



FACULTAD DE DERECHO

**INTELIGENCIA ARTIFICIAL Y DERECHO PENAL:  
EN PARTICULAR, LA IMPUTACIÓN OBJETIVA DE  
RESULTADOS LESIVOS**

Autor: Alejandro Pelet Pascual

Curso 2023-2024

5ºE-5

Área de Derecho Penal

Tutor: Javier Gómez Lanz

Madrid

Abril 2024

# ÍNDICE

<b>1. INTRODUCCIÓN.....</b>	<b>4</b>
1.1. LA INTELIGENCIA ARTIFICIAL: PRECISIÓN CONCEPTUAL .....	6
1.2. PERSPECTIVA HISTÓRICA Y EVOLUCIÓN DE LA INTELIGENCIA ARTIFICIAL 10	
1.3. TIPOS DE INTELIGENCIA ARTIFICIAL.....	12
1.4. MARCO REGULATORIO NACIONAL Y EUROPEO. LA PROPUESTA DE REGLAMENTO DE LA INTELIGENCIA ARTIFICIAL DE LA UNIÓN EUROPEA .....	13
<b>2. FENÓMENO DEL <i>ARTIFICIAL INTELLIGENCE CRIME</i> (AIC).....</b>	<b>14</b>
2.1. MERCADOS FINANCIEROS. FIJACIÓN DE PRECIOS Y COLUSIÓN.....	16
2.2. DELITOS CONTRA LA SALUD PÚBLICA. EN PARTICULAR EL TRÁFICO DE DROGAS. ....	17
2.3. DELITOS DE ACOSO.....	18
2.4. DELITOS DE TORTURA.....	19
2.5. DELITOS DE HOMICIDIO Y LESIONES .....	20
2.6. DELITOS CONTRA EL ORDEN SOCIOECONÓMICO .....	21
<b>3. LOS ACTOS Y OMISIONES DE LOS SISTEMAS DE INTELIGENCIA ARTIFICIAL Y LOS PROBLEMAS DE IMPUTACIÓN OBJETIVA. ....</b>	<b>22</b>
3.1. LA IMPUTACIÓN OBJETIVA DE RESULTADOS LESIVOS.....	22
3.1.1. <i>La relación de causalidad</i> .....	25
3.1.2. <i>El riesgo</i> .....	26
3.2. LOS MODELOS DE RESPONSABILIDAD PENAL .....	28
3.2.1. <i>Modelos de responsabilidad penal no directa</i> .....	30
3.2.1.1. El modelo de la autoría mediata .....	30
3.2.1.2. El modelo de la llamada “consecuencia natural y probable” .....	32
3.2.1.3. Modelos alternativos.....	35
3.2.2. <i>Modelo de responsabilidad penal directa</i> .....	37
<b>4. LAS PENAS Y LA INTELIGENCIAS ARTIFICIAL .....</b>	<b>41</b>
<b>5. CONCLUSIÓN.....</b>	<b>49</b>
<b>6. BIBLIOGRAFÍA.....</b>	<b>53</b>

## **LISTADO DE SIGLAS Y REFERENCIAS**

Código Penal (CP)

Comisión Nacional de los Mercados y de la Competencia (CNMC)

Constitución Española (CE)

Estrategia Nacional de Inteligencia Artificial (ENIA)

Inteligencia Artificial (IA)

Ley Orgánica de Protección de Datos y Garantía de Derechos Digitales (LOPDGDD)

Parlamento Europeo (PE)

Procesamiento del Lenguaje Natural (PLN)

Trabajo de Fin de Grado (TFG)

## 1. INTRODUCCIÓN

En el último lustro el desarrollo de la inteligencia artificial (en adelante “IA”) ha permitido al ser humano desarrollar sus tareas diarias de manera más eficiente y rápida. Los sistemas de inteligencia artificial están presentes en todos los aspectos de la vida, desde la sanidad al transporte, en lo que es una tendencia al alza. Claramente, los beneficios son innumerables e indiscutibles y, en diversas ocasiones, se ha alabado el progreso científico y técnico que supone.

La propia Declaración de Montreal de 2018 así lo manifiesta al explicar que: “La inteligencia artificial constituye una forma importante de progreso científico y tecnológico, que puede generar grandes beneficios sociales al mejorar las condiciones de vida y la salud, agilizar la justicia, crear riquezas, reforzar la seguridad pública [...]” (p. 7) o como la reciente Declaración Bletchley, en noviembre de 2023, que remarcaba la capacidad de la IA de mejorar el bienestar social, la paz y la prosperidad. Pero, de la misma manera que presenta beneficios y ventajas, también hay una infinidad de peligros y amenazas que trae consigo. Desafíos sociales, éticos y legales que afectan y afectarán de lleno en los años venideros.

De esta manera, los sistemas de IA pueden llegar a restringir las decisiones de los individuos, constreñir derechos fundamentales, alterar el mercado laboral, influenciar la política, o incrementar las desigualdades económicas y sociales (Dobrinou, 2019). Todo ello en un ambiente de innovación tecnológica que permitirá a los sistemas de IA desarrollar funciones autónomas e incluso cognitivas convirtiéndose en verdaderos agentes que interactúan con el entorno y con capacidad de alterarlo (Dobrinou, 2019).

Es en este contexto donde surge la necesidad crucial de atender a la responsabilidad legal que puede derivar de los hechos cometidos por sistemas de IA y en especial, los problemas derivados de la imputación objetiva de resultados lesivos, en particular a través de la gestión autónoma o manipulación de objetos de la realidad por parte de los sistemas de inteligencia artificial, afectando al entorno y en consecuencia, generando un resultado que puede ser lesivo de bienes jurídicos. Dicho de otra manera, se pretende analizar los problemas de

responsabilidad penal derivados de las acciones cometidas por sistemas de IA que alcanzan autonomía suficiente para tomar decisiones que pueden afectar a su entorno y la realidad que los rodea. Se pretende responder a las siguientes preguntas: ¿A quién se le puede imputar las acciones de un robot o agente artificial que deriven en un daño a un bien jurídico? ¿Al programador/diseñador del sistema de inteligencia artificial o al usuario? Y en los casos donde el robot es autónomo, ¿puede un individuo responder penalmente de las decisiones autónomas tomadas por un sistema de inteligencia artificial sin intervención humana? O, por el contrario, ¿puede responder el propio robot o sistema de inteligencia artificial?

Como se explicará más adelante, la IA puede y podrá intervenir en numerosos procesos criminales ya sea como instrumento o, llegado el momento, como autor. Si una persona comete un delito será declarado criminalmente responsable, siempre y cuando se cumplan con los requisitos exigidos por la ley para que así lo sea. No obstante, el robot o algoritmo inteligente (en este último caso, como ya se explicará, se hace referencia a los delitos de manipulación del mercado) no es destinatario de las normas penales (Blanco, 2019). De tal manera que, si la decisión de actuar ha sido autónoma, sin que medie intervención humana y, en consecuencia, en el ejercicio de su propia voluntad, no parece, *a priori*, que pueda existir sanción alguna (Blanco, 2019). En otras áreas como el Derecho civil existen debates acerca de cómo responder por los daños cometidos por entes inteligentes y en esta línea el Parlamento Europeo, en su Resolución del 16 de febrero de 2017, estableció unas recomendaciones sobre normas de Derecho civil sobre robótica (Blanco, 2019). Mientras tanto, en el ámbito del Derecho Penal no se ha producido una intervención en este aspecto, en tanto la tecnología no ha conseguido avanzar lo suficiente como para entender de la existencia de una inteligencia artificial que imite a la humana, esto es, que reúna las características que posee el ser humano como para ser objeto de sanción penal (Blanco, 2019). Sin embargo, antes o después, llegará dicho momento siendo, por tanto, necesario reflexionar sobre cómo actuar en los casos donde el sistema de Inteligencia Artificial comete un delito tras haber analizado el entorno y haber decidido actuar, con autonomía, para alcanzar determinados objetivos (Blanco, 2019).

## 1.1. LA INTELIGENCIA ARTIFICIAL: PRECISIÓN CONCEPTUAL

La conceptualización de la IA es complicada en tanto que no solo plantea diversos debates y problemas doctrinales en relación con lo que se considera inteligencia, sino que, además, cualquier intento de definición se ve afectada en función de la disciplina que la aborda y los objetivos perseguidos con la misma (Valls, 2022). El propio Comité Económico Social y Europeo en su Dictamen C-288 refleja esta dificultad por elaborar una definición consensuada (2017).

Esta problemática por conseguir un consenso deriva, según Amador, de dos hechos fundamentales. En primer lugar, el rechazo del ser humano a admitir que una máquina pueda llegar a incorporar “capacidades mentales en el más amplio sentido de la expresión” (Amador, 1996, p. 15). En segundo lugar, la inteligencia como concepto mal definido. En relación con este segundo hecho, De la Cuesta viene indicando que el concepto de inteligencia, como una de las características propias del género humano, “no es pacífico ni en su contenido (¿qué es la inteligencia?) ni en su extensión (¿solo son inteligentes los humanos?)” (2019, p. 52)

De esta manera, para poder construir un marco conceptual, es necesario empezar por lo que entendemos por “inteligencia” y “artificial”. Sin llevar a cabo un análisis muy extenso de los términos, ya que escapa al objetivo de esta investigación, acudiendo a la RAE, inteligencia viene definida como la “capacidad de entender, comprender o resolver problemas” (Real Academia Española, s.f.) mientras que artificial, se define como “Hecho por mano o arte del hombre” (Real Academia Española, s.f.), o “Producido por el ingenio humano” (Real Academia Española, s.f.). Es más, podemos encontrar una definición en la RAE de “inteligencia artificial” describiéndola como: “Disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico” (Real Academia Española, s.f.)

La delimitación conceptual de inteligencia artificial puede realizarse a través de dos perspectivas. Desde la primera perspectiva, la IA se puede entender como centrada en el

estudio de los procesos cognitivos “intentando obtener un desarrollo teórico sistematizado de las diversas actividades del intelecto que nos permitan un conocimiento más profundo y preciso del mismo” (Amador, 1996, p. 20). Entre los autores que parten desde esta perspectiva en su análisis de la inteligencia artificial encontramos a Nilsson. Según este autor: “La *Inteligencia Artificial (IA)*, en una definición amplia y un tanto circular, tiene por objeto el estudio del comportamiento inteligente en las máquinas” (Nilsson, 1998, p. 30).

Desde la segunda perspectiva, la IA tiene por objetivo la creación de sistemas automáticos. McCarthy, creador del concepto de la IA como se explicará más adelante, así lo definió: “It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable” (2007). De la misma manera, Rich afirma que: “Artificial intelligence (A.I.) is the study of how to make computers do things that people are better at” (1985, p. 120).

Nos encontramos, por tanto, con multitud de formas de definir la Inteligencia Artificial atendiendo a diversos criterios y perspectivas. No obstante, aquellas que puede encontrar un mayor encaje o que están más cercanas al ámbito del Derecho Penal, son las que se describen a continuación:

- Valls lo define como “la capacidad de la máquina de pensar, de razonar y actuar con inteligencia” (2022, p.5). Así, la IA se entiende como un sistema capaz de actuar en determinadas circunstancias de manera idéntica a un ser humano.
- El Grupo de Expertos de Alto Nivel en Inteligencia Artificial de la Unión Europea describe la IA como los sistemas de software (asistentes de voz, software de análisis de imágenes, motores de búsqueda, sistemas de reconocimiento facial y de voz, entre otros) o de hardware (robots avanzados, drones o coches autónomos) diseñados por humanos, que atendiendo a un objetivo complejo, actúan en el medio físico o digital percibiendo su entorno e interpretando los datos, razonando sobre el conocimiento

derivado de ellos y decidiendo la mejor acción a tomar para alcanzar el objetivo perseguido (2019).

- La Propuesta de Reglamento denominada “Ley IA”, define los sistemas de IA, en su artículo 3, como el software que se ha desarrollado utilizando “una o varias de las técnicas y estrategias [...] que puede, para un conjunto determinado de objetivos definidos por seres humanos, generar información de salida como contenidos, predicciones, recomendaciones o decisiones que influyan en los entornos con los que interactúa” (2021).
- Kaplan y Haenlein entienden la inteligencia artificial como la capacidad del sistema para interpretar correctamente los datos del entorno, aprender de ellos y utilizarlos para lograr objetivos y tareas específicas mediante una adaptación flexible (2019).

En definitiva, hay una serie de características comunes a todas las definiciones existentes tanto a nivel internacional como nacional de la IA y son: la percepción del entorno, incluida la consideración de la complejidad del mundo real; procesamiento de información, recopilando y analizando inputs; toma de decisiones (incluyendo razonamiento y aprendizaje), mediante la realización de acciones y ejecución de tareas cierto nivel de autonomía; y la consecución de objetivos específicos (siendo esta la razón de ser de los sistema de IA) (Samoili et al., 2020).

En cualquier caso, dentro de lo que hoy en día se encuentra bajo el concepto de IA son numerosas técnicas de procesamiento matemático de datos. Entre estas técnicas, se pueden destacar el *Big Data* para la efectiva gestión de volúmenes de datos, el *Data Mining* para encontrar patrones y sintetizar grandes cantidades de datos, el *Machine Learning* que permite el aprendizaje de las maquinas al ir incorporando datos actualizados y el Procesamiento del Lenguaje Natural (PLN) que permite que los sistemas informáticos puedan entender y manipular el lenguaje natural, de manera que puedan reconocer la voz humana y dar una respuesta lógica (Miró, 2018).

Hay dos elementos esenciales que ayuda a determinar el verdadero alcance de los sistemas de IA actuales y futuros: (i) el abanico de instrucciones que sean capaces de ejecutar y (ii) el grado de autonomía con el que lleve a cabo las mismas frente a la influencia del ser humano (Miró, 2018). Así, en referencia al primer elemento, no habría que centrarse en las utilidades concretas y específicas que sean capaces de hacer, sino que habría que analizar la capacidad de razonamiento de la máquina atendiendo al nivel de equivalencia entre lo complejo del procesamiento llevado a cabo por la máquina y el que desarrolla el cerebro humano (Miró, 2018).

Como se puede observar, lo más problemático es en aquellos casos en los que la Inteligencia Artificial evolucione hasta tal punto donde la intervención del ser humano y la intervención de sistema estén tan alejados que difícilmente se pueda atribuir las acciones y decisiones tomadas por estos sistemas a los programadores o usuarios (Miró, 2018). En otras palabras, que atendiendo a la constante evolución de la IA, lleguen a ser capaces de tomar decisiones y realizar acciones en nuestra realidad en base a parámetros distintos y distorsionados de los que en su momento fueron programados y todo ello, debido a su capacidad de autoaprendizaje (Miró, 2018). De tal manera que, estas acciones puedan ser lesivas de bienes jurídicos.

