencia de dichos defectos o la determinación de la responsabilidad por los mismos, el Grupo de Expertos propone invertir la carga de la prueba, de manera que corresponda al fabricante probar que el defecto no existía, no era predecible o se encuentra fuera de su ámbito de responsabilidad. Estas nuevas tecnologías son en esencia complejas, dinámicas e interconectadas, lo que puede conllevar que resulte especialmente difícil establecer un nexo de causalidad entre el origen del defecto y el daño ocasionado. Para proteger al usuario en este tipo de situaciones, se considera proporcional esa inversión probatoria, de forma que sea el operador o la persona a la que se dirija la reclamación por el defecto quien demuestre que no existía la deficiencia o el nexo causal que permite que le sea imputada la responsabilidad por el mismo.

Además de cuestionar la responsabilidad objetiva por productos con IA defectuosos, también debe plantearse la conveniencia de una posible alteración normativa con respecto a la responsabilidad subsidiaria. En concreto, se suscita la posibilidad de expandir el régimen de responsabilidad indirecta a los daños causados por la tecnología autónoma para los casos en que existe equivalencia funcional. Dicha equivalencia responde a una situación en la que el uso de una tecnología autónoma que causa daños pueda dar lugar a la responsabilidad indirecta del operador, si su uso es equivalente al empleo de auxiliares humanos, es decir, que la tarea ejecutada por una IA pudiese haber sido realizada por un ser humano.

Las anteriores cuestiones acerca de los riesgos sobre la responsabilidad serán tratadas en mayor profundidad junto a algunas aportaciones personales en el apartado donde se revisa

el régimen propuesto en el proyecto del Reglamento sobre Responsabilidad por el funcionamiento de los sistemas de Inteligencia Artificial[37].

4. Regulación de la inteligencia artificial

La implementación de sistemas de IA se ha extendido a campos tan diversos como el diagnóstico de enfermedades, la predicción de delitos y la evaluación de seguros[38]. Considerando los riesgos previamente descritos, y conforme a lo apuntado por la Comisión Europea, existe una necesidad imperante de examinar si la legislación actual es suficiente para dar respuesta a los riesgos de la IA y si puede ser efectivamente aplicada. Si no lo es, debe considerarse si procede hacer adaptaciones de la actual legislación o, si esto fuese insuficiente, elaborar nuevas normas.

En la línea de lo anterior, es razonable la conclusión que se alcanza ante la incertidumbre causada por la ausencia de un marco regulatorio claro. Si no se proporciona un enfoque legislativo a escala de la UE, existe un riesgo real de fragmentación en el mercado interior, lo que socavaría los objetivos de confianza, seguridad jurídica y aceptación del mercado.

Para armonizar la legislación de los Estados Miembros y anticipar una respuesta para las situaciones de riesgo o amenaza causadas por sistemas de IA, es necesario que se lleve a cabo una adaptación de la legislación existente o la elaboración de nuevas normas en el ámbito europeo. Más allá de las Directrices Éticas para una IA fiable[39], deben concretarse los riesgos que existen en las distintas áreas del Derecho y proponerse medidas concretas para dar una solución eficaz a los mismos. Esto proporciona un punto de partida para los diferentes Estados, que deben regular la IA con sus propias normas, pero ajustándose a los principios establecidos por la UE.

4.1. Posibles adaptaciones del marco legislativo europeo a la IA

4.1.1. Derecho de Propiedad Intelectual e Industrial

La inteligencia artificial puede llevar a cabo creaciones literarias y artísticas, como pueden ser canciones, cuadros o poemas. En Europa, mientras que algunos países han

[37] COMISIÓN DE ASUNTOS JURÍDICOS DEL PARLAMENTO EUROPEO, *Proyecto de Informe con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial*…cit.

[38] TAN, Z., YEOM, S., FREDRIKSON, M. y TALWALKAR, A., *Learning Fair Representations for Kernel Models,*ArXiv preprint, arXiv:1906.11813, 2019, pp. 1-15.

[39] GRUPO DE EXPERTOS DE ALTO NIVEL EN IA DE LA COMISIÓN EUROPEA (b), *Ethics Guidelines for Trustworthy AI*…cit.

optado por la protección de las obras creadas por una tecnología, otros se posicionan en contra de atribuir una autoría a este tipo de trabajos, considerando que dichas obras deben pertenecer a dominio público.

La Ley de Derechos de Autor de Reino Unido [40] da una definición de obras generadas por ordenador y establece su autoría, los derechos morales que corresponden a dichas obras y el periodo de duración de dicha protección. Se entiende como «trabajo generado por un ordenador» aquel que es creado por un ordenador sin la intervención de un autor humano y se considera que el autor es la persona por quien se toman las medidas necesarias para la creación de la obra. En la misma línea, la Ley de Derechos de Autor de Irlanda[41] establece un periodo de duración del copyright para obras generadas por ordenador de setenta años. Sin embargo, en la mayoría de los países europeos se tiende a no reconocer derechos de autor sobre las obras generadas por ordenador.

Es momento de establecer de forma clara cuál es el criterio que la UE comparte, para que así los Estados miembros adopten una postura de aceptación o rechazo sobre la existencia de una autoría para este tipo de trabajos. En el caso de reconocer a los trabajos creados por IA derechos de autor, debe determinarse a quién corresponden los mismos, mientras que, de rechazarse esta idea, habría que clasificar estas creaciones como obras de dominio público.

Los escritores y periodistas ahora pueden utilizar software de procesamiento de textos para escribir artículos periodísticos y libros. Los diseñadores gráficos pueden crear carteles y cuadros. Los compositores pueden ajustar su música y crear nuevas canciones. La IA, al igual que el ordenador, pero de forma más sofisticada y desarrollada, es una herramienta que sirve de ayuda o asistencia para los autores humanos. La pregunta que debe plantearse llegado este punto no es si debe reconocerse o no derechos de autor sobre las obras creadas por la IA, sino a quién corresponden esos derechos.

La IA no tiene consciencia de sí misma ni de su obra, ni personalidad jurídica a la que pueda ser atribuida una autoría. El autor en este tipo de supuestos debe ser quien ha diseñado la tecnología que finalmente ha dado lugar a la obra, como en el caso de que

---

[40] *Copyright, Designs and Patents Act*, Reino Unido, 1988, disponible en <http://www.legislation.gov.uk/ukpga/1988/48/part/I/chapter/I/crossheading/authorship-and-ownership-of-copyright>.

[41] *Copyright and Related Rights Act*, Irlanda, 2000, disponible en <http://www.irishstatutebook.ie/eli/2000/act/28/enacted/en/print>.

una persona desarrolle y entrene una red neuronal que cree cuadros o canciones. Podría también plantearse el escenario en el que se adjudiquen los derechos de autor al propietario de la tecnología, por ejemplo, en el caso de una editorial que haya adquirido una IA y la utilice para redactar libros o artículos.