Por todo ello, en la actualidad podemos hablar de tres modelos de IA atendiendo al grado de intervención e interacción del ser humano con la máquina. En primer lugar, *Man in the loop*. En estos casos, la IA requiere de la intervención humana en intervalos regulares de tiempo para llevar a cabo sus acciones (Miró, 2018). En segundo lugar, *Man on the loop*, que implica la capacidad de la máquina de actuar por sí misma debido a una programación previa, si bien el humano puede modificar o interrumpir la actividad de la máquina en cualquier momento (Miró, 2018). Por último, el modelo *Man out of the loop*, en virtud del cual, la máquina es capaz de actuar de manera independiente durante determinados intervalos de tiempo sin que el humano tenga ningún tipo de influencia sobre las acciones de la máquina (Miró, 2018).

## 1.2. PERSPECTIVA HISTÓRICA Y EVOLUCIÓN DE LA INTELIGENCIA ARTIFICIAL

La inquietud del ser humano por reproducir los comportamientos humanos se puede observar a lo largo de toda la historia. Conseguir un objetivo con el mínimo de acciones posibles era un fin que ya estaba presente incluso en los primeros juegos matemáticos, como es el de las Torres Hanói, siendo esta una de las cuestiones que más se ha trabajado en el ámbito de la IA (Rainer y Rodríguez, 2017). Es más, Aristóteles ya propuso la creación de máquinas inteligentes en el año 322 a.C. (Rainer y Rodríguez, 2017). Herón de Alejandría, en su obra *Autómata*, trató la figura del robot, describiendo numerosos aparatos que normalmente tenían fines religiosos, si bien, con mayor frecuencia, buscaban el esparcimiento y el acompañamiento (Rainer y Rodríguez, 2017). El Antiguo Egipto también fue un referente en la invención de nuevos mecanismos, normalmente, en un afán de atemorizar y sorprender a quienes los contemplasen. Claramente, los avances de cada época han estado limitados por el estado de la técnica de cada momento (Rainer y Rodríguez, 2017).

El origen exacto de la IA es difícil de determinar, aunque probablemente se puede remontar a los años 40, específicamente al año 1942, año en el cual Isaac Asimov publicó su libro *“Círculo Vicioso”* (Haenlein y Kaplan, 2019). La trama se desenvuelve en torno a un robot y las tres leyes de la robótica que implican: que un robot no puede lesionar a un humano o permitir, a través de su inacción, que un humano sea lesionado; que un robot tiene que seguir las órdenes dadas por los humanos salvo en aquellos casos donde dichas ordenes entren en conflicto con la primera norma; y finalmente, que un robot debe proteger su propia existencia siempre y cuando esta protección no entre en conflicto con la primera o segunda norma (Haenlein y Kaplan, 2019). En 1950, Alan Turing publicó su artículo *“Maquinaria computacional e Inteligencia”* en la que describía cómo crear máquinas inteligentes y, en particular, cómo examinar su inteligencia en lo que se denomina el Test de Turing, hoy todavía considerado como un referente a la hora de determinar la inteligencia de un sistema artificial (Haenlein y Kaplan, 2019). Esta prueba funciona bajo la siguiente premisa: si un humano interactúa con un humano y una máquina a la vez y es incapaz de distinguir entre ambos, entonces se puede decir que la máquina es inteligente (Haenlein y Kaplan, 2019).

No obstante, el término inteligencia artificial fue acuñado por Marvin Minsky y John McCarthy en 1956 durante la Conferencia de Dartmouth en la que se reunieron los que más tarde fueron considerados como padres de la IA. Tras esta Conferencia, vieron varias décadas de grandes avances en este ámbito, como así fue con el programa ELIZA creado en mitad de la década de los años 60 por Joseph Weizenbaum (Haenlein y Kaplan, 2019). ELIZA fue una herramienta de procesamiento del lenguaje natural con capacidad para simular una conversación con un ser humano siendo uno de los primeros programas que intentó superar el Test de Turing (Haenlein y Kaplan, 2019). No obstante, en la década de los 70, surgieron críticas debido al enorme gasto en investigación sobre IA. Esta situación dio lugar a lo que se denominó el Invierno de la Inteligencia Artificial. La particular forma en que los primeros sistemas, como ELIZA, intentaron replicar el intelecto humano explica, en gran parte, la falta de mayores avances en el campo durante años. Estos sistemas eran sistemas expertos que son un conjunto de reglas basadas en la idea de que el intelecto humano puede reducirse a un conjunto de afirmaciones "si-entonces" (Haenlein y Kaplan, 2019). Así, en aquellas áreas que permiten tal formalización, los sistemas expertos se desempeñaban perfectamente. Sin embargo, en aquellas áreas que no permitían semejante formalización, se presentaban más complicadas para este tipo de sistemas, como es por ejemplo el reconocimiento facial (Haenlein y Kaplan, 2019). Consecuentemente, estos sistemas no pueden ser considerados verdaderos sistemas de Inteligencia Artificial (Haenlein y Kaplan, 2019).

En 1969, se empezó a investigar sobre redes neuronales artificiales. Sin embargo, debido a la falta de capacidad de procesamiento de los ordenadores del aquel momento, esta investigación se estancó (Haenlein y Kaplan, 2019). Fue ya en 2015, cuando a través de *AlphaGo*, un programa desarrollado por Google, que este campo de investigación y desarrollo sobre la IA reapareció en forma de *Deep Learning* (Haenlein y Kaplan, 2019). Hoy en día, las redes neuronales y *Deep Learning* son la base de las mayorías de aplicaciones que hoy calificamos como inteligencia artificial. Es la base de algoritmos como el usado por Facebook para el reconocimiento de imágenes o el de los coches autónomos (Haenlein y Kaplan, 2019).

### 1.3. TIPOS DE INTELIGENCIA ARTIFICIAL

Se suele hablar o referenciar tres tipos de IA atendiendo a su capacidad de imitar características humanas. La IA débil o estrecha (“narrow AI”) que tiene un rango limitado de habilidades, la general o fuerte (“strong or general AI”) que tiene habilidades a la par de los humanos y la superinteligente que desarrolla habilidades superiores a las humanas.

La inteligencia artificial débil hace referencia a todos los modelos de IA desarrollados hoy en día y que han tenido mayor éxito. Este tipo de inteligencia artificial está hecha para llevar a cabo tareas y actividades específicas como el reconocimiento facial, asistentes de voz o la conducción (Escott, 2017). Dicho de otra manera, está altamente capacitada para realizar una tarea específica y concreta para lograr un objetivo determinado, siguiendo una serie de pasos predeterminados que permanecen invariables y constantes frente a cualquier tipo de inputs sin ir, en ningún caso, más allá de su programación original (Del Rosal, 2023).

La inteligencia artificial general es aquella que replica la inteligencia y/o comportamientos humanos con la capacidad de aprender y aplicar su inteligencia para resolver problemas (Escott, 2017). Es, por tanto, capaz de pensar, comprender y actuar de forma indistinguible a la de un humano, en lo que se denomina *machine learning* (Del Rosal, 2023). Se hace posible el aprendizaje autónomo sin necesidad de que la propia máquina haya sido programada para ello (Del Rosal, 2023). Este tipo de sistemas de inteligencia artificial no han sido desarrollados plenamente todavía, en tanto que implicaría conseguir hacer las máquinas conscientes (Del Rosal, 2023). Para alcanzar este hito sería necesario programar las máquinas con un conjunto completo de habilidades cognitivas como es aplicar la experiencia para resolver problemas (Del Rosal, 2023). El objetivo final de un sistema de IA general es enseñar a las máquinas a comprender a los humanos a un nivel verdaderamente humano y no una mera réplica o simulación del comportamiento humano (Del Rosal, 2023).

Está aún lejos el momento para conseguir un modelo de inteligencia artificial general capaz de actuar y tomar decisiones que afecten al entorno, pero el desarrollo y evolución de los sistemas de aprendizaje automático es cada vez más rápido (Del Rosal, 2023). En este

sentido, el *Stanford AI Index Report 2023*, recoge que los sistemas de IA de aprendizaje automático más significativos en 2022, que se desarrollaron y se lanzaron, fueron los de lenguaje. En 2022, se pusieron en marcha 23 sistemas de lenguaje de inteligencia artificial, cerca de seis veces más que los sistemas multimodales (Stanford University, 2023). La inteligencia artificial general todavía no se ha conseguido, pero el número de sistemas de inteligencia artificial que se programan son cada vez más complejos y de mayor envergadura (Stanford University, 2023). Además, crear sistemas de inteligencia artificial requiere de grandes cantidades de datos, poder de computación y dinero (Stanford University, 2023).

Finalmente, la superinteligencia, que es el modelo hipotético en virtud del cual, no solo replicaría o comprendería el comportamiento o la inteligencia humana, sino que las máquinas se convertirían en seres conscientes de sí mismos llegando a superar la inteligencia y habilidad humana (Del Rosal, 2023). Se estaría hablando de un plano totalmente distinto donde la IA evoluciona hasta tal punto que se asemejaría tanto a las emociones y experiencias humanas que no solo las comprende, sino que también evoca sus propias emociones, necesidades, creencias y deseos (Escott, 2017).

#### 1.4. MARCO REGULATORIO NACIONAL Y EUROPEO. LA PROPUESTA DE REGLAMENTO DE LA INTELIGENCIA ARTIFICIAL DE LA UNIÓN EUROPEA

La regulación sobre inteligencia artificial hasta el momento ha sido bastante escasa. A nivel nacional, se presentó en mayo de 2021 la Estrategia Nacional de Inteligencia Artificial (ENIA) por la que se establece un plan futuro para el desarrollo de la IA en España proporcionando un claro marco de referencia para el desarrollo de la IA (Ministerio de Economía, Comercio y Empresa, 2020). Esta estrategia forma parte del Plan de Recuperación, Transformación y Resiliencia de la economía española (Ministerio de Economía, Comercio y Empresa, 2020). Además, se incluyen una serie de medidas con el fin de promover la innovación, transparencia y ética en el ámbito de la IA. Asimismo, a nivel regulatorio se pueden encontrar leyes y la normativa específica en determinados ámbitos que influyen en el desarrollo de la IA como Ley Orgánica de Protección de Datos y Garantía de Derechos Digitales (LOPDGDD) o la Ley de Seguridad Ferroviaria (Ministerio de Economía, Comercio y Empresa, 2020).

No obstante, el marco regulatorio más importante es el que se está desarrollando en el seno de la Unión Europea. El 21 de abril de 2021, la Comisión Europea presentaba una propuesta del primer Reglamento de Inteligencia Artificial conocida hoy en día como Ley IA. Esta Propuesta de Reglamento tiene por objeto, conforme a su artículo 1: establecer unas normas armonizadas para el desarrollo, la introducción en el mercado y la utilización de sistemas de Inteligencia Artificial en la Unión Europea; prohibir determinadas prácticas; establecer determinados requisitos para sistemas de IA de alto riesgo, así como obligaciones para los operadores de dichos sistemas; normas de transparencia aplicables a los sistemas IA; y normas de control y vigilancia del mercado.

En diciembre de 2023, la Presidencia del Consejo y el Parlamento Europeo alcanzaron un acuerdo sobre esta propuesta, modificando ciertos aspectos de la propuesta de la Comisión (Consejo de la Unión Europea, 2023). Los principales cambios de este acuerdo han sido el establecimiento de normas sobre modelos de IA de uso general que pueda llegar a causar riesgos sistémicos, un revisado sistema de gobernanza, una ampliación de la lista de prohibiciones y mayor protección de los derechos fundamentales que puedan verse afectados por los sistemas de IA (Consejo de la Unión Europea, 2023).

El 13 de marzo de 2024 el Parlamento Europeo aprobó el Reglamento acordado con los Estados Miembros en diciembre de 2023, aunque todavía está sujeto a una serie de comprobaciones jurídico-lingüísticas y se requiere la adopción formal por parte del Consejo (Parlamento Europeo, 2024).

## **2. FENÓMENO DEL *ARTIFICIAL INTELLIGENCE CRIME* (AIC)**

La inteligencia artificial juega un rol cada vez más preponderante en los actos criminales, generando riesgos digitales, físicos y políticos. La IA tiene la capacidad de aumentar la frecuencia de una variedad de nuevas formas de delincuencia, del mismo modo que la aparición del Internet lo hizo al facilitar una amplia gama de delitos no tradicionales. De esta

manera, la forma en que la inteligencia artificial puede incidir en la comisión de hechos delictivos es muy variada y plural, lo que ha provocado que parte de la doctrina use el término *AI-Crime* para hacer referencia a los delitos cometidos a través de la IA (King et al., 2019).

Los criminales pueden utilizar la IA, por ejemplo, para agilizar y mejorar sus ataques aprovechándose de nuevas víctimas, maximizando el potencial de beneficios en menos tiempo y desarrollando modelos de negocio ilegales más creativos a la vez que disminuyen la probabilidad de ser descubiertos (Europol, 2021). Además, a medida que prolifere la IA como servicio, disminuirán las habilidades y los conocimientos técnicos necesarios para utilizarla, lo que reducirá la barrera de entrada (Europol, 2021). La rápida adopción de las nuevas tecnologías por parte de criminales y de grupos de crimen organizado en su *modus operandi* no solo plantea diversos y continuos cambios en el entorno delictivo mundial, sino que también provoca grandes problemas a las fuerzas de seguridad y ciberseguridad (Europol, 2021). Así, los delincuentes aprovechan las nuevas tecnologías como la IA como catalizador de sus actividades delictivas (Europol, 2021). En muchos casos se debe a que facilita a los criminales acceder a herramientas y servicios que aumenta su capacidad de ataque (Europol, 2021).

De este modo, la IA puede ser utilizada para fines criminales de distintas formas. Puede ser utilizada como herramienta para la delincuencia, es decir, aquellos casos donde la IA se utiliza para cometer un delito tradicional, como el robo o la intimidación; o, también puede ser el objeto de la actividad delictiva, como son los intentos de eludir los sistemas de protección de IA o para hacerlos fallar o que actúen de forma errática (Caldwell et al., 2020). Incluso existe la posibilidad de utilizar la IA como contexto para la actividad criminal, es decir, hacer creer a la víctima que la IA tiene una serie de funcionalidades que no tienen como es predecir el mercado de valores o manipular a los votantes (Caldwell et al., 2020).

En aras de hacer un análisis de los ámbitos donde puede tener mayor incidencia la Inteligencia Artificial, se va a separar, a modo de ejemplo, en varios apartados que permiten visualizar el amplio campo de actividades delictivas en las que puede intervenir la IA: (i) Mercados financieros, fijación de precios y colusión; (ii) delitos contra la salud pública, en

particular el tráfico de droga; (iii) delitos de acoso; (iv) delitos de tortura; (v) delitos de homicidio y lesiones; y (vi) delitos contra el orden socioeconómico.

## 2.1. MERCADOS FINANCIEROS. FIJACIÓN DE PRECIOS Y COLUSIÓN

En este apartado se analizará la intervención de la IA en la manipulación del mercado, fijación de precios y colusión. De esta manera, la IA puede tener una gran influencia en la manipulación de mercados, entendiendo esta como la conducta que puede lesionar el objetivo de la libre formación de los precios (Real Academia Española, s.f.). Dicho de otra manera, todas aquellas acciones y/u operaciones de los operadores del mercado que intentan influir artificialmente en los precios del mismo (Spatt, 2014). Es un delito regulado en los artículos 284 y ss. del Código Penal (en adelante “CP”) cuyo bien jurídico protegido es la integridad de los mercados y la confianza de los inversores que actúan en ellos como se indica en la LO 1/2019, de 20 de febrero. Los engaños utilizados para alterar y manipular el precio y los mercados pueden provenir de agentes artificiales que están diseñados para operar en nombre de un usuario y esto se debe a que un agente artificial puede llegar a adquirir la capacidad de, a partir de observaciones reales o simuladas, generar señales que engañen al resto (King et al., 2019). En los modelos de mercados simulados que incluyen agentes artificiales, se ha demostrado que, mediante el aprendizaje por refuerzo, estos agentes pueden llegar a aprender a suplantar el libro de órdenes, conocido como “*spoofing*” (King et al., 2019). Del mismo modo, los *chatbots* también pueden provocar disrupciones en el mercado pudiendo ser parte de lo que se conoce como *pump and dump*, una maquinación por la cual la parte manipuladora adquiere una posición determinada en un producto financiero, como son las acciones, inflándolas artificialmente a través de una promoción fraudulenta antes de vender su posición a terceros, desplomándose tras la venta (Lin, 2017). Los conocidos como *social bots* (un tipo de *chatbot*) han resultado ser muy útiles para este tipo de maquinaciones pudiendo diseminar desinformación sobre una empresa (Lin, 2017). Si bien es poco probable que la mayoría de los profesionales humanos se vean influenciados por este tipo de *spam* en las redes sociales, los agentes de negociación algorítmica responden precisamente a este tipo de desinformación (King et al., 2019).

Otro reto que plantea la IA y que afecta a la seguridad del mercado y de los consumidores, es la colusión algorítmica. El uso extendido de algoritmos puede derivar en comportamientos anticompetitivos permitiendo a las empresas coordinarse sin necesidad de un acuerdo formal o interacción humana (CNMC, s.f.). Pueden ser utilizados para monitorizar una estrategia coordinada preestablecida. No obstante, en estos casos es necesario una comunicación explícita entre las empresas con el fin de crear el cartel y a continuación, aplicar los algoritmos para vigilar el acuerdo alcanzado (CNMC, s.f.). Asimismo, existe la posibilidad de utilizar algoritmos de precios que provocan una coordinación tácita que no requiere de comunicación entre los competidores, como puede ser en el caso de que varias empresas soliciten a la misma empresa de software el diseño de su algoritmo (CNMC, s.f.). Por último, puede darse el caso de que las empresas opten por utilizar algoritmos de *Deep Learning* para decidir sobre los precios (CNMC, s.f.). Son estos casos los que más preocupación pueden causar, dado que los rivales diseñan de manera unilateral sus algoritmos de fijación de precios imponiendo un determinado objetivo (CNMC, s.f.). Dicho algoritmo va aprendiendo y, en un determinado momento, escoge cual es la estrategia más óptima para cumplir con el objetivo, pudiendo de esta manera llegar a la conclusión de que la mejor estrategia radica en coludir (CNMC, s.f.). Por tanto, no es necesario que un algoritmo esté diseñado específicamente para coludir para que se produzca dicha colusión, puesto que la inteligencia artificial juega un rol cada vez más importante en la toma de decisiones y los algoritmos a través de la prueba y error pueden llegar al resultado de coludir (CNMC, s.f.).

## 2.2. DELITOS CONTRA LA SALUD PÚBLICA. EN PARTICULAR EL TRÁFICO DE DROGAS.

La IA puede ser un instrumento clave de apoyo a la distribución y venta de sustancias prohibidas y reguladas en los artículos 368 y ss. del CP siendo el bien jurídico protegido, la salud pública. El uso de vehículos no tripulados que utilizan la IA y navegación autónoma exagera las tasas de éxito del contrabando (King et al., 2019). Atendiendo a que las redes de contrabando se desmantelan vigilando e interceptando líneas de transporte, la utilización de este tipo de vehículos hace mucho más complicado el trabajo de las Fuerzas y Cuerpos de seguridad del Estado (King et al., 2019). Tal es el caso, que incluso los drones se configuran

como una amenaza en la forma de contrabando automatizado y en las últimas décadas, las redes de crimen organizado están utilizando submarinos controlados remotamente para traficar con cocaína, habiendo sido captados este tipo de submarinos no tripulados por los Cuerpos de Seguridad de Estados Unidos (King et al., 2019). Los vehículos submarinos no tripulados son un ejemplo perfecto de la doble vertiente que tiene la IA. Este tipo de vehículos puede ser usados lícitamente por motivos de defensa, para la protección de las fronteras o para patrullaje marítimo, entre otras aplicaciones (King et al., 2019). Sin embargo, también se han mostrados efectivos para desarrollar actividades ilegales. En el caso de tráfico de drogas, los criminales pueden evitar su implicación debido a que este tipo de vehículos pueden actuar independientemente de un operador (King et al., 2019). Así, no hay un nexo que se pueda verificar entre el vehículo y quien lo desplegó si tanto el software como el hardware carecen de un rastro que permita averiguar quién lo obtuvo y cuando, al igual que en los casos donde las pruebas son destruidas al ser interceptados los vehículos (King et al., 2019).

### 2.3. DELITOS DE ACOSO

Los actores maliciosos también pueden usar los *bots* sociales como instrumentos directos e indirectos para cometer delitos de acoso recogido en el artículo 172 ter del CP. Utilizar herramientas como dar “me gusta” o retuitear son tácticas indirectas, como también es la manipulación de encuestas para crear la impresión de que una persona es objeto de hostilidad generalizada (King et al., 2019). Asimismo, los actores criminales pueden manipular los *bots* sociales de otro actor mediante la interacción con el usuario socavando las estructuras de datos de clasificación y generación aprendidas por el *bot* (King et al., 2019). Un claro ejemplo es Tay, un *chat bot* de inteligencia artificial creado por Microsoft. Este *chat bot*, al aprender de sus interacciones con usuarios de Twitter empezó a escribir tweets incendiarios y obscenos utilizando palabras e imágenes inapropiadas (Neff y Nagy, 2016). Al poco tiempo, Microsoft cerró la cuenta de Twitter de Tay (Neff y Nagy, 2016). Hay que remarcar que, en estos casos, lo que se podría considerar acoso puede entrelazarse con el uso de los *bots* sociales para manifestar una opinión, dicho de otra manera, para ejercer la libertad de expresión (King et al., 2019). En estos casos tendrá que ser los Tribunales quienes resuelvan esta controversia.

Por lo tanto, algunas de estas actividades pueden ser constitutivas de acoso en el sentido de un comportamiento social pero no jurídicamente inaceptable, mientras que otras actividades pueden alcanzar el umbral del acoso delictivo (King et al., 2019).

Asimismo, la IA tiene la capacidad de producir contenidos falsos cada vez más complejos, lo que abre la puerta a nuevos tipos de acoso. Los desarrolladores han producido software que crea videos sintéticos (King et al., 2019). Estos videos tienen como base un video real en el que aparece una persona si bien el programa cambia la cara de esta persona por la de otra (King et al., 2019). Muchos de estos videos son pornográficos existiendo el riesgo de que sean utilizados por los usuarios para acosar a las víctimas (King et al., 2019).

#### 2.4. DELITOS DE TORTURA

La inteligencia artificial juega un papel importante en la labor policial. La utilización de robots por parte de la policía es un fenómeno cada vez más extendido (King et al., 2019). En un tiroteo en Dallas en 2016, la policía utilizó un robot armado con explosivos para matar a un tirador (King et al., 2019). Del mismo modo, también han utilizado robots para desarmar bombas o incluso entregar teléfonos móviles para facilitar una negociación. No obstante, el uso de los robots se puede expandir también al ámbito de la interrogación debido, por un lado, a la incapacidad de los seres humanos de ser detectores fiables de mentiras y la confianza en la tecnología para mejorar las capacidades humanas y por otro lado, los crecientes avances en las herramientas de monitorización fisiológica y la investigación sobre la interacción entre el ser humano y el ordenador (McAllister, 2017). Por tanto, el uso de la IA en los interrogatorios está motivado por la capacidad para detectar engaños, imitar rasgos humanos y poder manipular al interrogado (McAllister, 2017). Así, un desarrollador puede incorporar las capacidades de autonomía y análisis de la IA en un robot o agente artificial de interrogatorio abriéndose la posibilidad de que puedan llegar a torturar a una víctima (King et al., 2019). El riesgo radica en que el agente artificial sea desplegado para aplicar técnicas de tortura psicológica o física, incluso pudiendo ocurrir sin que haya intervención humana (King et al., 2019).

## 2.5. DELITOS DE HOMICIDIO Y LESIONES

Existe la posibilidad de que máquinas controladas por IA ocasionen homicidios y lesiones imprudentes atentando de esta manera contra dos bienes jurídicos fundamentales, la vida y la integridad física. Diversas son las maneras a través de las cuales la IA podría atentar contra estos bienes jurídicos, pero especial atención merece los sistemas de armas autónomas (AWS por sus siglas en inglés). Se lleva tiempo estudiando este campo dentro del mundo militar dado que las máquinas controladas por IA permiten procesar datos, analizar información y tomar decisiones en menos tiempo que los seres humanos, lo cual es especialmente útil en el ámbito de la defensa (Blauth et al., 2022). Así, los AWS se pueden definir como sistemas de IA programados especialmente para seleccionar (búsqueda y detección) y atacar (uso de la fuerza correspondiente) objetivos sin requerir control o intervención humana tras su activación (Blauth et al., 2022). Hay diversas armas autónomas letales que usan la IA ya desarrolladas y en desarrollo. En este sentido, se puede destacar los robots militares controlados por IA. En ciertos casos, la orden inicial de atacar sería dada por un ser humano, pero ya se están desarrollando sistemas donde la interacción humana no sería requerida delegándose la decisión de emplear fuerza letal en la máquina (Dresp-Langley, 2023). Asimismo, es relevante señalar el papel que juegan dentro de los AWS los drones autónomos. Se estima que los enjambres de drones completamente autónomos se conviertan en armas de destrucción masiva al combinar dos propiedades únicas: daño masivo y falta de control humano que permita garantizar que no se dañe a civiles (Dresp-Langley, 2023). Además, claramente la incorporación de IA a las armas y su configuración como sistemas de armas autónomas conlleva otros riesgos como puede ser que el software integrado en el hardware militar (como por ejemplo los drones) sea alterado por agentes maliciosos (Blauth et al., 2022).

Mas allá de los sistemas de armas autónomas, cabe hablar de los coches autónomos. El avance en el desarrollo de estos es cada vez mayor. Estos coches solo pueden funcionar correctamente si interactúan de forma segura con su entorno (Gless, 2016). Los coches autónomos pueden ser incapaces de reaccionar frente a una crisis imprevista o, simplemente, su tecnología puede fallar, causando accidentes y posiblemente daños, lesiones e incluso la

muerte (Gless, 2016). Además, los coches autónomos pueden conllevar una expansión de lo que se conoce como terrorismo vehicular al no requerir el reclutamiento de conductores (Caldwell et al., 2020). De esta manera, un solo perpetrador puede llegar a realizar múltiples atentados, o incluso coordinar un gran número de vehículos a la vez con fines terroristas (Caldwell et al., 2020)

## 2.6. DELITOS CONTRA EL ORDEN SOCIOECONÓMICO

Respecto al delito de robo y estafa, la IA puede ser incorporada al proceso delictivo cuando no cometer estos delitos por sí misma. La IA puede recabar datos personales y utilizarlos en contra del titular, así como utilizar diversos métodos para falsificar una identidad que pueda confundir a las autoridades bancarias y permitir una determinada transacción (King et al., 2019). Existen diferentes métodos a través de los cuales la IA facilita la comisión de este tipo de delitos. Los sistemas de IA pueden intentar obtener información personal a través de redes sociales o incluso llegar a manipular a la víctima para obtener información o acceder a su ordenador (King et al., 2019). Si bien hoy en día las capacidades de la IA conversacional son limitadas, en un futuro es un problema a tener en cuenta. Es más, actual e independientemente de estas limitaciones, los actores criminales pueden llegar a utilizar *bots* sociales, lo suficientemente amplios, para descubrir a personas susceptibles a este tipo de engaños y verse ayudados gracias a la IA para llevar a cabo estos delitos (King et al., 2019). Esto se debe a que la IA puede ayudar a “to produce more intense cases of simulated familiarity, empathy, and intimacy, leading to greater data revelations” (Graeff, 2014, p. 5). También pueden verse ayudados a través de lo que se conoce como *phishing* automatizado (King et al., 2019). En estos casos, los investigadores han demostrado que es posible utilizar técnicas de aprendizaje automático para elaborar mensajes personalizados para atacar a un usuario concreto, lanzando ataques de *phishing* específicos (conocido como *spear phishing*) (King et al., 2019). Igualmente, la IA puede ayudar a falsificar una identidad basándose en nuevas técnicas de síntesis de voz (King et al., 2019). Este último aspecto es importante destacarlo en tanto que las tecnologías de síntesis de voz asistidas por IA son una verdadera amenaza en términos de robos y estafas (King et al., 2019). Utilizando estas

técnicas se podrían desbloquear puertas y vehículos al poder reproducir la voz de un usuario o cliente (King et al., 2019).