Buena parte de los expertos en inteligencia artificial[42] consideran que la negativa al otorgamiento de derechos de autor sobre las obras generadas por ordenador supondría un desincentivo económico, puesto que, al pertenecer la obra creada al dominio público, el diseñador de la IA que ha creado la obra no podría lucrarse. Uno de los objetivos planteados por la Comisión Europea responde al impulso del desarrollo de la inteligencia artificial, para lo cual se necesita dotar de incentivos a quienes la diseñan. En este sentido, la desprotección de las obras literarias y artísticas supondría posiblemente un freno al progreso y avance tecnológicos a causa de la falta de incentivos económicos.

Es por ello por lo que el debate en cuanto a los derechos de propiedad intelectual e industrial debe abordarse desde la perspectiva de la tecnología como herramienta utilizada por el ser humano, existiendo una autoría detrás de la obra creada por IA. La tarea pendiente es revisar el concepto de autor, dar cabida a la existencia de obras generadas por tecnologías –dentro de las cuales se incluye la IA–, establecer el periodo de duración de la protección y examinar los posibles escenarios para determinar a quién corresponde la autoría en función de las circunstancias concretas.

4.1.2. Derecho de Transporte Terrestre

Las tecnologías de IA integradas en productos o aplicadas a prestaciones de servicios pueden presentar nuevos riesgos de seguridad para los usuarios en el área del transporte terrestre. Un error en el reconocimiento de objetos por parte de un automóvil autónomo puede causar un accidente con lesiones físicas y daños materiales. Estos riesgos pueden ser causados por defectos en el diseño de la IA, por una insuficiente disponibilidad o calidad de los datos o debido a otros problemas derivados del aprendizaje automático.

Dos convenios internacionales de tráfico por carretera, a saber, el Convenio de Ginebra sobre el tráfico por carretera de 1949[43] y el Convenio de Viena sobre el tráfico por

[42] HRISTOV, K., *Artificial Intelligence and the Copyright Dilemma*. The Journal of the Franklin Pierce Center for Intellectual Property, Vol. 57, 3 (438), disponible en <https://law.unh.edu/about/unh-law-publications/idea-journal-franklin-pierce-center-intellectual-property>. Fecha de consulta: 15 de mayo de 2020.

[43] Convención de Ginebra, de 19 de septiembre de 1949, sobre la circulación vial.

carretera de 1968[44], son la base sobre la que se asientan la mayor parte de las leyes nacionales de tráfico de los Estados miembros. Las normas de tráfico que se recogen en estas convenciones fueron elaboradas bajo el paradigma de conducción de un vehículo por una persona humana. La aparición de vehículos de conducción autónoma que pueden operar sin interferencia humana actualmente no es compatible con este planteamiento. Por lo tanto, un vehículo totalmente automatizado, en el sentido de la Convención de Ginebra y la Convención de Viena, no tiene conductor.

Para dar solución a esta situación, debe revisarse el planteamiento de la normativa de tráfico terrestre, dando cabida a los supuestos de vehículos autónomos, en los que no hay una persona que intervenga en la conducción. El sistema de IA dirige la conducción del vehículo, por lo que la figura del conductor que se recoge en la Convención de Ginebra y la Convención de Viena queda superada. La responsabilidad por la conducta del vehículo recae en su caso sobre el sistema de IA instalado en el automóvil, pero al carecer de personalidad jurídica esto crea una laguna.

Una posibilidad para colmar la misma es desviar la responsabilidad hacia la parte que tiene más influencia sobre la conducta que ha generado el sistema de IA. Dependiendo de las circunstancias, el responsable podría ser el fabricante del vehículo, que es quien decide sobre el hardware y el software del mismo; el diseñador de la IA, que establece los objetivos y entrena las redes neuronales; el proveedor de los datos que sirven de base a ese entrenamiento o incluso la persona que haya conseguido hackear el sistema de IA y controlarlo a distancia. El propietario también podría ser responsable, por ejemplo, si la conducta ha tenido lugar como consecuencia de que se hayan ignorado las instrucciones para instalar una actualización de software o, en el caso de vehículos semiautónomos, se hayan desoído las advertencias del sistema en las que se requiere su intervención en la conducción.

Debe, por tanto, revisarse la legislación de la Unión Europea en materia de tráfico terrestre, integrando el concepto de vehículo autónomo y semiautónomo en la normativa vigente. Es posible que surjan nuevas obligaciones y derechos para el fabricante del transporte, el conductor y el pasajero. Asimismo, el régimen de responsabilidad en el caso

[44] Convención de Viena, de 8 de noviembre de 1968, sobre la circulación vial.

de accidentes puede verse alterado por esta inclusión, de manera que también debe adaptarse la respuesta que se ofrece en este tipo de supuestos.

4.1.3. Derecho de Protección de Datos

El Reglamento General de Protección de Datos[45] no aborda específicamente la IA. A pesar de que se han tenido en cuenta los avances y los riesgos de los entornos digitales en la regulación de la protección de datos, la Unión Europea ha optado por lo que podría denominarse como «legislación independiente de la tecnología»[46]. Las reglas y principios del RGPD son suficientemente flexibles para cubrir futuros cambios tecnológicos y perdurar en el tiempo. No obstante, la consecuencia de un exceso de generalidad que procure la inclusión de nuevos avances puede dar lugar a grandes divergencias en la interpretación de la ley y, en consecuencia, generar cierta incertidumbre jurídica.

El artículo 4 del RGPD incluye una serie de definiciones como «datos personales», «elaboración de perfiles» o «tratamiento» que, a priori, son válidas y aplicables a la IA. Asimismo, lo previsto en cuanto a las obligaciones de los implementadores de sistemas de IA coincide con lo establecido en la disposición 24 del Reglamento. Se sugiere que, independientemente de la ubicación del establecimiento y de la pertenencia a la UE, todo actor esté sujeto a la legislación de protección de datos si se realiza oferta de bienes o servicios a los interesados en el territorio de la Unión.

Sin embargo, el principio de limitación de la finalidad y el principio de minimización de datos, señalados ambos en el artículo 5.1 apartados b) y c), respectivamente, podría ser demasiado acotado teniendo en cuenta las capacidades de procesamiento de IA. El uso de algoritmos y la utilidad del aprendizaje automático se basa en la tendencia a recopilar la mayor cantidad de datos posible y la generación de nuevos datos. De hecho, la reutilización de la información es una característica principal de las aplicaciones de IA en relación con el análisis de *big data*. Aunque a priori podría entenderse que estos principios actúan como una restricción o impedimento para el desarrollo de los sistemas de IA por

[45] Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE.

[46] MITROU, L., *Data Protection, Artificial Intelligence and Cognitive Services is the General Data Protection Regulation (GDPR) "Artificial Intelligence-Proof"?*, 2019, pp. 1-90, disponible en <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE2PdYu>. Fecha de consulta: 19 de mayo de 2020.

la prohibición de reutilización de datos y limitación de su uso, una interpretación adecuada de este artículo puede solventar este inconveniente[47].

Un desarrollador o implementador de IA está sujeto a los principios de protección de datos que se recogen en el RGPD, debiendo respetar todos los preceptos de dicha norma que puedan aplicarse a su situación. Para contribuir a esta adaptación normativa, dos son los retos que los desarrolladores e implementadores de la IA deben afrontar. Por una parte, deben definir de forma suficiente los fines perseguidos en la utilización de los datos empleados en el entrenamiento del sistema de IA. Por otro lado, sería conveniente que al recoger datos se cerciorasen de seleccionar cuáles son relevantes para el entrenamiento del sistema de IA, de manera que trate de respetarse el principio de minimización de datos descrito en el párrafo anterior.

Adoptando esta perspectiva, el Reglamento de Protección de Datos puede ser directamente aplicable a la IA, si bien es conveniente que se interprete parte de su contenido de forma más detallada o extensiva, como puede ser el caso de los principios referidos.

4.1.4. Derecho de Consumo y Responsabilidad por Productos Defectuosos

Las recomendaciones sobre responsabilidad por defectos en productos que incorporen IA que se recogen en el informe de la Comisión Europea[48] podrían implicar la modificación de la Directiva de Responsabilidad del Producto[49]. Teniendo en cuenta la participación de diferentes agentes en el diseño, desarrollo, utilización y aplicación de los sistemas de IA, debe ampliarse el concepto de operador que determina el ámbito objetivo de la norma. Para ello, deben considerarse no sólo las formas de inteligencia artificial actualmente existentes, sino que es preciso redefinir el concepto de operador, producto defectuoso y defecto, de forma que tengan cabida las nuevas tecnologías que surjan en los próximos años, sean o no consecuencia de una evolución de la IA.

---

[47] MITROU, L., *Data Protection, Artificial Intelligence and Cognitive Services is the General Data Protection Regulation (GDPR) "Artificial Intelligence-Proof"?*...cit.

[48] GRUPO EXPERTO EN RESPONSABILIDAD Y NUEVAS TECNOLOGÍAS DE LA COMISIÓN EUROPEA, *Liability for Artificial Intelligence…cit.*

[49] CONSEJO DE LAS COMUNIDADES EUROPEAS, *Directiva del Consejo, de 25 de julio de 1985, relativa a la aproximación de las disposiciones legales, reglamentarias y administrativas de los Estados miembros en materia de responsabilidad por los daños causados por productos defectuosos (85/374/CEE*), Diario Oficial de las Comunidades Europeas, 13(19) disponible en <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:31985L0374&from=EN>.

La aceptación de la clasificación de operadores sugerida en el informe sobre Responsabilidad por la IA[50] debe ir acompañada del establecimiento de obligaciones específicas para el operador frontal y el operador de backend, así como un marco de responsabilidad diferenciado para cada operador por el incumplimiento de lo que se haya previsto. A pesar de que la opacidad y la autonomía de los sistemas de IA puedan suponer un obstáculo o incluso imposibilitar en la práctica la prueba de la existencia o el origen del defecto, esta limitación puede superarse haciendo responsables a las personas que crean, mantienen o controlan los riesgos asociados al sistema de inteligencia artificial. De ahí la importancia de definir el término operador e incluir dentro de esta acepción no sólo al fabricante, sino también al proveedor de los datos, el desarrollador y el implementador –quien en última instancia utiliza la IA–, para el caso de que estos agentes sean diferentes personas.

El proveedor de datos es la persona que recoge y/o entrega o proporciona los datos a quien desarrollará la inteligencia artificial, pudiendo existir una diferenciación entre quien recoge en un principio los datos y quien los estructura posteriormente para su utilización.

Podría entenderse el desarrollador como la persona que diseña la inteligencia artificial usando los datos proporcionados por el proveedor para su entrenamiento y fijando los objetivos que condicionan la actuación de la IA y la creación del algoritmo.

El fabricante o productor puede coincidir con el desarrollador o no, según la función de la IA y su posible integración en un objeto o máquina. Por ejemplo, en el caso de los vehículos autónomos podría darse la situación de que el fabricante del automóvil externalizase la función de diseño del sistema de conducción autónoma y se encargarse únicamente del montaje del transporte integrando el sistema de IA en el mismo.

Finalmente, el implementador podría definirse como la persona que decide sobre el uso del sistema de IA, ejerce control sobre el riesgo y se beneficia de su funcionamiento; entendiendo por ejercicio del control cualquier acción del implementador que afecte al modo de funcionamiento de la IA o que altere funciones o procesos del sistema.

[50] GRUPO EXPERTO EN RESPONSABILIDAD Y NUEVAS TECNOLOGÍAS DE LA COMISIÓN EUROPEA, *Liability for Artificial Intelligence…cit.*

En la línea de lo previsto en la Directiva sobre Responsabilidad de Productos Defectuosos[51], la existencia de un defecto implica que el resultado del producto o servicio no es exactamente lo que se esperaba, sin que ello suponga necesariamente que el producto o servicio no funciona. Por producto o servicio defectuoso debe entenderse aquel que no ofrezca la seguridad que cabría legítimamente esperar teniendo en cuenta todas las circunstancias y, especialmente, el uso que razonablemente se prevé del mismo.

La responsabilidad debería abarcar todas las operaciones de los sistemas de IA, independientemente del lugar donde se realice la operación o que la misma sea física o virtual. De esta forma se evita que la localización de los operadores o la tangibilidad de sus acciones desprotejan a los usuarios que pueden verse perjudicados por los potenciales daños y perjuicios causados por la inteligencia artificial.

En el caso de que coexistan diferentes acciones o comportamientos entre los agentes responsables que hayan dado lugar al riesgo o defecto, si la responsabilidad es objetiva debería entenderse como solidaria, facilitando así la posibilidad del sujeto afectado de dirigirse contra uno, algunos o todos los agentes implicados. En el ámbito de la responsabilidad subsidiaria, sería necesario delimitar la extensión de dicha responsabilidad, tanto desde el punto de vista de los actores que pasarían a ser responsables subsidiarios, como para determinar los defectos o riesgos por los cuales debe surgir esta responsabilidad.

Por último, la inversión de la carga de la prueba habría de ser recogida como excepción a la regla general de prueba por parte del usuario que reclama la responsabilidad del operador por defectos en el producto o servicio que ha adquirido. El informe de la Comisión Europea[52] hace alusión a esta inversión de la carga de la prueba en los casos en los que pueda entenderse que existe presunción de culpa por parte del operador. Dicha presunción podría devenir del incumplimiento de las normas de registro o seguridad que sean establecidas para los sistemas de IA e impuestas a los operadores. Se sugiere, como una de las posibilidades, que las tecnologías de IA incorporen algún tipo de sistema de registro que permita identificar la fuente del mal funcionamiento que causó el daño. La

[51] CONSEJO DE LAS COMUNIDADES EUROPEAS, *Directiva del Consejo, de 25 de julio de 1985, relativa a la aproximación de las disposiciones legales, reglamentarias y administrativas de los Estados miembros en materia de responsabilidad por los daños causados por productos defectuosos…cit.*

[52] GRUPO EXPERTO EN RESPONSABILIDAD Y NUEVAS TECNOLOGÍAS DE LA COMISIÓN EUROPEA, *Liability for Artificial Intelligence…cit.*

inexistencia de este sistema de registro o seguridad podría constituir una presunción refutable de responsabilidad, pudiendo probarse la improcedencia de dicha acusación con la aportación de la información que explique el funcionamiento del sistema de IA o el origen del defecto.