### **3. LOS ACTOS Y OMISIONES DE LOS SISTEMAS DE INTELIGENCIA ARTIFICIAL Y LOS PROBLEMAS DE IMPUTACIÓN OBJETIVA**

#### **3.1. LA IMPUTACIÓN OBJETIVA DE RESULTADOS LESIVOS**

Para abordar la problemática de la imputación objetiva en relación con la inteligencia artificial, es necesario determinar los elementos que configuran el delito. Dos grandes macroelementos lo configuran: la antijuricidad penal y la culpabilidad. Ambos elementos son necesarios para poder atribuir responsabilidad penal e imponer una sanción. Dicho de otra manera, para poder aplicar una pena, se requiere, además de un injusto grave (acción penalmente antijurídica), que el sujeto que lo lleva a cabo se relacione de alguna manera con su conducta ilícita, es decir, que reúna los requisitos determinados (imputabilidad); que su voluntad haya envuelto de determinada manera el ataque al bien jurídico (dolo) y que se encuentre en una situación normal (exigibilidad de una conducta distinta). La antijuricidad penal y la culpabilidad determinan la existencia de un delito. Resulta evidente la necesidad de centrarse en la antijuricidad penal que está compuesto a su vez por la acción, la tipicidad, y la falta de justificación (Obregón y Gómez, 2023). La tipicidad es el elemento básico y crucial de la antijuricidad penal que determina qué hechos son constitutivos de un ilícito penal. De esta manera, el juicio de la tipicidad presupone la existencia de unos hechos calificables como típicos (la acción) que debe ser completado, para afirmar la antijuricidad penal, con un examen de la ausencia de un permiso jurídico específico para llevar a cabo la acción típica (la falta de justificación). La tipicidad es un predicado principalmente objetivo, es decir, recurre a la hora de delimitar legalmente los comportamientos típicos a elementos objetivos que describen el proceso causal, si bien también se sirve en determinadas ocasiones de elementos normativos y subjetivos. Implica una valoración negativa de la conducta tipificada en tanto que se considera que dicha conducta constituye, según marca el legislador, una agresión (lesión o puesta en peligro) a bienes jurídicos (Obregón y Gómez, 2023).

Los elementos de la conducta típica son a grandes rasgos dos. Por un lado, la acción en sentido propio, descrita en el tipo y, por otro lado, la producción de un resultado separable espacial y temporalmente de la acción (Obregón y Gómez, 2023). No obstante, hay diferentes clases de tipos, pudiéndose distinguir entre los tipos sin resultado, que se centran en desvalorar la consumación de la acción sin que se exija la producción de un resultado ulterior y los tipos de resultado donde el desvalor recae en la producción de un resultado como consecuencia de la realización de la acción de la que es separable espacial y temporalmente. La distinción entre acción y resultado deriva de los artículos 11 y 16 del CP. En dichos artículos se encuentran referencias al “resultado” y a la existencia, por tanto, de delitos que necesitan un resultado para la consumación de los mismos (Obregón y Gómez, 2023).

Es en los delitos de resultado donde se hace necesario poder conectar la manifestación de voluntad (acción u omisión) con el resultado o efecto, puesto que solo en aquellos casos donde el resultado provenga de dicha manifestación podrá considerarse al autor de la acción causante del resultado (Rodríguez, 2010). Es decir, si el tipo exige para su consumación, la producción de un resultado resulta necesario poder atribuir el resultado a la acción del sujeto. De esta manera, la problemática residía en la relación de causalidad, cuya existencia determinaba la consumación del delito. Se desarrollaron numerosas teorías como la teoría de la equivalencia de las condiciones o de la causalidad adecuada. No obstante, estas teorías fueron completadas con un aspecto normativo o valorativo que dio pie a la teoría hoy imperante, la teoría de la imputación objetiva (Obregón y Gómez, 2023). En virtud de la misma, un sujeto solo responderá del resultado causado por su acción en aquellos casos donde el resultado pueda ser puesto a cargo de la conducta entendiendo esta como obra de su autor. Dicho de otra manera, la imputación objetiva en palabras de Obregón García y Gómez Lanz es “la atribución de un resultado a la acción del sujeto como obra suya, debido a la existencia de una determinada relación de riesgo entre la acción y el resultado” (2023, p.79).

En cualquier caso, para poder afirmar la imputación objetiva, es presupuesto necesario una relación de causalidad entre acción y resultado que se analiza en virtud de la teoría de la equivalencia de las condiciones. Sin embargo, la relación de causalidad no agota por sí misma

la relación entre acción y resultado, sino que se requiere, en virtud de esta teoría, que entre la acción y resultado se constituya una relación de riesgo que implica la concurrencia de varios requisitos: (i) la acción tiene que ser peligrosa, es decir, que objetivamente haya creado o incrementado el riesgo o peligro de producción del resultado, y (ii) la acción debe ser lo suficientemente peligrosa, esto es, que el riesgo creado por la acción haya producido efectivamente el resultado, de tal manera que sea posible afirmar que éste es la concreción o, mejor dicho, el correlato lógico de aquel (Obregón y Gómez, 2023).

A continuación, se expondrán ambos elementos de la imputación objetiva y, de manera general, cómo se interrelacionan con los sistemas de inteligencia artificial. En apartados sucesivos, se estudiarán los diferentes modelos de atribución de responsabilidad que la doctrina ha introducido para dar respuesta a los problemas de imputación que la IA plantea.

No obstante, y antes de todo, conviene destacar que dentro de las diferentes modalidades de *AI Crime* que se han expuesto en el apartado en el anterior, los problemas de imputación objetiva son múltiples y diversos. A modo de ejemplo, dentro de los delitos de manipulación de mercados, el principal problema de imputación objetiva es atribuir la alteración artificial de los precios del mercado a la acción realizada; en los delitos de homicidios y lesiones, los problemas de imputación objetiva residen en atribuir el resultado lesivo de muerte o lesión a la acción realizada; o en los delitos de orden socioeconómico, imputar el perjuicio patrimonial a la acción desarrollada.

En cualquier caso, los problemas de imputación objetiva se predicán de los tipos de resultado, no siendo, por tanto, objeto de estudio los tipos sin resultado como son los tipos mera actividad. A modo de ejemplo, quedaría excluido del análisis y de los problemas de imputación objetiva, un delito de seguridad vial por exceso de velocidad (art. 379.1 del CP), donde un coche con conducción automática al no registrar correctamente la velocidad máxima permitida la sobrepasa.

### 3.1.1. La relación de causalidad

La teoría de la equivalencia de las condiciones fue formulada por Glaser. Sostiene que toda condición necesaria para la producción de un determinado resultado tiene el valor de causa. Esta teoría es comúnmente conocida como teoría de la *condicio sine qua non*, y se aplica de la siguiente manera: si se suprime mentalmente la acción y desaparece el resultado, entonces la acción es causa del resultado. Esta teoría presenta un gran inconveniente que es la excesiva extensión de la esfera de tipicidad, en tanto que amplía hasta el infinito la relación de causalidad ya sea desde el punto de vista de las causas como desde el punto de vista de los efectos (Obregón y Gómez, 2023).

Como ya se ha comentado, por el momento no se reconoce personalidad a las máquinas y por tanto no les es atribuible ninguna responsabilidad penal por el daño causado por las mismas. La atribución de responsabilidad cuando se trata de sistemas poco avanzados suele ser más sencilla aplicando las doctrinas tradicionales. No obstante, incluso en estos casos pueden surgir problemas dado que en el desarrollo, programación, producción, distribución, implementación y uso de estos sistemas interviene una cantidad considerable de personas e incluso de empresas que pueden ser potencialmente responsables (Diamantis, 2023). De ahí que, en el desarrollo de sistemas de IA, esta situación se complica todavía más, especialmente cuando se aplica la teoría de la equivalencia de las condiciones. A esta complejidad se le denomina *many hands problem* que representa la dificultad de fijar unos criterios claros de atribución de responsabilidad a las personas físicas en estos casos (Diamantis, 2023). Este concepto es desarrollado en profundidad por Diamantis en el contexto de la criminalidad de la empresa. Este autor explica cómo en las operaciones corporativas de gran escala, hay un gran número de empleados participantes, con equipos distribuidos de miles de empleados que se encargan de diseñar y ejecutar algoritmos cooperativos (Diamantis, 2023). En este sentido, un operador puede convertir la función corporativa en algo perjudicial, bien a propósito o bien por negligencia. El problema de las “demasiadas manos” se origina cuando es difícil o prácticamente imposible demostrar que existe el empleado que por su conducta deficiente contribuyó al daño producido por el algoritmo (Diamantis, 2023). Igualmente, Diamantis también habla del *no hands problem*. En virtud del mismo, hace referencia a

aquellas operaciones empresariales complejas que pueden salir mal incluso cuando todos los empleados se hayan comportado de forma responsable (Diamantis, 2023). Dicho de otra manera, se darán los supuestos donde un algoritmo se comporte de manera inusual o incorrecta, aun cuando todas aquellas personas que intervienen en el proceso, ya sea el programador o el usuario que lo maneja, se hayan comportado de manera correcta, irreprochable. Debido al desarrollo tecnológico, los algoritmos incorporan cada vez más un cierto margen de imprevisibilidad (Diamantis, 2023).

Los problemas que se derivan para la relación de causalidad son claros por cuanto se hace muy complicado saber cuál es la acción que efectivamente ha llevado a la producción de un daño. Hay que preguntarse si efectivamente ha sido quien ha programado el sistema de IA o quien lo ha utilizado o, si por el contrario, ninguno de los anteriores es el responsable, entrando de lleno en el terreno de la responsabilidad de los propios sistemas de IA. No obstante, en virtud de la teoría de la *condicio sine qua non* explicada, en estos casos, si se elimina cualquier intervención, el resultado desaparece.

### 3.1.2. *El riesgo*

En lo que afecta al riesgo, se van a analizar los dos requisitos anteriormente descritos y cómo se ven afectados cuando se introduce la inteligencia artificial. Empezando por un análisis del primer requisito: la creación o incremento del riesgo. Este requisito tiene que ser valorado desde una perspectiva *ex ante* con el fin de determinar si efectivamente el peligro de producción del resultado es mayor a causa de la realización de la acción (Obregón y Gómez, 2023). De la misma manera, la jurisprudencia exige, cumulativamente, que el riesgo que ha generado la acción esté “jurídicamente desaprobado” o que no esté permitido. Así y en este sentido, se puede afirmar que en los casos donde este riesgo este permitido, no es factible la imputación objetiva del resultado. Es decir, no todo riesgo creado debe ser jurídicamente desaprobado puesto que la sociedad en determinadas circunstancias y contextos está dispuesta a tolerar el riesgo generado por determinadas conductas por considerarlas necesarias para el desarrollo de contactos sociales (Caro, 2023). Por ejemplo, la compraventa de un automóvil y su conducción. En el caso referido, el riesgo derivado de

la conducción es manifiesto, pero el legislador lo permite por cuanto tiene mayor peso el bien común. No obstante, el carácter permitido del riesgo no significa que sea causa de exclusión de la imputación objetiva. Siempre que la acción sea lo suficientemente peligrosa, el resultado será imputable objetivamente a la acción del sujeto (presuponiendo que se cumple la relación de causalidad). Otra cosa es que exista algún tipo de permiso legal que permita excluir la tipicidad de la conducta (Obregón y Gómez, 2023). Igualmente, se produce una exclusión de la imputación objetiva del resultado en aquellos casos en los que la conducta desarrollada por el autor no supone un riesgo suficiente *ex ante*, es decir, que el resultado no era previsible objetivamente (Llonín, 2022).

Lo descrito anteriormente es trasladable al campo de la Inteligencia Artificial y de la robótica tanto desde lo que respecta al conocimiento *ex ante* de un riesgo como también en lo que respecta a la asunción de un nivel determinado de riesgo. En relación con el conocimiento *ex ante*, en el supuesto de un ataque a los bienes jurídicos protegidos procedente del descontrol del dispositivo robótico, el principal problema sería especialmente de carácter probatorio con el fin de determinar si a la persona física (el programador o usuario) le era previsible objetivamente el resultado (Llonín, 2022). En lo que atañe a la asunción de un determinado nivel de riesgo permitido, a nivel europeo ya se discuten y exponen las ventajas e inconvenientes del desarrollo y utilización de sistemas de inteligencia artificial, elaborando una serie de niveles de riesgo atendiendo a su posible afeción a los derechos fundamentales (Llonín, 2022). En este sentido, cabe plantearse cuál es el nivel de riesgo asumible para la sociedad con respecto a los avances tecnológicos. La búsqueda del riesgo cero llevará inevitablemente a los programadores y desarrolladores a abandonar los proyectos relativos a IA. Es una consecuencia presumible puesto que pocos estarán dispuestos a asumir sanciones penales por los errores que se hayan producido en la programación de los sistemas de IA o por las decisiones tomadas por dichos sistemas que hayan evolucionado más allá del programa base. Por el contrario, no puede recaer en la sociedad la asunción de la totalidad del riesgo. El problema es balancear los intereses que hay en juego.

En relación con el segundo requisito: la correlación entre riesgo y resultado. Para que efectivamente se pueda afirmar la imputación objetiva, aparte de constatar la creación o incremento de un riesgo por la acción realizada, es necesario la existencia de una correlación entre riesgo y resultado. Esto es, que el resultado producido realiza el peligro inherente a la acción. Se está ante un juicio valorativo sobre la intensidad del peligro y su relación con el resultado (Obregón y Gómez, 2023). Es decir, en qué manera, por ejemplo, el programador por el mero hecho de haber creado el sistema de inteligencia artificial puede ser considerado responsable penal de una acción y decisión tomada por este sistema, teniendo en cuenta que su participación en la toma de la decisión es bastante cuestionable. Siguiendo con el ejemplo anterior, el problema evidente resulta en analizar de qué manera el programador por el mero hecho de haber creado la IA ha generado un riesgo suficiente con correlación evidente entre dicho riesgo generado y la decisión y acción realizada por la IA.

Descritos los elementos principales de la imputación objetiva, procede analizar los diferentes modelos de responsabilidad penal que la doctrina ha confeccionado para dar respuesta a estos problemas de imputación. Conviene señalar que los modelos desarrollados por la doctrina se centran en la imputación subjetiva para la atribución de responsabilidad penal, al tener como paradigma principal el dolo.