Finalmente, la Directiva de Maquinaria[53] y la Directiva de Seguridad General del Producto[54] deben asimismo revisarse y actualizarse, de forma que ambas estén adaptadas a la IA y en concordancia con el resto de normativa europea que sea modificada, especialmente en lo referente al régimen de responsabilidad por daños y perjuicios.

4.1.5. Derecho de Seguros

Considerando la existencia de riesgos potenciales de los sistemas de IA y en previsión de minimizar los daños que pueden ocasionar los mismos a la víctima o al responsable por los mismos, debe estudiarse la posibilidad de exigir un seguro obligatorio de terceros para ciertas tecnologías emergentes, entre las cuales debe incluirse la IA. Una cobertura adecuada de los riesgos es además necesaria para garantizar la confianza de los ciudadanos en esta tecnología. Estos seguros podrían ayudar a las víctimas de manera más fácil o fructífera a reclamar una indemnización por daños, sin que se prejuzgue el derecho de la aseguradora a recurrir contra los responsables de actos de responsabilidad civil.

Una posibilidad es obligar a los implementadores de la IA a contar con un seguro para sus sistemas de alto riesgo, o quizá para todos ellos, aunque haciendo una diferenciación en la cobertura en función del riesgo que tenga el sistema. El mercado de los seguros podría adaptar los productos existentes o crear un nuevo seguro para los numerosos sectores y las diferentes tecnologías, productos y servicios que conllevan los sistemas de IA. Otra opción es que se establezcan fondos de compensación para las víctimas que no podrían reclamar efectivamente una indemnización debido a las dificultades para

[53] PARLAMENTO EUROPEO y CONSEJO DE LA UNIÓN EUROPEA, *Directiva del Parlamento Europeo y del Consejo, de 17 de mayo de 2006, relativa a la maquinaria (2006/42/EC*), Diario Oficial de la Unión Europea, L157/24, disponible en <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:157:0024:0086:EN:PDF>.

[54] PARLAMENTO EUROPEO y CONSEJO DE LA UNIÓN EUROPEA, *Directiva del Parlamento Europeo y del Consejo, de 3 de diciembre de 2001, relativa a la seguridad general de los productos (2001/95/EC*), Diario Oficial de la Unión Europea, L11/4, disponible en <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32001L0095&from=EN>.

identificar al responsable por los daños ocasionados o en el caso de que la tecnología no esté asegurada.

4.2. Elaboración de nuevas normas sobre IA

4.2.1. Directrices Éticas para una IA fiable

El 8 de abril de 2019 la Comisión publicó las Directrices Éticas para una IA fiable[55], primer instrumento orientativo europeo acerca de los pasos a seguir en los conflictos surgidos por la utilización de esta nueva tecnología. El objetivo de este instrumento pasa por maximizar los beneficios y minimizar los riesgos que surgen por la adopción de la inteligencia artificial. Para cumplir esta misión, se considera que el punto de partida es promover la fiabilidad de la IA en el sentido de extender la confianza en la misma, lo cual se pretende consiguiendo que esta tecnología sea lícita, ética y robusta, técnica y socialmente. Se trata de que estos tres componentes sean considerados de forma conjunta y simultánea, estableciéndose un marco de fiabilidad para la protección de los Derechos Fundamentales recogidos en la Carta de la UE.

Las Directrices Éticas superan el plano abstracto de la enunciación de principios generales, haciendo referencia a medidas concretas que deberían ser objeto de un estudio pormenorizado. No obstante, más allá de los principios éticos que recoge es necesaria una mayor profundización y concreción en algunos aspectos, especialmente respecto a las medidas que pueden adoptarse para mitigar los riesgos de la IA.

4.2.2. Reglamento sobre Responsabilidad por los sistemas de IA

El 27 de abril de 2020 fue publicado el Proyecto de Informe con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial incluyendo la propuesta del Parlamento Europeo para el Reglamento sobre Responsabilidad por el funcionamiento de los sistemas de Inteligencia Artificial[56]. En este documento, el Parlamento Europeo apuesta por la creación de un marco jurídico horizontal basado en principios comunes con el fin de establecer una igualdad de normas en toda la Unión y proteger eficazmente nuestros valores europeos. Considera que las nuevas normas comunes para los sistemas de IA deben recogerse en un reglamento,

[55] GRUPO DE EXPERTOS DE ALTO NIVEL EN IA DE LA COMISIÓN EUROPEA (b), *Ethics Guidelines for Trustworthy AI*…cit.

[56] COMISIÓN DE ASUNTOS JURÍDICOS DEL PARLAMENTO EUROPEO, *Proyecto de Informe con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial*…cit.

haciendo especial alusión a la cuestión de la responsabilidad civil en caso de daño o perjuicio causado por la IA.

Teniendo en cuenta lo anterior, parece que la regulación de la IA se llevará a cabo no solamente a través de adaptaciones de normativas ya existentes, sino mediante la creación de nuevos instrumentos, como las Directrices Éticas o este futuro Reglamento sobre Responsabilidad por el funcionamiento de los sistemas de inteligencia artificial.

En materia de responsabilidad, la adaptación de la Directiva de Responsabilidad por Productos Defectuosos[57] podría llegar a resultar insuficiente. El Parlamento Europeo apunta en su Informe que, de conformidad con los sistemas de responsabilidad estricta de los Estados miembros, el Reglamento propuesto solo debe cubrir los daños que afecten a la vida, la salud, la integridad física y la propiedad, y debe establecer las cantidades y el alcance de la indemnización, así como el plazo de prescripción para las reclamar dichos daños. Esto permite dotar a los Estados miembros de un punto de partida común para los derechos fundamentales más importantes, dejando un margen de libertad con respecto a otros derechos, para que los países regulen las particularidades del régimen de responsabilidad de la forma que más se adecúe a su ordenamiento jurídico.

El proyecto del Reglamento comienza presentando en su artículo 3 ciertas definiciones. Sin embargo, hay ciertos conceptos que deberían ser incluidos y otros que conviene matizar. El sistema de IA de alto riesgo es una clasificación que necesita una descripción más detallada, prescindiendo de la mención sobre la aleatoriedad y atendiendo a los diferentes criterios que determinan si, efectivamente, el riesgo es elevado y el sistema requiere de mayor protección. Recapitulando lo expuesto anteriormente, un sistema de IA debería ser calificado de alto riesgo en función de tres criterios: la gravedad del posible perjuicio, la probabilidad de que el riesgo se materialice y el modo en que se utiliza el sistema.

Por otra parte, la existencia de un anexo donde se recojan los sistemas de IA de alto riesgo puede generar más incertidumbre que seguridad. Debe cuestionarse si resulta más eficaz a efectos de su aplicación en la práctica la creación de una lista cerrada, aunque revisable, de sistemas de riesgo o el examen caso por caso del sistema en función de unos criterios concretos. Si bien es cierto que una lista puede dar una sensación de mayor concreción,

[57] CONSEJO DE LAS COMUNIDADES EUROPEAS, *Directiva del Consejo, de 25 de julio de 1985, relativa a la aproximación de las disposiciones legales, reglamentarias y administrativas de los Estados miembros en materia de responsabilidad por los daños causados por productos defectuosos…cit.*

existe el riesgo de que determinados sistemas no se encuentren identificados en la misma y queden fuera del marco de protección, además de la posibilidad de que sean causados daños y no se aplique el régimen de responsabilidad o medidas para sistemas de alto riesgo hasta su inclusión. De aceptarse el método de enumeración en el anexo de los sistemas de alto riesgo, debe estudiarse el efecto retroactivo de las medidas de prevención, análisis y corrección, así como de la aplicación del régimen de responsabilidad.