### 3.2. LOS MODELOS DE RESPONSABILIDAD PENAL

Conforme se ha venido explicando a lo largo del trabajo, resulta claro que los sistemas de IA autónomos conllevan un nivel de riesgo real y cierto y que pueden llegar a provocar daños. Es más, en la Declaración sobre Inteligencia artificial, robótica y sistemas “autónomos” del Grupo Europeo sobre Ética de la Ciencia y las Nuevas Tecnologías se puso de manifiesto que los sistemas de IA más avanzados no son transparentes dado que sus acciones, en gran parte, han dejado de ser programadas por seres humanos (2018). Se encuentran ejemplos de esta opacidad en casos como Google Brain que ha desarrollado una IA que a su vez es capaz de generar otras en menos tiempo y mejores o AlphaZero. Gracias a las redes generativas antagónicas, se hace posible que las máquinas se enseñen a sí mismas permitiendo incorporar en su análisis nuevas estrategias (Grupo Europeo sobre Ética de la

Ciencia y las Nuevas Tecnologías, 2018). Por ello, en estos casos, las acciones llevadas a cabo por las máquinas pueden convertirse en indiscifrables, escapando del control humano. Esto se debe a dos razones. Por un lado, suele ser imposible determinar cómo se han generado ciertos resultados más allá de los algoritmos iniciales (Grupo Europeo sobre Ética de la Ciencia y las Nuevas Tecnologías, 2018). Y por el otro lado, el desempeño de los robots o máquinas está basado en la información obtenida durante todo el aprendizaje pudiendo no estar disponible dicha información o datos que alberga (Grupo Europeo sobre Ética de la Ciencia y las Nuevas Tecnologías, 2018). Es por ello que, frente a los atentados más graves contra los bienes jurídicos protegidos, sea necesaria una intervención del Derecho Penal.

Quintero ha dado un paso para poder definir cómo se debe responder por los daños causados por un sistema de IA o máquina. Señala que no habría ningún problema para valorar la responsabilidad penal de los actos cometidos por los robots que han sido programados para ello (Quintero, 2017). Tampoco habría ningún problema en atribuir la responsabilidad penal por el daño causado a los bienes jurídicamente protegidos si quienes crean los sistemas de IA o robots sabían y aceptaban la posibilidad de que se desviarán de su tarea (aquí el autor entra en el terreno de la imputación subjetiva para atribuir la responsabilidad penal) (Quintero, 2017). Asimismo, Quintero pone de manifiesto que en los casos donde hay una desviación en la conducta previsible causando daños por razones totalmente imprevistas, habrá que calificarlos como un acontecimiento fortuito (Quintero, 2017). Por último, indica que no será posible invocar el principio de precaución con el fin de atribuir responsabilidad penal en aquellos casos donde el estado de la ciencia no permita predecir si la utilización de un robot puede causar daños o no (Quintero está acudiendo a la imputación subjetiva y a la infracción del deber de cuidado que afecta a la tipicidad) (Quintero, 2017). A todas luces, estas propuestas parecen simplificadores del hecho ante el que nos encontramos. Como señala Del Rosal, mantener que los daños causados por las máquinas debido a desviaciones totalmente impredecibles es un acontecimiento fortuito, supone “aceptar que unos daños pueden ser imputables a una conducta humana [ya sea la del programador o la del usuario], pero que el resultado no puede ser abarcado por su dolo o por su imprudencia” (2023, p. 14) (Del Rosal acude también a la imputación subjetiva). Sin embargo, un sistema de IA puede generar daños por muchas más razones que las condiciones ambientales como parece alegar Quintero,

pudiendo ser toda una incógnita la razón por la que ha causado dichos resultados. En consecuencia, el problema presente es de imputación de la conducta a los resultados lesivos, en tanto que existe la posibilidad de que no se pueda ni si quiera afirmar la relación de causalidad entre los resultados lesivos y la conducta humana. Por ello, se procede a analizar diversos modelos en aras de averiguar si efectivamente se resuelve este problema de imputación.

El principal proponente de los modelos de responsabilidad penal que se van a analizar es Gabriel Hallevy. Entre estos modelos, se encuentran aquellos que proponen la responsabilidad directa de sistema de IA y aquellos otros que rechazan esta posibilidad y ofrecen otras alternativas. Es importante destacar que la mayoría de los modelos que sigue la doctrina se centran principalmente en la imputación subjetiva y tienen como base para el desarrollo de estos a la Inteligencia Artificial general. Son modelos basados en una proyección hacia el futuro para resolver problemas que ahora todavía no existen. En cualquier caso, se van a examinar estos modelos propuestos, pero sin perder la vista de que utilizan como paradigma el dolo, aunque la mayoría de los problemas de IA pueden derivarse de imprudencias. Por tanto, si bien habrá que hacer mención de los elementos de imputación subjetiva para explicar los modelos propuestos, el objetivo principal es hacer un análisis de la imputación objetiva a partir de estos.

### *3.2.1. Modelos de responsabilidad penal no directa*

#### *3.2.1.1. El modelo de la autoría mediata*

Entre los modelos que propuso Hallevy, se encuentra el *The Perpetration-via-Another Liability Model*. Este modelo considera que los sistemas de IA no tienen ningún atributo humano, por lo que es considerado un agente inocente. De acuerdo con este modelo, la IA no tiene capacidad suficiente para ser considerada autor de un delito por cuanto, en palabras del propio Hallevy, las capacidades de la IA se asemejarían a: “the capabilities of a mentally limited person, such as a child, a person who is mentally incompetent, or one who lacks a criminal state of mind” (2010, p. 179). Este modelo supone aplicar los conceptos y

fundamentos de la autoría mediata que como expone el artículo 28 CP: “Son autores quienes realizan el hecho [...] por medio de otro del que se sirven como instrumento”. La IA, por tanto, es considerada un instrumento.

El problema surge en determinar quién efectivamente es el autor mediato que se sirve de la IA para cometer los delitos. Esto es fundamental en tanto en cuanto el autor mediato responde del hecho como si él mismo lo hubiera realizado, mientras el autor inmediato (en este caso la IA) no responde penalmente por cuanto, atendiendo a sus capacidades, concurre una causa de exención de la responsabilidad (Obregón y Gómez, 2023). Hallevy propone dos candidatos como posibles autores mediatos, el programador del sistema de IA o el usuario. Un programador podría diseñar la IA para cometer delitos. Así, a modo ejemplo, el programador podría diseñar el software de un robot operativo que es intencionalmente colocado en una fábrica y el software del robot está diseñado para incendiar la fábrica cuando no haya nadie. Si bien el robot ha incendiado la fábrica, el programador es el verdadero autor mediato del delito (Hallevy, 2010). Respecto del usuario, aunque no programa el sistema de IA, sí que lo usa para su propio beneficio. Póngase el ejemplo del usuario que compra un robot-servidor diseñado para ejecutar órdenes. Dicho robot asume al usuario como maestro y éste le ordena atacar a cualquier invasor de la casa. El robot, en consecuencia, ejecuta la orden. Igual que con el programador, el robot en este ejemplo comete el acto que atenta contra los bienes jurídicos protegidos, pero es el usuario el que es considerado autor mediato (Hallevy, 2010). En consecuencia, si un programador crea un sistema de IA para realizar hechos delictivos y lo usa a tal fin, el programador es autor mediato. Igualmente, si un usuario adquiere un sistema de IA sabiendo que puede ser utilizado para cometer delitos y decide usarlo a tal fin, se estaría de nuevo ante una autoría mediata del usuario.

Como se puede observar, se utilizan los factores de la imputación subjetiva como medio principal para resolver los problemas de atribución de responsabilidad penal que aquí se plantean, pero no se hace mención de los problemas de imputación objetiva. No obstante, no hay una verdadera problemática dado que la imputación objetiva, en estos casos, está clara. Si el programador no hubiera creado el sistema para cometer un delito, o si el usuario no hubiera dado la orden, no habría resultado. De manera que, efectivamente, se cumple la

relación de la causalidad. Asimismo, el programa diseñado y la orden dada crea un riesgo y existe una correlación entre riesgo y resultado puesto que el resultado producido ha realizado el riesgo inherente a la acción.

Sin embargo, es conviene resaltar que no parece preciso tener que acudir a la teoría de la autoría mediata para poder atribuir responsabilidad penal en estos casos. Bastaría con considerar a la IA como un instrumento del delito sin necesidad de atribuirle la condición de agente inocente inimputable como se pretende con este modelo. Atendiendo al estado actual de la ciencia, resulta lo más convincente. El programador o usuario, en este escenario, utiliza o se aprovecha de la IA para cometer delitos, siendo a estos a los que se les imputaría los resultados lesivos derivados de los hechos cometidos por la IA.

#### 3.2.1.2. El modelo de la llamada “consecuencia natural y probable”

El segundo sistema propuesto por Hallevy es el llamado *The Natural-Probable-Consequence Liability Model*. Este modelo asume una intervención profunda de los programadores o usuarios en las actividades diarias de un sistema de IA, pero sin intención alguna de cometer algún delito a través del mismo (Hallevy, 2010). Es el modelo aplicable a los casos donde la IA comete delitos que debieran haber sido previstos o que eran previsibles. Dicho de otra manera, los programadores o usuarios, en estos casos, no tenían conocimiento del hecho delictivo hasta que se ha cometido, sin que hayan participado en su planeamiento ni en ninguna parte del delito cometido (Hallevy, 2010). De nuevo, se recurren a los factores de imputación subjetiva para la atribución de responsabilidad penal. Hallevy explica este modelo a través de un ejemplo. Parte de un sistema de IA que ha sido creado para funcionar como un piloto automático. El sistema de IA está programado para proteger la misión como parte de su tarea de pilotar el avión. Llegado el momento, durante el vuelo, el piloto decide activar el piloto automático. En un momento dado, tras su activación, el piloto observa condiciones meteorológicas adversas e intenta volver a la base, abortando la misión (Hallevy, 2010). Sin embargo, la IA considera las acciones humanas como un atentado contra la misión que tiene que proteger y en consecuencia, toma las medidas necesarias para eliminar dicha amenaza contra la misión (es decir, el ser humano). Puede cortar el suministro de oxígeno o

eyectarlo del avión, etc. Como resultado, el piloto muere a causa de la acción de la IA (Hallevy, 2010). En virtud de este modelo, los programadores o usuarios deberían responder penalmente por su capacidad para prever la comisión de un delito por parte del sistema de IA.

Sin embargo, la doctrina de la consecuencia natural y probable es aplicada en el mundo anglosajón para imponer responsabilidad penal a los cómplices de los delitos que se derivan de una empresa delictiva o, dicho de otra manera, de las conspiraciones para cometer un delito cuando el delito original previsto cambia de alguna forma (Del Rosal, 2023). Es decir, se trata de aquellos casos donde varios individuos tienen la intención de cometer un delito, pero alguno de ellos acaba cometiendo un delito diferente o adicional. Por tanto, la responsabilidad penal de los cómplices (o cualquiera de las formas de participación) se extiende a los delitos no planeados que son razonablemente previsibles o que, en su virtud, son consecuencia natural de otro hecho delictivo (Del Rosal, 2023). Dicho de otra manera, se constituye una ficción legal a raíz de la cual se considera que el sistema de IA es cómplice del programador, siendo este responsable de aquellos actos que, si bien han sido ejecutados por la máquina, son consecuencia natural probable del plan que ha ideado (Morales, 2021).

No obstante, esta doctrina no tiene aplicación dentro del ordenamiento jurídico español para los casos de complicidad, puesto que se requiere el llamado doble dolo de la participación. Esto es, dolo en cuanto a la propia complicidad y dolo respecto del delito principal (Del Rosal, 2023). Por ello, para trasladar este modelo al ordenamiento jurídico, podrían utilizarse dos vías para la imputación de responsabilidad penal: la imprudencia o el dolo eventual de los programadores y usuarios.

Tal y como ha quedado reflejado, Hallevy recurre de manera reiterada a los factores de la imputación subjetiva para explicar su modelo de atribución de responsabilidad penal. Así, recurre, en este caso, tanto al dolo como la imprudencia. Ahora bien, dejando de lado la imputación subjetiva, se procede a hacer un análisis de la imputación objetiva a partir del modelo planteado.

Para empezar, en referencia a la imprudencia descrita en este modelo, es pertinente matizar algunos aspectos en relación con la tipicidad, puesto que en los tipos imprudentes se exigen unos requisitos diferentes a los tipos dolosos. En primer lugar, es necesario que se produzca una infracción del deber de cuidado exigible por parte del sujeto, que se produce cuando el comportamiento del sujeto no es acorde a las medidas prescritas para la realización de la acción peligrosa (Obregón y Gómez, 2023). En segundo lugar, para poder afirmar la imputación objetiva, es preciso que el resultado sea, en el momento de realizar la acción, previsible y evitable. Los supuestos de imprevisibilidad o inevitabilidad se darán cuando no sea apreciable una correlación suficiente entre resultado e intensidad del riesgo creado por la acción, o bien, cuando ni si quiera haya una relación de causalidad. Retomando el ejemplo anterior, para poder imputar objetivamente al programador a título de imprudencia el ilícito penal llevado a cabo por la IA (si se produjese la muerte del piloto o de algún pasajero) sería necesario afirmar que hubo una infracción del deber de cuidado al desarrollar el sistema de IA, es decir, que no actuó con la prudencia necesaria ni tomó todas las medidas prescritas para la realización de la acción (Hallevy, 2010). Cabe señalar que, en muchos casos, la decisión sobre si una conducta o acción peligrosa es diligente o no exige tomar en consideración normas consuetudinarias o sociales que no han sido objeto de formalización (Obregón y Gómez, 2023). En este caso, y atendiendo al estado actual de la ciencia, habría que preguntarse hasta qué punto llega la diligencia en el desarrollo de software y sistemas de inteligencia artificial. Adicionalmente, respecto del riesgo generado por la programación del sistema, sería necesario analizar hasta dónde llega el riesgo permitido en consonancia con la diligencia debida. Habría que preguntarse hasta qué punto es predecible por parte del programador el comportamiento de la IA que, si bien en los sistemas menos avanzados el problema es menor, en los que lleguen a una mayor autonomía, la solución se complica. Imagínese un sistema de IA programado en base a una serie de instrucciones a seguir, pero con el paso del tiempo evoluciona en su forma de actuar atendiendo a la experiencia adquirida. No solo se aleja la relación de causalidad en estos casos, sino que el riesgo creado es del todo imprevisible ante las diversas acciones que pueda cometer una IA de estas características.

Distinto es el caso del dolo eventual (si bien la frontera con la imprudencia es cada vez más difusa). En estos casos, dentro del ámbito de la imputación subjetiva, el sujeto no busca la realización de la acción típica, ni tampoco es consecuencia segura o inevitable de la conducta desarrollada, pero se la representa como probable y el sujeto asume su realización (Obregón y Gómez, 2023). A modo de ejemplo, piénsese en un sistema de IA diseñada para cometer un robo con violencia en un banco (Morales, 2021). El programador no busca causar ninguna muerte, pero sí que se le presenta como algo probable, previsible. Es más, lo está asumiendo al configurar el programa para que ejerza la violencia necesaria para conseguir el objetivo (Morales, 2021). En estos casos, atribuir responsabilidad penal a título de imprudencia parece poco conveniente, de ahí que se aplique el dolo eventual.