En el anexo del Proyecto de Reglamento se recogen cinco sistemas de IA calificados de alto riesgo: aeronave no tripulada, vehículos con niveles de automatización 4 y 5, sistemas autónomos de gestión del tráfico, robots autónomos y dispositivos autónomos de limpieza de lugares públicos. Sin perjuicio de los posibles matices que puedan ser añadidos a estos sistemas, la categoría de robots autónomos presenta sin duda cierta problemática. Un robot autónomo es una máquina inteligente capaz de realizar tareas en el mundo por sí misma, sin control humano explícito[58]. Por tanto, cuando se habla de robots autónomos se hace referencia desde un robot que lleva a cabo operaciones quirúrgicas hasta un robot aspirador. Sin embargo, es evidente que el nivel de riesgo en estos dos ejemplos es muy diferente, tanto atendiendo a la magnitud de los daños que puede generar como al impacto de los mismos en la esfera personal y social. Si se mantiene la creación de este anexo, procede crear categorías más específicas, a salvo de que el objetivo en un principio sea clasificar por lo general los sistemas como de alto riesgo.

Por otra parte, existen sistemas de IA que no pertenecen a las anteriores categorías y que, sin embargo, pueden tener alto riesgo. En concreto, no se reflejan los sistemas de decisión autónomos más allá de los integrados en vehículos dedicados a la conducción autónoma. Sin embargo, pueden existir sistemas autónomos de toma de decisiones que tengan alto riesgo, como por ejemplo un sistema que determine la medicación o dieta que debe seguir un paciente. Si bien es cierto que, por el momento, estos sistemas todavía no están muy extendidos y requieren de mayor desarrollo, en un futuro cercano la IA se implementará de forma prolífera en sectores más allá de la conducción y el tráfico.

En cuanto a las definiciones que se refieren a los sujetos afectados por este Reglamento, es importante incluir a todos los actores que puedan ser potencialmente responsables de

[58] BEKEY, G.A., *Autonomous Robots: From Biological Inspiration to Implementation and Control*, 2017, disponible en <https://mitpress.mit.edu/books/autonomous-robots>. Fecha de consulta: 17 de mayo de 2020.

daños causados por la inteligencia artificial. No sólo el implementador puede ser responsable, sino también el productor, el desarrollador o el proveedor de los datos.

Se entiende que implementador es quien ejerce en mayor medida el control sobre los riesgos asociados y se beneficia de las operaciones del sistema de IA. Esto genera grandes dudas en cuanto a quién es el implementador, puesto que todos los agentes ejercen en parte control sobre los riesgos asociados, siendo mayor o menor en cada fase en función de las características del sistema. Para un sistema de IA supervisado, el mayor riesgo puede estar en la calidad de los datos que han entrenado el sistema, mientras que, en un sistema no supervisado, el riesgo puede venir dado de la exposición del sistema a nuevos datos que modifican su comportamiento. Por tanto, atendiendo al control del riesgo, el implementador podría ser en cada caso un sujeto diferente, en función del tipo de sistema de IA y la fase donde se ha originado el defecto. Por otra parte, se conjuga este elemento de control de riesgo con el criterio de beneficio obtenido por el uso del sistema. Con respecto a este segundo aspecto, parece que quien generalmente obtendría beneficios de su utilización es la persona que pone a disposición del usuario el producto o servicio en el cual se ha integrado la IA, lo cual también puede identificarse con el operador final.

La definición de productor tampoco es exacta, puesto que el productor puede ser quien desarrolla el sistema, pero también quien simplemente lo integra cuando un tercero lo ha diseñado. Asimismo, el productor no tiene por qué coincidir con el operador final, que será quien finalmente ponga a disposición del usuario el sistema de IA.

El ejemplo ya mencionado sobre los vehículos autónomos puede ilustrar con claridad lo anterior. El proveedor de datos es la persona que pondría a disposición los datos de entrenamiento de la IA. El desarrollador, una vez recibidos dichos datos, es el encargado de fijar los objetivos y añadir las restricciones al sistema de IA según proceda, llevar a cabo el entrenamiento de la IA minimizando el margen de error y recalibrando el sistema para reducir el nivel de incertidumbre. Una vez la IA funcione correctamente, podría ponerse a disposición del productor o fabricante del vehículo, quien ensambla las piezas y realiza el montaje del automóvil, integrando el sistema de IA cuya función es la conducción autónoma. Finalmente, este productor, que puede ser o no el operador final, venderá el vehículo ya terminado a los clientes finales.

Para que el régimen de responsabilidad funcione correctamente, es importante conocer las diferentes etapas del desarrollo de un sistema de IA y examinar cuáles son las tareas, obligaciones y riesgos concretos que recaen sobre cada uno de los actores. Por tanto, una

revisión de los sujetos definidos en este precepto ayudaría a una mejor comprensión de este proceso completo y aportaría una mayor claridad a la hora de determinar la responsabilidad. De las definiciones mencionadas parece extraerse que el implementador es quien oferta el producto o servicio con IA al usuario final, mientras que el productor sería, en su caso, el desarrollador.

En el artículo 4 se establece la responsabilidad objetiva del implementador, exonerando al mismo de responsabilidad solamente en el caso de que exista fuerza mayor. Teniendo en cuenta la naturaleza y funcionamiento de estos sistemas, debe concretarse qué se entiende por fuerza mayor en el caso de la IA. En ocasiones es evidente que ciertos daños causados por el sistema no podían haberse previsto, como en el caso de que un vehículo autónomo no reconozca un objeto o cuerpo cuya aparición sea nueva y del cual no se tenían datos hasta el momento. Sin embargo, otros escenarios pueden dar lugar a mayor confusión, por ejemplo, un comportamiento imprevisto cometido dentro del margen de error del sistema que había superado los requisitos de seguridad y calidad establecidos. La utilización de sistemas de IA implica que, dentro del funcionamiento normal, hay ciertas posibilidades de resultados erróneos o anormales. Debe determinarse si dichos resultados serán considerados como fuerza mayor o, por el contrario, dan lugar a una responsabilidad objetiva incluso cuando el agente responsable realizó su trabajo con la máxima diligencia debida.

Por su parte, el artículo 8 prevé la responsabilidad subjetiva del implementador para los sistemas de IA que no sean de alto riesgo. La responsabilidad objetiva para los sistemas de alto riesgo y subjetiva para los sistemas de bajo riesgo, dándose en ambos casos la inversión de la carga de la prueba, puede llegar a resultar excesivamente gravosa para el implementador. Es cierto que debe facilitarse para el usuario la posibilidad de reclamar los daños y defectos que hayan sido ocasionados y, en caso de que sea difícil probar el origen de los mismos, la inversión de la carga de la prueba puede ser conveniente. Sin embargo, los fallos del sistema de IA pueden darse en diversas etapas en las que el implementador no ha tenido intervención alguna, por lo que debe regularse el régimen de responsabilidad teniendo en cuenta todas las posibilidades.

Existiendo inversión de la carga de la prueba, la responsabilidad subjetiva para sistemas de IA tanto de alto riesgo como de bajo de riesgo no limita las posibilidades de accionar contra la persona afectada. Permite que dicha persona pueda dirigirse hacia el implementador y que, si este demuestra que el daño no le es imputable, se redirija la

reclamación contra el verdadero responsable. Esta solución facilita que el cliente final no se vea perjudicado y que el implementador no soporte una carga excesiva por esa responsabilidad objetiva. De lo contrario, podría darse una situación de desincentivo por temor a las graves consecuencias en caso de producirse daños, disminuyendo así la implementación y utilización de sistemas de IA. Esto frustraría en parte el objetivo de promover la adopción de la inteligencia artificial en la Unión Europea y conseguir una posición fuerte en este campo en el panorama internacional.

En la misma línea, el artículo 10 prevé como supuesto de negligencia concurrente únicamente el caso en el que el daño se haya debido a las actuaciones del implementador y la persona afectada en conjunto. Considerando el régimen de responsabilidad establecido para el implementador, una vía que ofrece una solución menos gravosa podría ser la previsión de negligencia concurrente con otros actores como el productor o el proveedor de datos en el caso de que sus actuaciones hayan sido el origen del daño o perjuicio. De esta forma se alivia la carga soportada por el implementador en los casos en los que su intervención, más allá de poner a disposición del usuario final el producto o servicio, no ha causado el daño. Esta opción permite que, sin desproteger al usuario, se dirima la responsabilidad de una forma más justa desde el inicio del procedimiento de reclamación, evitando la dilación que puede suponer un procedimiento posterior de repetición por parte del implementador contra el verdadero responsable.

Ese procedimiento de repetición o recurso de indemnización se contiene en el artículo 12 del Reglamento. Sin embargo, una revisión de dicha disposición podría resultar beneficiosa tanto para el implementador como para la persona afectada. La imposibilidad de iniciar la reclamación de indemnización por parte del implementador antes de finalizar íntegramente el pago a la persona perjudicada presenta desventajas para ambas partes.

Por un lado, el implementador se ve privado de la acción de repetición hasta que finalice un proceso en el que se le imputa responsabilidad objetiva sin opción de desviar incluso con prueba la responsabilidad hacia otra persona. Debe tenerse en cuenta, a estos efectos, que las indemnizaciones previstas en los artículos 5 y 6, a las que se hará mención más adelante, alcanzan cuantías considerablemente elevadas. Ello implica que el implementador, sin importar lo diligente que haya sido su actuación, debe responder por toda indemnización que se le reclame sin posibilidad de reclamar al responsable hasta el abono. Esto supone una carga excesiva y desproporcionada hacia esta figura, que incluso en el supuesto de contar con un seguro o un fondo para este tipo de contingencias, podría

tener dificultades para hacer frente al pago de todas las indemnizaciones que le sean reclamadas. Con esta observación no se trata de desviar o disminuir la responsabilidad del implementador, puesto que, en caso de ser responsable por el daño, respondería abonando la indemnización correspondiente. Simplemente se trata de facilitar el procedimiento de reclamación y dar la posibilidad de que el mismo sea más directo, de forma que el implementador, si consigue probar que la responsabilidad recae sobre otra persona, no deba responder por el daño y sea la persona responsable quien haga frente a esa indemnización.

Esta solución no beneficia únicamente al implementador, que puede liberarse de la responsabilidad objetiva si consigue probar la responsabilidad de otro actor, sino que facilita el cobro de la indemnización por la persona afectada. Esto es así en tanto en cuanto un cúmulo de reclamaciones hacia una única persona puede llegar a saturar o sobrecargar el sistema de compensaciones que se haya previsto, de forma que no sólo se perjudique la situación económica de un agente que quizá no sea responsable de daño alguno, sino que además frustre el cobro de parte de las indemnizaciones solicitadas.

La responsabilidad objetiva tiene sentido en el caso de que solo esa persona pueda ser responsable sobre la actuación del objeto o ser que causa el daño, como puede ocurrir con un animal. Sin embargo, la IA debe ser tratada como cualquier otro objeto en cuyo montaje o desarrollo intervienen varias personas. El daño puede haber sido causado por diferentes agentes y se debe dar la posibilidad a la persona afectada de reclamar individual o conjuntamente contra todos ellos, así como facilitar que la persona que ha puesto a disposición del perjudicado el bien o servicio pueda probar la ausencia de responsabilidad.

En el caso de que el usuario final sea un consumidor, es razonable que se siga un régimen de responsabilidad similar al establecido para productos defectuosos, de forma que se reclame contra el productor o fabricante el daño causado por el producto o servicio defectuoso. Sin embargo, el artículo 7 f) la Directiva de Responsabilidad por Productos Defectuosos[59] prevé la posibilidad de que el productor no sea responsable cuando el defecto proceda de una parte integrante del producto cuya fabricación corresponda a un

[59] CONSEJO DE LAS COMUNIDADES EUROPEAS, *Directiva del Consejo, de 25 de julio de 1985, relativa a la aproximación de las disposiciones legales, reglamentarias y administrativas de los Estados miembros en materia de responsabilidad por los daños causados por productos defectuosos…* cit.

tercero. Este sería el caso de un fabricante de vehículos autónomos que no ha diseñado el sistema de IA, pero lo ha adquirido de un tercero para integrarlo en el automóvil.

Existe cierto temor a que la persona afectada, en el caso de existir el régimen de responsabilidad descrito, se encuentre en una posición de vulnerabilidad en la cual no pueda reclamar a ninguna persona. Sin embargo, el único supuesto en el que no habría lugar a la indemnización es el de fuerza mayor. Existiendo la inversión de carga de la prueba, en caso de que sea difícil o imposible observar el funcionamiento de la IA por falta de trasparencia y, por tanto, determinar dónde se originó el daño y quién responde por el mismo, no habría exoneración o liberación de responsabilidad.

Una de las obligaciones establecidas para los desarrolladores de la IA debe ser la transparencia y explicabilidad del sistema, por lo que, en los casos en los que no se pueda determinar el origen del daño y, por tanto, no se pueda probar que no hubo responsabilidad, esta podría recaer sobre el desarrollador e incluso el implementador. El desarrollador debe asegurarse de que el sistema de IA que diseña y vende es suficientemente transparente y explicable, por lo que si no se puede determinar el origen del daño es responsable por haber faltado a dicha obligación.

La tarea del implementador que no participa en el desarrollo de la IA es entregar el producto terminado o servicio al usuario. En el proceso de compra de la inteligencia artificial e integración de la misma en el producto o servicio, tiene el deber de tratar de verificar por todos los medios a su alcance que el sistema de IA cumple con los requisitos legalmente establecidos. Por tanto, la responsabilidad del implementador tendrá lugar cuando incurra en dolo, culpa, falta de diligencia o negligencia al utilizar un sistema de IA que no sea suficientemente robusto, transparente o explicable en su producto o servicio. Sólo procedería la liberación de responsabilidad si, habiendo tomado todas las cautelas posibles y empleando toda la diligencia debida, no pudo conocer que el sistema de IA no cumplía con los estándares de calidad previstos.