Por ello, en lo que respecta a la imputación objetiva si se trata de un delito doloso, habrá que estar a los requisitos ya explicados, siendo necesarios la relación de causalidad y el riesgo. El riesgo generado con la creación de un sistema dispuesto a utilizar violencia para la consecución de su objetivo supone crear o incrementar un riesgo y el peligro de esta acción efectivamente se realiza en el resultado.

En cualquier caso, es necesario diferenciar entre los supuestos donde existe una experiencia mínima que permita prever la posible producción de un resultado dañino, con independencia del índice de probabilidad, de aquellos en los que no se sabe que puede ocurrir quedando únicamente la resistencia científica a declarar la imposibilidad de que nada suceda (Morales, 2021).

#### 3.2.1.3. Modelos alternativos

Del Rosal propone dos modelos más a los propuestos por Hallevey: el modelo de responsabilidad por el producto y el modelo de responsabilidad objetiva.

El modelo de responsabilidad por el producto toma su base de la responsabilidad civil por el producto defectuoso. El Tribunal Supremo en reiteradas ocasiones ha precisado el marco jurídico en el que se encuadra la responsabilidad civil por el producto defectuoso. En

virtud de los requisitos que se marcan, se podrían hacer responsables civilmente en determinados supuestos a las empresas por los daños causados por los sistemas de IA o autómatas (Del Rosal, 2023). No obstante, este patrón de responsabilidad objetiva no tendría encaje si se aplica al conjunto de la imputación, tanto objetiva como subjetiva. Ahora bien, quizá sí que tendría cabida si se aplicase respecto de la primera, es decir, este carácter objetivo no la inhabilita como criterio de imputación objetiva. De tal manera que, conforme a este modelo, cualquier resultado derivado del funcionamiento de la IA sería objetivamente imputable al que pone el producto en el mercado. Sin perjuicio de tener que examinar si obró con dolo o imprudencia para poder atribuir responsabilidad penal.

En cualquier caso, en el ámbito penal, el encaje de la responsabilidad penal por los productos defectuosos no tiene fácil solución y ha generado notables controversias al respecto (Del Rosal, 2023).

Hay dos momentos cruciales en la responsabilidad penal por el producto. En primer lugar, la oferta en el mercado de los productos peligrosos/defectuosos que afectan a la salud pública y frente a lo cual el ordenamiento jurídico responde a través de los delitos contra la salud pública (artículos 359 y ss. del Código Penal). En segundo lugar, tras haber sido utilizados dichos productos lesionando la salud, vida o integridad física de las personas, reaccionando el ordenamiento jurídico mediante los delitos de homicidio o lesiones (Del Rosal, 2023).

No obstante, para poder aplicar este modelo es necesario que se pueda definir a los sistemas de IA como productos comerciales, lo cual es discutible. Además, es necesario la existencia de un defecto en el producto o que sus propiedades estén erróneamente representadas, cosa que, en el caso de la IA, es complejo de probar dado que la IA puede generar un daño sin que medie defecto alguno, al menos en el sentido que la responsabilidad del producto exige (Del Rosal, 2023).

En relación con el modelo de responsabilidad objetiva, Del Rosal toma como punto de partida el régimen de responsabilidad civil en materia de inteligencia artificial recomendado por el Parlamento Europeo (en adelante “PE”) en su resolución del 20 de octubre de 2020.

Conforme a dicha resolución, el PE considera oportuno establecer un régimen de responsabilidad objetiva para los que considera sistemas de inteligencia artificial de alto riesgo (en este sentido, el PE considera que los sistemas de IA autónomos pueden suponer en un alto riesgo para el público) (Del Rosal, 2023). Así, atendiendo a este régimen propuesto, será responsable objetivamente el operador del sistema de IA respecto de cualquier daño causado por acción física o virtual del sistema de IA (Del Rosal, 2023). No merece entrar en mayor profundidad respecto de este modelo puesto que no tiene encaje la responsabilidad objetiva dentro del sistema penal.

### 3.2.2. *Modelo de responsabilidad penal directa*

En los modelos anteriores la responsabilidad penal se acaba exigiendo a la persona que se encuentra detrás del sistema. No obstante, se dan los casos donde ni siquiera esas personas (programadores o usuarios) tienen la posibilidad objetiva de prever la producción de resultados lesivos. A estos casos, hay que añadir aquellos otros donde los sistemas de IA causantes del daño han sido programados por otro sistema de IA, sin que haya por tanto una persona humana detrás, sino otra máquina (Morales, 2021). Para intentar dar respuesta a estos casos, Hallevy introdujo el modelo de responsabilidad penal directa de los sistemas IA. Este modelo no asume ningún tipo de dependencia de la IA respecto un programador o usuario, centrándose en la propia IA (Morales, 2021). De tal manera que, no existiría ningún impedimento para reconocer a un sistema de IA como responsable a título individual de sus propios actos (Morales, 2021).

Hallevy, en este intento atribuir responsabilidad penal a los sistemas de IA, distingue entre los siguientes elementos del delito: elemento externo, *actus reus*, compuesto por la conducta, el resultado, la relación de causalidad y circunstancias externas que exigen determinados tipos; y el elemento interno, *mens rea*, en el que se encuentra subsumido el dolo y la negligencia (Del Rosal, 2023).

Respecto del *actus reus*, Hallevy considera que no es muy complicado su atribución a la IA. Así, estima que en tanto que la IA controle el mecanismo, mecánico o de otro tipo, para

mover sus partes móviles, entonces cualquier acto podría considerarse realizado por esta entidad (Hallevy, 2010). La propuesta de Hallevy retrotrae a un concepto causal naturalista de acción correspondiente al clasicismo propugnando por Bebel y Von Liszt. Conforme a esta teoría, la acción se define como un movimiento corporal (en este caso mecánico) reconducible a un impulso de la voluntad humana que provoca una modificación en el mundo externo perceptible por los sentidos (Obregón y Gómez, 2023). Se desprovee, por tanto, a la acción de cualquier contenido de voluntad. De modo que, la conducta si es considerada como una modificación externa del mundo exterior, entonces claramente la IA es capaz de satisfacer este requisito. En consecuencia, si un robot dotado de IA realiza cualquier tipo de movimiento y golpea a una persona, se puede considerar que se cumple el requisito de *actus reus* del delito de lesión. Igualmente, en los casos de omisión, si se impone a la IA una obligación de actuar y no actúa, el requisito de *actus reus* se cumple por omisión (Del Rosal, 2023).

Como se puede observar este concepto clásico de acción es superado por la regulación actual del ordenamiento jurídico español. Dos serían los elementos necesarios para poder hablar de una acción, que los hechos sean externos y voluntarios. Es más, hoy en día cualquier hecho no humano no es acción a efectos penales por no concurrir en ellos un ser humano de cuya voluntad pudieran depender (Obregón y Gómez, 2023). Por lo tanto, este sería el primer obstáculo al trasladar este modelo al ordenamiento jurídico español, por cuanto esta concepción puramente descriptiva y objetiva de la acción que propugna Hallevy no es predicable de las conductas tipificadas en el Código Penal. Sin embargo, el propio Mir Puig ya expresó: “que en nuestro Derecho penal el delito deba ser obra de un ser humano no se debe a razones ontológicas ni a la naturaleza de las cosas, sino a una decisión del Derecho positivo” (Obregón y Gómez, 2023, p. 55).

Los tipos delictivos cuentan con elementos que rodean la conducta, aunque no derivan de ella (a estos elementos Hallevy los denomina *circumstances*). En tanto que son circunstancias externas a la conducta, en los delitos cometidos por la IA también son satisfechos (Del Rosal, 2023). Igualmente, Hallevy considera que la imputación objetiva (*causation*) de resultados lesivos es atribuible a la conducta y no a la persona (en este caso la

IA) por lo que, bajo su perspectiva, el elemento de la relación de causalidad es fácilmente satisfecho (Del Rosal, 2023).

Más complejo es, por el contrario, que la IA satisfaga el elemento interno, *mens rea*, conocido como elemento mental integrado por el conocimiento y voluntad. Hallevy define el conocimiento como la percepción sensorial de datos fácticos y su comprensión (se estaría ante el elemento intelectual del dolo) (Hallevy, 2010). Atendiendo al desarrollo tecnológico actual, los sistemas de IA son capaces de percibir estos datos fácticos que son absorbidos y transferidos a unidades centrales de procesamiento que los analizan (Hallevy, 2010). En consecuencia, este primer sub-elemento es atribuible a la IA. Es decir, la IA tendría capacidad cognitiva, salvando las diferencias, para conocer su entorno y analizarlo.

En lo que afecta al elemento volitivo, hay que señalar que la intención conlleva la voluntad de llevar a cabo una acción calificada como delito además de la conciencia de realizar dicha acción. Esta voluntad supone un auténtico querer, no un mero deseo (Obregón y Gómez, 2023). Es aquí donde surge el mayor problema para atribuir el elemento volitivo a la IA. La voluntad se predica de un sujeto que sea capaz de tener sentimientos o estados mentales que le lleven a actuar de una determinada manera. Sentimientos tales como la envidia, el amor, el odio, etc., y que, en el estado actual de la ciencia, no se puede atribuir a ningún sistema de IA (Del Rosal, 2023).

Asimismo, dentro del *mens rea*, se presume la capacidad del sujeto de comportarse de manera distinta a como lo hizo recibiendo un reproche por su actuación ilícita, en tanto que se ha demostrado que el sujeto podría haber actuado de conformidad con la ley (Del Rosal, 2023). Como se puede observar, dentro del elemento mental al que hace referencia Hallevy, se estructura lo que en la teoría del delito es el dolo (con su elemento intelectual y volitivo) y la imputabilidad. Ambos elementos fundamentales para predicar la culpabilidad en la conducta del sujeto y necesarios para que surja la responsabilidad penal por la acción desarrollada.

Ciertos sistemas de IA, como ya se ha indicado, están siendo desarrollados a través de *deep learning*, dotados de redes neuronales artificiales que les permite identificar y evaluar diferentes escenarios y tomar decisiones en base a los mismos. Además, el *machine learning* facilita un aprendizaje inductivo a los sistemas de IA, adquiriendo experiencia y aprendiendo de la misma (Del Rosal, 2023). Lógicamente, la adquisición de nuevos conocimientos en base a dichas experiencias afectará a su toma de decisiones. De modo que, un sistema de IA fuerte o general podrá tomar decisiones y ejecutarlas tras un análisis y evaluación de las distintas posibilidades y formas de conducta (Del Rosal, 2023). En consecuencia, si una acción del sistema de IA es constitutiva de delito, entonces se podrá asumir que la IA tenía intención de cometerlo. Esto se debe a que, si la máquina tiene capacidad de analizar la probabilidad con mayor precisión que un ser humano, razón de más para concluir que la IA era consciente de su actividad delictiva (Del Rosal, 2023). Del Rosal llega a esta conclusión expresando: “el aspecto volitivo puede cumplirlo un ordenador siempre que esté dotado de un sistema de redes neuronales artificiales (*deep learning*), es decir, un sistema de IA de los que hemos denominado fuerte” (2023, p. 18).

El modelo propuesto por Hallevy es lo más cercano a una humanización de la IA. Es más, para poder hacer posible esta modelo sería indispensable atribuir algún tipo de personalidad jurídica a los sistemas de IA. No es reciente el planteamiento de atribuir personalidad a los entes artificiales. El Parlamento Europeo, en su resolución de 16 de febrero de 2017 con recomendaciones destinadas a la Comisión sobre normas de Derecho Civil sobre robótica, abordó la posibilidad de dotar a los robots de personalidad jurídica manifestando que “cuanto más autónomos sean los robots, más difícil será considerarlos simples instrumentos en manos de otros agentes (como el fabricante, el operador, el propietario, el usuario, etc.)” (p.5) por ello, “la autonomía de los robots suscita la cuestión de su naturaleza y de si pertenecen a una de las categorías jurídicas existentes o si debe crearse una nueva categoría con sus propias características jurídicas” (p.5). Atendiendo a estas razones, el Parlamento Europeo solicitó a la Comisión Europea crear:

una personalidad jurídica específica para los robots, de forma que como mínimo los robots autónomos más complejos puedan ser considerados personas electrónicas responsables de reparar

los daños que puedan causar, y posiblemente aplicar la personalidad electrónica a aquellos supuestos en los que los robots tomen decisiones autónomas inteligentes o interactúen con terceros de forma independiente (2017, p. 17)

En cualquier caso, el modelo propuesto por Hallevy implicaría que una vez alcancen los sistemas de IA el aprendizaje automatizado, pudiendo hablar propiamente de una IA general o fuerte, responderán por las decisiones que tomen y por las consecuencias que de estas se deriven. No obstante, atendiendo al estado actual de la ciencia, de momento resulta inviable. En palabras de Del Rosal: “es imposible poder afirmar que un autómatas o un sistema de IA tenga capacidad de acción, capacidad de culpabilidad y capacidad de pena” (2023, 19). Cabe señalar que tampoco se predicaba de las personas jurídicas antes de la reforma del Código Penal en el año 2010. Aunque son casos distintos. Si bien es cierto que la persona jurídica se constituye como centro de imputación de determinadas decisiones, las mismas son tomadas por personas físicas; mientras que en el caso de los sistemas de IA, como se viene explicando, es el propio sistema quien toma las decisiones de manera autónoma para las que en principio no está programado y sin que intervenga el ser humano.

El modelo que aquí se ha expuesto implica la atribución de personalidad a los sistemas IA como si de personas físicas se trataran. Partiendo de este punto no cabe hacer un análisis exhaustivo de la imputación objetiva por cuanto habría que estar a cada caso concreto. Bastaría con apreciar los requisitos ya mencionados para imputar el resultado a la acción u omisión del sistema de IA.