Asimismo, posiblemente un mayor desarrollo del artículo 11 sobre la responsabilidad solidaria y conjunta facilitaría la tarea de determinar la responsabilidad entre los distintos actores desde un principio. Solamente se recogen dos supuestos: la responsabilidad de varios implementadores o la responsabilidad del implementador y el productor. Dado que ampliar la definición de implementador a otras figuras supondría una mayor incertidumbre a la hora de la aplicación práctica del Reglamento, posiblemente la opción

más factible sea añadir otros supuestos en los que existan varios responsables de forma solidaria y conjunta, como los expuestos previamente.

En el artículo 5 y 6 se establece el importe y el alcance de la indemnización que procede por los daños y perjuicios causados por sistemas de IA. De nuevo, esta disposición hace únicamente referencia al supuesto del implementador como responsable, no haciendo mención alguna acerca del proveedor de los datos, el desarrollador o el productor o fabricante. Dado que las cuantías tanto por daños materiales como personales son elevadas, debe revisarse lo señalado en cuanto al régimen de responsabilidad del implementador. Por otra parte, la concreción de la cuantía indemnizatoria parece quedar a discrecionalidad del juez o tribunal que conozca del procedimiento. Teniendo en cuenta el amplio rango de la indemnización, y para evitar que existan resoluciones contradictorias o muy diferentes, podría ser conveniente establecer ciertos criterios de moderación de la cuantía compensatoria. Esto no supone una limitación de la discrecionalidad del juez, sino una orientación para procurar que la indemnización se determine siguiendo unos criterios uniformes en todos los Estados miembros.

Finalmente, el plazo de prescripción especial recogido en el artículo 7 es de diez o treinta años, en función del supuesto. Este término debe ponerse en contexto con los plazos de prescripción generalmente establecidos para delitos que pueden causar daños similares. La prescripción para delitos como el homicidio o las lesiones se regula de forma considerablemente distinta en los países europeos, por lo que este plazo de prescripción especial podría llegar a resultar controvertido si carece de una justificación.

Dice la Exposición de Motivos del Reglamento que «cualquier marco en materia de responsabilidad civil orientado al futuro debe aspirar a lograr un equilibrio entre la protección eficaz de las potenciales víctimas de daños y, al mismo tiempo, ofrecer un margen de maniobra suficiente para posibilitar el desarrollo de nuevas tecnologías, productos o servicios». Sin embargo, el estricto régimen de responsabilidad que se hace recaer en exclusiva sobre el implementador puede suponer un gran desincentivo para las empresas dispuestas a apostar por la adopción de sistemas de IA. El objetivo de protección de las personas afectadas puede perseguirse desde una óptica más amplia, donde se de cabida a otros actores potencialmente responsables. Pueden ofrecerse las mismas o incluso más posibilidades de reclamación al usuario perjudicado haciendo la situación del implementador menos gravosa, para los casos en los que la responsabilidad no recaiga sobre su actuación, sino sobre otros sujetos.

5. Conclusión

La adaptación de la legislación vigente y la elaboración de nueva normativa sobre inteligencia artificial es un proceso complejo que requiere de un conocimiento de la materia en profundidad y un diseño de medidas de previsión, control, corrección y responsabilidad adecuadas a la naturaleza y funcionamiento de esta tecnología. Dada la dificultad que puede entrañar anticipar los riesgos de la IA en todos sus ámbitos de aplicación, las recomendaciones y sugerencias planteadas desde diferentes puntos de vista y disciplinas científicas es posiblemente la vía que permite una mejor comprensión de los problemas existentes.

La Comisión Europea ha comenzado un gran trabajo en el camino hacia la regulación de la IA, optando por la adaptación de la normativa existente en la mayor parte de los ámbitos, pero dando paso a la creación de un nuevo Reglamento que recoja el régimen de Responsabilidad por los daños causados por la inteligencia artificial. Si bien esta perspectiva puede ser acertada, algunos de los aspectos que responden a estos cambios o novedades pueden resultar controvertidos, por lo que conviene que se lleve a cabo una revisión o matización de los mismos.

Sin perjuicio de las recomendaciones y medidas propuestas, el planteamiento general de la responsabilidad en materia de IA es especialmente relevante. Encontrar el equilibrio entre la seguridad del marco de responsabilidad y la protección del desarrollo tecnológico y la innovación es un doble objetivo fundamental para la Unión Europea. En este sentido, debe ponerse especial cuidado en la determinación de los posibles sujetos responsables, el carácter de la responsabilidad que se pretende atribuir y los supuestos en los que la misma procede. Un marco de responsabilidad demasiado amplio o restringido puede ocasionar que uno de los objetivos mencionados pueda verse perjudicado, con el impacto negativo que ello conlleva para la protección de los usuarios de la IA, en el caso del primero, o para el papel de la UE como impulsor de la adopción de esta tecnología, si se atiende al segundo.

Esta propuesta no es una guía de la futura regulación de la IA, pero sí ha sido elaborada con el objetivo de apoyar ese esfuerzo legislativo. El debate es el motor del avance, por lo que las cuestiones que en este documento se tratan deben ser interpretadas como puntos controvertidos y aportaciones constructivas. La inteligencia artificial y, por extensión, su regulación, nos afecta a todos. Es momento de colaborar para intentar que, teniendo en

cuenta todas las contribuciones posibles, la normativa de IA permita cumplir ese doble objetivo de la Unión Europea: proporcionar seguridad, estimulando el progreso.

6. Bibliografía

ANTORÁN CABISCOL, J., *Understanding Uncertainty in Bayesian Neural Networks*, Departamento de Ingeniería de la Universidad de Cambridge, 2019, *Pro manuscript*, pp. 1-94.

APPLE, «Apple Card launches today for all US customers», *Apple Newsroom*, 2019, disponible en <https://www.apple.com/newsroom/2019/08/apple-card-launches-today-for-all-us-customers/>. Fecha de consulta: 6 de abril de 2020.

BEKEY, G.A., *Autonomous Robots: From Biological Inspiration to Implementation and Control*, 2017, disponible en <https://mitpress.mit.edu/books/autonomous-robots>. Fecha de consulta: 17 de mayo de 2020.

COECKELBERGH, M., «Ethics of artificial intelligence: Some ethical issues and regulatory challenges», *Technology and Regulation*, 2019, pp. 31-34. DOI: 10.26116/techreg.2019.003.

COLE, G.W. y WILLIAMSON, S.A., *Avoiding Resentment Via Monotonic Fairness*, ArXiv preprint, arXiv:1909.01251v1, 2019, pp. 1-16.

DOSHI-VELEZ, F., KORTZ, M., BUDISH, R., BAVITZ, C., GERSHMAN, S., O'BRIEN, D., SCHIEBER, S., WALDO, J., WEINBERGER, D. y WOOD. A., *Accountability of AI under the law: The role of explanation.* ArXiv preprint, arXiv:1711.01134, 2017, pp. 1-15.