#### **4. LAS PENAS Y LA INTELIGENCIA ARTIFICIAL**

Dando por hecho que el modelo de autoría directa es aceptado y se otorgase a los sistemas de IA personalidad jurídica, resulta necesario hacer un análisis de las penas que pudiesen ser de aplicación a estos. El problema no es baladí. Si atendiendo al delito cometido, la pena apropiada es una pena privativa de libertad, hay que preguntarse cómo puede un sistema de IA cumplir semejante pena, especialmente si dicho sistema de IA no está instalado en un robot (o cualquier otro cuerpo físico). De la misma manera, habría que considerar como se

puede imponer la pena de multa o, incluso, la libertad condicional. Estos problemas ya surgieron en su momento respecto de las personas jurídicas. Pero, a diferencia de los sistemas de IA, en el núcleo de las personas jurídicas, hay personas físicas. Es por ello que el artículo 31 bis del CP establece la responsabilidad penal de las personas jurídicas en los siguientes términos: “De los delitos cometidos en nombre o por cuenta de las mismas, y en su beneficio directo o indirecto, por sus representantes legales o por aquellos que actuando individualmente o como integrantes de un órgano de la persona jurídica [...]”. Dicho de otra manera, la imputación de responsabilidad penal a la persona jurídica es consecuencia de la comisión de un delito por parte de una persona física que mantiene una vinculación especial con la persona jurídica. Está fundado, aunque sea parcialmente, en una responsabilidad por el hecho ajeno (Obregón y Gómez, 2023). En cualquier caso, parece simple la solución para las personas jurídicas por cuanto si se puede imponer una pena en los mismos términos que a las personas físicas, entonces no es necesario realizar ningún ajuste (Hallevy, 2010). Así, cuando se impone una multa a la persona jurídica, ésta la paga de la misma manera que la persona física paga su multa. El conflicto surge en aquellos casos donde no se puede aplicar la misma pena que se aplica a los seres humanos siendo necesario un ajuste como se efectuó con el establecimiento de diversas penas para las personas jurídicas. Esta es la situación en la que se encontrarían los sistemas de IA.

Las consideraciones sobre el ajuste de las penas son examinadas en profundidad por Hallevy aplicando los fundamentos teóricos de cualquier pena. Así, en su análisis, menciona tres etapas que se pueden trasladar a las siguientes preguntas: (i) ¿Cuál es el significado de la pena para el ser humano? (ii) ¿Cómo afectan dichas penas a los sistemas de IA? y (iii) ¿Qué penas pueden alcanzar el mismo significado cuando se imponen a entidades de IA? (2010)

Hallevy señala que el significado de las penas para los sistemas de IA sería el mismo que para las personas físicas haciendo especial referencia a las siguientes penas: pena de muerte, prisión, suspensión de la pena, trabajos en beneficio de la comunidad y multas.

Respecto de la pena de muerte, hay que partir del hecho de que en España está prohibida por la Constitución de 1978 en su artículo 15. Hallevy refleja la falta del consenso en las diversas jurisdicciones en cuanto a su constitucionalidad (2010). No obstante, señala que, efectivamente, es la forma más eficaz de incapacitar a los criminales en relación con la posible reincidencia pues, una vez ejecutada, el delincuente claramente no puede volver a cometer ningún crimen (Hallevy, 2010). En cualquier caso, la pena de muerte implica la pérdida de la vida para los seres humanos. El traslado de esta pena a los sistemas de IA puede tener cierto encaje. Esto se debe a que la vida para los sistemas de IA es su existencia independiente como entidad pudiendo ser una existencia abstracta (en caso de ser el software instalado en un servidor de red) o física (por ejemplo, un robot) (Hallevy, 2010). La pena de muerte para un sistema de IA, teniendo los mismos efectos, esto es, incapacitarla para que no pueda cometer más acciones delictivas, sería el borrado del software (Hallevy, 2010). La erradicación del software supondría el cese de la existencia independiente del sistema de IA.

Respecto a las penas privativas de libertad, el artículo 35 del CP hace referencia a “la prisión permanente revisable, la prisión, la localización permanente y la responsabilidad personal subsidiaria por impago de multa”. Sin embargo, procede hacer especial hincapié en este punto a las penas de prisión. Las penas de prisión implican la privación de la libertad para las personas físicas, además de la imposición de limitaciones a su libre comportamiento, movimiento y libertad para autogestionarse (Hallevy, 2010). La libertad para un sistema de IA es poder actuar como tal en un área determinada, como por ejemplo un robot en el ámbito médico tiene la libertad de participar en cirugías o un sistema de IA en una fábrica tiene libertad para fabricar (Hallevy, 2010).

Hallevy propone que para conseguir los mismos efectos que el encarcelamiento para un sistema de IA habría que inutilizar al sistema durante un periodo determinado de tiempo (2010). De tal manera que, durante dicho periodo, el sistema de IA no puede actuar constriñendo así su libertad de actuación.

En lo que atañe a la suspensión de la pena privativa de libertad, hay que indicar que es un beneficio que se concede al reo, pero bajo una serie de condiciones (como se dispone en

el artículo 80 CP). Es más, esta medida puede ser revocada conforme a lo dispuesto en el artículo 86 del CP. Entre los motivos que pueden llevar a revocar esta medida y ejecutarse la pena prevista es la comisión de otro delito durante el periodo de suspensión. Es esta amenaza de revocación lo que Hallevy considera como el verdadero significado de esta medida (2010). Dicho de otra manera, la amenaza de revocación de la suspensión de la pena disuade a las personas físicas de cometer un delito. Además, cuando se acuerda la suspensión de la pena, no se lleva a cabo ninguna acción sobre la persona como tal (Hallevy, 2010). En este sentido, Hallevy considera que se podría aplicar de la misma manera a los sistemas de IA sin que requiera hacer un esfuerzo mayor para su adaptación.

En relación con los trabajos en beneficio de la comunidad, el significado para los humanos de esta pena no es otro que la contribución obligatoria de mano de obra a la comunidad (Hallevy, 2010). Hallevy considera que la imposición de esta pena a un sistema de IA no parece tener mayor complicación por cuanto él mismo puede ser empleado como trabajador en diversos ámbitos. Así, cuando una entidad de IA trabaja en una empresa o fábrica, el trabajo que desarrolla es en beneficio de los dueños de la empresa o fábrica o quizá en beneficio del resto de trabajadores al facilitar el desarrollo de sus tareas (Hallevy, 2010). Por ello, del mismo modo que puede trabajar en beneficio de determinadas personas, también podría ser utilizada para el beneficio colectivo (Hallevy, 2010). Solo faltaría determinar qué tipo de trabajo podría desarrollar una IA para el cumplimiento de esta pena.

Finalmente, queda hablar de las penas de multa. Este tipo de pena implica la privación de propiedad a las personas físicas o jurídicas, ya sea dinero u otra propiedad (decomiso). La imposición de la pena de multa a una persona jurídica es idéntica a la imposición de una multa a una persona física, puesto que tanto unos como otros son propietarios de inmuebles o titulares de cuentas bancarias (Hallevy, 2010). Por ello, el pago de la multa no varía. Sin embargo, los sistemas de IA, en principio, no son propietarios ni titulares de cuentas bancarias por lo que habría una mayor dificultad para aplicar esta pena (Hallevy, 2010). Si por el contrario lo fuesen, su aplicación sería conforme a lo expuesto. En este sentido, quizá sería posible otorgar titularidad a la IA de los beneficios de su actividad autónoma (por ejemplo, derechos de autor).

En líneas generales, la mayoría de las penas explicadas serían aplicables a un sistema de IA si su personalidad fuese reconocida. De modo que, imponer estas penas a estas entidades de IA no negaría la propia naturaleza de estas penas en comparación con su imposición a las personas físicas o jurídicas; sin perjuicio de los ajustes que fuesen precisos y necesarios para hacer estas penas efectivas.

Sin embargo, resulta necesario y conveniente hacer un análisis de las finalidades de la pena para comprobar si tienen encaje cuando las penas se imponen a los sistemas de IA. Las teorías sobre la finalidad de las penas se suelen dividir en tres grandes grupos: absolutas, relativas y mixtas. Estas teorías versan sobre las dos perspectivas que se vienen planteando al respecto: el fin de la pena es castigar porque se ha cometido un delito, o bien, el fin de la pena es prevenir que se cometan delitos en el futuro.

Las teorías absolutas se caracterizan por no asignar ninguna finalidad a la pena más allá del propio castigo. La pena se entiende como una reacción necesaria contra el delito cometido imponiendo un mal (privación de bienes jurídicos) con el objetivo de compensar un mal previo (el delito). Dicho de otra manera, la pena no se justifica por sus consecuencias sociales, su utilidad o incluso su necesidad, sino por la exigencia de justicia de quien ha cometido un delito reciba lo que se merece (Peñaranda y Basso, 2019). Por tanto, la pena es un fin en sí misma atendiendo al aforismo latino *punitur, quia peccatum est*. Los autores adheridos a estas teorías hablan de dos fines de la pena, la expiación o la retribución. La retribución ha ganado mayor arraigo (Peñaranda y Basso, 2019). Los partidarios destacados de esta concepción de la pena y que mejor fundamentan este paradigma son los filósofos alemanes Kant y Hegel. Es más, doctrinalmente se considera que la obra de Kant, *Metafísica de las Costumbres* es la mejor exponente de esta teoría proporcionando una fundamentación ética a la justificación absoluta de la pena (Peñaranda y Basso, 2019). Así, “según la interpretación más generalizada de su concepción, para Kant solo es admisible fundar la pena en el merecimiento (demérito) del delincuente por el hecho cometido (*quia peccatum est*)” (Peñaranda y Basso, 2019, p. 166).

Por tanto, estas teorías están caracterizadas por atribuir a la pena un fin retributivo abandonando cualquier consideración utilitaria. De modo que, si el objetivo es conseguir justicia, la pena entonces tendrá que ser proporcional al injusto culpable atendiendo, en todo caso, a la gravedad del hecho y la culpabilidad del sujeto (Peñaranda y Basso, 2019). En consecuencia, se desprende una relación de igualdad entre la pena y la gravedad del mal cometido.

Las teorías relativas a diferencia de las absolutas se caracterizan por atribuir a la pena una función o finalidad más allá del propio castigo. Se atribuye a la pena una finalidad utilitaria consistente en la evitación o prevención de futuros delitos (*punitur ut ne peccetur*) (Peñaranda y Basso, 2019). Estas teorías “fundamentan el castigo en sus consecuencias sociales, en necesidades de prevención, en la protección de bienes jurídicos y/o en su utilidad para los individuos o para la comunidad” (Peñaranda y Basso, 2019, p. 169). Se puede distinguir entre la prevención general y la prevención especial.

La prevención general implica que la finalidad preventiva actúa sobre la colectividad. El fin de la pena radica en evitar futuros delitos incidiendo principalmente mediante la amenaza de una pena y, en su caso, su imposición y cumplimiento, no solo a la propia persona física que deberá sufrirla sino sobre el conjunto de la sociedad (Peñaranda y Basso, 2019). Dentro de la prevención general, hay dos grandes tendencias: prevención general negativa y prevención general positiva. La función de la pena en las teorías de prevención general negativa es disuadir a la comunidad de cometer delitos por miedo o temor a ser castigados (Peñaranda y Basso, 2019). Está basada, por tanto, en la intimidación. En base a esta tendencia, la ley compele a su cumplimiento bajo la amenaza de ser sancionados con una pena; en consecuencia, la generalidad se ve disuadida de cualquier tipo de intención criminal al ejercer sobre la misma una coacción psicológica (Peñaranda y Basso, 2019).

La prevención general positiva, por el contrario, puede albergar varios significados. Por un lado, se puede poner el acento en la capacidad de la pena para incidir en la ciudadanía a través del aprendizaje y de otros mecanismos de psicología profunda (Peñaranda y Basso, 2019). Es decir, se entiende que la ley penal cumple con una función pedagógica, incluso

motivadora, con el fin de fomentar en el ciudadano una actitud de respeto hacia las normas jurídicas (Peñaranda y Basso, 2019). Por el otro lado, se puede entender que la principal función atribuida a la pena es mantener la confianza en las normas influyendo de esta manera en otros procesos de control que tienen como objetivo preservar la integración y cohesión social. En este sentido, la convivencia genera unas expectativas, una confianza en que el comportamiento humano se desarrollara de manera regular; la norma implica la garantía de esas expectativas y su infracción es una decepción de estas (Peñaranda y Basso, 2019). Así, la pena es la réplica ante dicha infracción reforzando la confianza social en el ordenamiento jurídico.

En la prevención especial, la pena no se dirige a la comunidad sino a la persona determinada que ha cometido el delito con el fin de evitar que vuelva a incurrir en él (Peñaranda y Basso, 2019). De la misma manera que la prevención general, la prevención especial puede tener diversas interpretaciones. Puede tener efectos correctivos, procurando una transformación moral del delincuente implantando motivos prosociales o incluso altruistas mediante la ejecución de la pena (Peñaranda y Basso, 2019). También se puede interpretar que tiene efectos de intimidación especial e individualizada, de modo que la pena como medio de privación de bienes jurídicos frene la tentación del delincuente a reincidir (Peñaranda y Basso, 2019). Otra interpretación es entender que la prevención especial tiene como efecto eliminar al delincuente (no tiene por qué ser una supresión física de la persona, puede implicar su destierro) o inocuizarlo (expulsión o asilamiento de la sociedad) mediante la ejecución de las penas (Peñaranda y Basso, 2019). Por último, también hay aquellos que, en virtud de la prevención especial, otorgan a la pena una finalidad resocializadora, atendiendo a los factores que generan o fomentan la comisión de delitos e incidiendo sobre ellos en relación con el sujeto que los comete; con el fin de que tras la ejecución de la pena y cumplimiento de la misma, el sujeto pueda reintegrarse en la sociedad (Peñaranda y Basso, 2019). Esta última finalidad ha estado en el centro de atención de los ordenamientos jurídicos y, en especial, en España, dado que tal y como dispone el art. 25. 2 de la CE: “Las penas privativas de libertad y las medidas de seguridad estarán orientadas hacia la reeducación y reinserción social [...]”.

Por último, y como corolario de estas teorías sobre los fines de las penas, están las teorías mixtas. Sin entrar en mayor detalle, estas teorías concilian las posiciones ya expuestas combinando de formas alternativas aspectos de retribución y prevención (ya sea general o especial). Así, la pena debe ser tanto utilitaria como justa debiéndose dirigir hacia la prevención de nuevos delitos, aunque no puede ir más allá de la culpabilidad del infractor (Peñaranda y Basso, 2019).

Como se ha expuesto, el principal objeto de estas teorías es el infractor como persona física que tiene unos sentimientos y procesa unas emociones. No obstante, no es comparable la persona física con la posible personalidad que se reconozca a los sistemas de IA. Si bien estos pueden llegar a tener autonomía tomando sus propias decisiones en base a su experiencia y aprendizaje, no parece lógico pensar que vayan a ser capaces de procesar emociones de la misma forma que los seres humanos, al menos en un futuro próximo.