ELDRED, C., ZYSMAN, J. y NITZBERG, M., *AI and Domain Knowledge: Implications of the Limits of Statistical Inference*, BRIE / WITS Technology Briefing, Berkeley, 2019, pp. 1-11, disponible en <https://ssrn.com/abstract=3479479>. Fecha de consulta: 18 de mayo de 2020.

HRISTOV, K., *Artificial Intelligence and the Copyright Dilemma*. The Journal of the Franklin Pierce Center for Intellectual Property, Vol. 57, 3 (438), disponible en <https://law.unh.edu/about/unh-law-publications/idea-journal-franklin-pierce-center-intellectual-property>. Fecha de consulta: 15 de mayo de 2020.

KLEINBERG, J., LUDWIG, J., MULLAINATHAN, S. y SUNSTEIN, C.S., «Discrimination in the Age of Algorithms», *Journal of Legal Analysis*, Vol. 10, 2018, pp. 113-174. DOI: 10.1093/jla/laz001.

KUSNE, M., LOFTUS, J., RUSSELL, C. y SILVA, R., *Counterfactual Fairness*, ArXiv preprint, arXiv:1703.06856v3, 8 de marzo de 2018, pp. 1-18.

MARTIN, N., *Uber Charges More If They Think You're Willing To Pay More*, Forbes, 2019, disponible en <https://www.forbes.com/sites/nicolemartin1/2019/03/30/uber-charges-more-if-they-think-youre-willing-to-pay-more/#1825e1747365>. Fecha de consulta: 21 de mayo de 2020.

MAYSON, S.G., «Bias in, Bias out», *Yale Law Journal* 128, 2019, pp. 1-84, ¿disponible en <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3257004>. Fecha de consulta: 3 de abril de 2020.

MITROU, L., *Data Protection, Artificial Intelligence and Cognitive Services is the General Data Protection Regulation (GDPR) "Artificial Intelligence-Proof"?*, 2019, pp. 1-90, disponible en <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE2PdYu>. Fecha de consulta: 19 de mayo de 2020.

NEDLUND, E., «Apple Card is accused of gender bias. Here's how that can happen», *CNN Business*, 2019, disponible en <https://edition.cnn.com/2019/11/12/business/apple-card-gender-bias/index.html>. Fecha de consulta: 6 de abril de 2020.

NELSON, G.S., «Bias in artificial intelligence», *North Carolina Medical Journal*, Vol. 80(4), 2019, pp. 220-222. DOI: 10.18043/ncm.80.4.220.

NEWCOMER, E., *Uber Yield Management: Uber Starts Charging What It Thinks You're Willing to Pay*, Bloomberg, 2017, disponible en <https://www.bloomberg.com/news/articles/2017-05-19/uber-s-future-may-rely-on-predicting-how-much-you-re-willing-to-pay>. Fecha de consulta: 20 de mayo de 2020.

PERC, M., OZER, M. y HOJNIK, J., «Social and juristic challenges of artificial intelligence», *Palgrave Communications*, 5:61, 2019, pp. 1-7. DOI: 10.1057/s41599-019-0278-x.

TAN, Z., YEOM, S., FREDRIKSON, M. y TALWALKAR, A., *Learning Fair Representations for Kernel Models,*ArXiv preprint, arXiv:1906.11813, 2019, pp. 1-15.

VAN GERVEN, M. y BOHTE. S., *Artificial Neural Networks as Models of Neural Information Processing*, Frontiers in Computational Neuroscience 11:114, 2017, pp. 1-2. DOI: 10.3389/fncom.2017.00114.

VIGDOR, N., «Apple Card Investigated After Gender Discrimination Complaints», *The New York Times Business*, 2019, disponible en <https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html>. Fecha de consulta: 6 de abril de 2020.

YEOM, S., DATTA, A. y FREDRIKSON, M., «Hunting for discriminatory proxies in linear regression models», *Advances in Neural Information Processing Systems*, 2018, pp. 4568-4578, disponible en <http://papers.nips.cc/paper/7708-hunting-for-discriminatory-proxies-in-linear-regression-models.pdf>. Fecha de consulta: 14 de abril de 2020.

7. Documentación

Carta de los Derechos Fundamentales de la Unión Europea, 2000, C 364/01, disponible en <https://www.europarl.europa.eu/charter/pdf/text_es.pdf>. Fecha de consulta: 18 de mayo de 2020.

COMISIÓN DE ASUNTOS JURÍDICOS DEL PARLAMENTO EUROPEO, *Proyecto de Informe con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial* (2020/2014 (INL)), disponible en <https://www.europarl.europa.eu/doceo/document/JURI-PR-650556_ES.pdf>.

COMISIÓN EUROPEA, *White Paper on Artificial Intelligence: a European approach to excellence and trust*, 19 de febrero de 2020, disponible en <https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en>. Fecha de consulta: 21 de marzo de 2020.

CONSEJO DE LAS COMUNIDADES EUROPEAS, *Directiva del Consejo, de 25 de julio de 1985, relativa a la aproximación de las disposiciones legales, reglamentarias y administrativas de los Estados miembros en materia de responsabilidad por los daños causados por productos defectuosos (85/374/CEE*), Diario Oficial de las Comunidades Europeas, 13(19) disponible en <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:31985L0374&from=EN>.

Convención de Ginebra, de 19 de septiembre de 1949, sobre la circulación vial.

Convención de Viena, de 8 de noviembre de 1968, sobre la circulación vial.

Copyright, Designs and Patents Act, Reino Unido, 1988, disponible en <http://www.legislation.gov.uk/ukpga/1988/48/part/I/chapter/I/crossheading/authorship-and-ownership-of-copyright>.

Copyright and Related Rights Act, Irlanda, 2000, disponible en <http://www.irishstatutebook.ie/eli/2000/act/28/enacted/en/print>.

GRUPO DE EXPERTOS DE ALTO NIVEL EN IA DE LA COMISIÓN EUROPEA (a), *A Definition Of AI: Main Capabilities And Disciplines*, 8 de abril de 2019, disponible en <https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines>. Fecha de consulta: 22 de marzo de 2019.

GRUPO DE EXPERTOS DE ALTO NIVEL EN IA DE LA COMISIÓN EUROPEA (b), *Ethics Guidelines for Trustworthy AI*, 8 de abril de 2019, disponible en <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>. Fecha de consulta: 17 de noviembre de 2019.

GRUPO EXPERTO EN RESPONSABILIDAD Y NUEVAS TECNOLOGÍAS DE LA COMISIÓN EUROPEA, *Liability for Artificial Intelligence and other emerging digital technologies*, 2019, pp. 1-70. DOI:10.2838/573689.

Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE.

PARLAMENTO EUROPEO y CONSEJO DE LA UNIÓN EUROPEA, *Directiva del Parlamento Europeo y del Consejo, de 17 de mayo de 2006, relativa a la maquinaria (2006/42/EC*), Diario Oficial de la Unión Europea, L157/24, disponible en <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:157:0024:0086:EN:PDF>.

PARLAMENTO EUROPEO y CONSEJO DE LA UNIÓN EUROPEA, *Directiva del Parlamento Europeo y del Consejo, de 3 de diciembre de 2001, relativa a la seguridad general de los productos (2001/95/EC*), Diario Oficial de la Unión Europea, L11/4, disponible en <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32001L0095&from=EN>.