Las teorías relativas utilizan la intimidación, el temor, la confianza social o la resocialización como elementos para dar una finalidad a la pena. Cabe preguntarse de qué manera procesaría el temor la IA o se vería intimidada; cómo se vería coaccionada psicológicamente para no cometer delitos. De la misma manera, la confianza que se desprende del ser humano y las expectativas que ésta genera, no parece que se pueda también predicar del comportamiento o acciones que desarrollen los sistemas de IA. Igualmente, resulta complicado aplicar la finalidad de la resocialización a la IA. Por tanto, resulta lógico asumir que la gran mayoría de las teorías relativas no sirven como fundamentación de la finalidad de las penas aplicadas a los sistemas de IA. A excepción de las teorías de prevención especial, si se interpreta la finalidad de la pena como puramente reductora o eliminadora de la peligrosidad criminal (la prevención especial de aislamiento, inocuización o eliminación).

Las teorías absolutas, por el contrario, no asignan ninguna finalidad a la pena más allá del propio castigo justificándose en la exigencia de justicia. No entran a valorar la afección de la pena en los procesos internos de la persona física (no se entra a valorar la intimidación o temor que provoca en la persona). Por ello, parece esta teoría como la más apropiada para fundamentar la finalidad de la pena respecto de los sistemas de IA (a excepción de la

prevención especial mencionada). La pura retribución como finalidad de la pena respecto de la IA.

## 5. CONCLUSIONES

El desarrollo de la Inteligencia Artificial ha traído consigo numerosos beneficios, pero también plantea numerosos retos e incluso riesgos a los que se tiene que hacer frente. Estos retos son especialmente relevantes en el ámbito del Derecho Penal. Este Trabajo de Fin de Grado trata de poner de manifiesto la complejidad de atribuir responsabilidad penal por los hechos cometidos por la IA. El enfoque general de la doctrina, en relación con este problema, resalta los problemas de imputación subjetiva (como ocurre en los modelos expuestos), pero esto entraña un escenario que actualmente no existe (al fundamentarse en la IA fuerte). Por esta razón, se ha optado por centrar este trabajo en los problemas asociados a la imputación objetiva de resultados lesivos de los hechos cometidos por sistemas de Inteligencia Artificial, en aras de determinar, a quién se le debe imputar las acciones de un robot o agente artificial que deriven en un daño a un bien jurídico.

Se ha procedido a realizar un análisis general de los elementos de la imputación objetiva y su interrelación con los sistemas de IA. De este análisis se derivan una serie de reflexiones y conclusiones. Respecto a la relación de causalidad, el principal problema que se ha identificado ha sido la multiplicidad de las personas que intervienen (programador, usuario o incluso terceras personas) y la distancia entre la instrucción con la que se programa la IA y el resultado lesivo que se produce. No obstante, en virtud de la teoría de la *condicio sine qua non*, en estos casos, si se elimina cualquier intervención, el resultado desaparece, por lo que no parece que haya que centrarse exclusivamente en los problemas de causalidad.

Consecuentemente, merece especial atención el elemento del riesgo de la imputación objetiva. Se ha analizado el mismo atendiendo a sus dos factores fundamentales: creación o incremento del riesgo y la correlación entre riesgo y resultado. En lo que afecta a la creación o incremento del riesgo, se ha podido concluir en relación con el conocimiento *ex ante*, que

el problema, en los casos de daños derivados del descontrol del dispositivo robótico o sistema de IA, es principalmente probatorio en atención a determinar si a la persona física (el programador o usuario) le era previsible objetivamente el resultado. En lo referente a la asunción de un determinado nivel de riesgo, el mayor inconveniente que se plantea es conseguir determinar cuál es nivel de riesgo aceptable y asumible por la sociedad, en tanto que, si se busca el riesgo cero, esto provocará un abandono de los proyectos relativos a IA por parte de los programadores y desarrolladores. Es una consecuencia presumible puesto que, pocos estarán dispuestos a asumir sanciones penales por los errores que se hayan producido en la programación de los sistemas de IA, o por las decisiones tomadas por dichos sistemas que hayan evolucionado más allá del programa base.

En lo que atañe a la correlación entre riesgo y resultado, habría que analizar cada caso concreto. A modo ejemplo, el problema reside en poder determinar si el programador, por el hecho de haber creado un sistema de IA, ha generado un riesgo suficiente con correlación evidente entre dicho riesgo generado y la decisión y acción realizada por la IA generadora de daños.

Asimismo, se han analizado los principales modelos propuestos por la doctrina de atribución de responsabilidad penal por los hechos cometidos por la IA. Como se ha indicado, estos modelos son en gran medida una proyección de futuro para problemas que hoy en día todavía no existen dado que tienen como base a la inteligencia artificial general. Además, los modelos que propone la doctrina se centran en la imputación subjetiva al tener como paradigma principal el dolo y no entran a analizar de manera pormenorizada los problemas de imputación objetiva. Sin embargo, sirven de base para poder hacer un análisis más detallado sobre a quién se le puede imputar objetivamente los resultados lesivos.

Estos modelos se pueden dividir en dos grupos: modelos de responsabilidad penal no directa (autoría mediata, consecuencia natural y probable y modelos alternativos) y directa. Se ha llevado a cabo un análisis de la imputación objetiva en relación con cada uno de estos modelos derivándose una serie conclusiones y reflexiones.

En el modelo de autoría mediata, la imputación objetiva no presenta grandes problemas. Sería suficiente con afirmar que, a consecuencia de la programación de la IA o el uso que se le ha dado, se ha provocado un resultado lesivo. No obstante, no parece necesario tener que recurrir a la teoría de la autoría mediata para poder atribuir responsabilidad penal. Bastaría con considerar a la IA como un instrumento del delito sin necesidad de atribuirle la condición de agente inocente inimputable como se pretende con este modelo. Atendiendo al estado actual de la ciencia, resulta lo más convincente. El programador o usuario, en este escenario, utiliza o se aprovecha de la IA para cometer delitos, siendo estos a los que se les imputaría los resultados lesivos derivados de los hechos cometidos por la IA.

Para poder resolver los problemas de imputación objetiva en el modelo de consecuencia natural y probable, es necesario distinguir si se está ante un delito imprudente o doloso. En el caso de estar ante un delito imprudente, habrá que tener en cuenta que es necesario que se produzca la infracción de un deber de cuidado y, además, para afirmar la imputación objetiva es necesario que el resultado sea previsible y evitable. Llegados a este punto surgen diversas problemáticas. Habría que preguntarse hasta qué punto llega la diligencia en el desarrollo de software y sistemas de inteligencia artificial; y respecto del riesgo generado por la programación del sistema, sería necesario analizar hasta dónde llega el riesgo permitido en consonancia con la diligencia debida. Respondiendo a estas preguntas se podría aclarar con mayor facilidad la posibilidad de imputar los actos cometidos por las IA a los programadores o usuarios. Si se estuviese ante un delito doloso, solo sería necesario que se cumpliesen con los requisitos exigidos en la imputación objetiva para poder imputar los hechos cometidos por la IA a los programadores o usuarios.

Se han tratado también diferentes modelos alternativos de responsabilidad objetiva siendo el de mayor transcendencia el de responsabilidad por el producto. En este modelo, desde un principio, parece claro la imposibilidad de su aplicación para resolver los problemas de atribución de responsabilidad penal. No obstante, esto solo es cierto si se exige una responsabilidad objetiva en todo el ámbito de imputación (tanto objetiva como subjetiva). Ahora bien, quizá sí que tendría cabida si se aplicase respecto de la primera (imputación objetiva), es decir, este carácter objetivo no la inhabilita como criterio de imputación objetiva.

De tal manera que, conforme a este modelo, cualquier resultado derivado del funcionamiento de la IA sería objetivamente imputable al que pone el producto en el mercado. Sin perjuicio de tener que examinar si obró con dolo o imprudencia para poder atribuir responsabilidad penal.

En el modelo de responsabilidad penal directa, la imputación objetiva no presenta ningún problema específico dado que al considerar a los sistemas de IA como si de personas físicas se trataran, habría que estar a cada caso concreto. Dicho de otra manera, bastaría con apreciar los requisitos de la imputación objetiva para atribuir los resultados lesivos al propio sistema de IA.

Finalmente, en el último apartado de este TFG se han tratado las penas y su finalidad aplicadas a los sistemas de IA. Partiendo de que se reconociese personalidad jurídica a los entes de IA, se puede concluir que, en términos generales, las mismas penas que se aplican a las personas físicas podrían ser aplicables a los sistemas de IA. Esto se debe a que no se niega la propia naturaleza de estas penas en comparación con su imposición a las personas físicas o jurídicas. Todo ello sin perjuicio de los ajustes que fuesen precisos y necesarios para hacer estas penas efectivas.

En lo que respecta a la finalidad de las penas aplicadas a la IA, resulta necesario aclarar que, si bien los sistemas de IA pueden llegar a tener autonomía tomando sus propias decisiones en base a su experiencia y aprendizaje, no parece lógico pensar que vayan a ser capaces de procesar emociones de la misma forma que los seres humanos, al menos en un futuro próximo. En consecuencia, se considera que las teorías que mejor fundamentan la finalidad de estas, son principalmente las teorías absolutas, dado que no asignan ninguna finalidad a la pena más allá del propio castigo, justificándose en la exigencia de justicia. Además, no entran a valorar la afección de la pena en los procesos internos de la persona física. Sin embargo, también podrían servir de fundamento las teorías de prevención especial, si se interpreta la finalidad de la pena como puramente reductora o eliminadora de la peligrosidad criminal (la prevención especial de aislamiento, inocuización o eliminación).

## 6. BIBLIOGRAFÍA

- **LEGISLACIÓN**

Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal, disponible en <https://www.boe.es/buscar/act.php?id=BOE-A-1995-25444>.

Ley Orgánica 1/2019, de 20 de febrero, por la que se modifica la Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal, para transponer Directivas de la Unión Europea en los ámbitos financiero y de terrorismo, y abordar cuestiones de índole internacional, disponible en [https://www.boe.es/diario\\_boe/txt.php?id=BOE-A-2019-2363](https://www.boe.es/diario_boe/txt.php?id=BOE-A-2019-2363).

Propuesta de Reglamento del Parlamento Europeo y del Consejo 2021/0106, por el que se establecen normas armonizadas en materia de inteligencia artificial (ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión, disponible en <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A52021PC0206>.

- **OBRAS DOCTRINALES**

AI Safety Summit. (2023). Declaración de Bletchley.

Amador, L. (1996). *Inteligencia Artificial y Sistemas Expertos*. Universidad de Córdoba.

Blanco, I. (2019). *Homo Sapiens y ¿Machina Sapiens?: Un Derecho Penal para los robots dotados de Inteligencia Artificial*. En C. Mallada (Ed.). *Nuevos retos de la ciberseguridad en un contexto cambiante* (63-80). Thomson Reuters Aranzadi

- Blauth, T. F., Gstrein, O. J., y Zwitter, A. (2022). Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI. *IEEE Access*, 10, 77110-77122. <https://doi.org/10.1109/ACCESS.2022.3191790>.
- Caldwell, M., Andrews, J. T. A., Tanay, T. y Griffin L. D. (2020). AI-enabled future crime. *Crime Science*, 9, 14. <https://doi.org/10.1186/s40163-020-00123-8>.
- Caro, J.A. (2023). Algunas consideraciones sobre el riesgo permitido en el Derecho penal. *Forseti-Revista de Derecho*, 12(18), 41-66.
- Comisión Nacional de los Mercados y de la Competencia (CNMC). (s.f.) Inteligencia Artificial y Competencia.
- Comité Económico y Social Europeo. (2017). *Inteligencia artificial: las consecuencias de la inteligencia artificial para el mercado único (digital), la producción, el consumo, el empleo y la sociedad* (C-288).
- Consejo de la Unión Europea. (9 de diciembre, 2023). Reglamento de Inteligencia Artificial: el Consejo y el Parlamento alcanzan un acuerdo sobre las primeras normas del mundo en materia de inteligencia artificial [Comunicado de Prensa]. <https://www.consilium.europa.eu/es/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/>
- De la Cuesta Aguado, P. (2019). Inteligencia artificial y responsabilidad penal. *Revista Penal México*, 16-17, 51-62. <https://dialnet.unirioja.es/servlet/articulo?codigo=7675764>.
- Del Rosal, B. (2023). ¿El modelo de la responsabilidad penal de las personas jurídicas para los daños punibles derivados del uso de la inteligencia artificial? *Revista Electrónica de Responsabilidad Penal de Personas Jurídicas y Compliance*, 2(2) <https://dialnet.unirioja.es/servlet/articulo?codigo=9148639>.

- Diamantis, M.E. (2023). Employed Algorithms: A Labor Model of Corporate Liability for AI. *Duke Law Journal*, 72, 797-859. <https://scholarship.law.duke.edu/dlj/vol72/iss4/2/>.
- Dobrinou, M. (2019). The Influence of Artificial Intelligence on Criminal Liability. *Challenges of the Knowledge Society*, 48-52. <https://www.proquest.com/scholarly-journals/influence-artificial-intelligence-on-criminal/docview/2263228271/se-2>
- Dresp-Langley, B. (2023) The weaponization of artificial intelligence: What the public needs to be aware of. *Frontiers in Artificial Intelligence*, 6. <https://www.frontiersin.org/articles/10.3389/frai.2023.1154184/full>.
- Escott, E. (24 de octubre, 2017). What Are the 3 Types of AI? A Guide to Narrow, General and Super Artificial Intelligence. *Codebots*. <https://codebots.com/artificial-intelligence/the-3-types-of-ai-is-the-third-even-possible>.
- Europol's European Cybercrime Centre. (2021). *Malicious Uses and Abuses of Artificial Intelligence*. <https://www.europol.europa.eu/publications-events/publications/malicious-uses-and-abuses-of-artificial-intelligence#downloads>.
- Ezrachi, A., y Stucke, M. E. (2017). Two artificial neural networks meet in an online hub and change the future (of competition, market dynamics and society). *Oxford Legal Studies Research Paper*. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2949434](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2949434).
- Gless, S., Silverman, E. y Weigend, T. (2016). If robots cause harm, who is to blame? Self-driving cars and criminal liability. *New Criminal Law Review*, 19(3), 412–436. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2724592](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2724592).

- Graeff, E. C. (2014). What we should do before the social bots take over: Online privacy protection and the political economy of our near future. *Media in transition 8: Public Media, Private Media*.
- Grupo de Expertos de Alto Nivel en Inteligencia Artificial de la Unión Europea: Comisión Europea. (2019). *A definition of AI: Main capabilities and scientific disciplines*.
- Grupo Europeo sobre Ética de la Ciencia y las Nuevas Tecnologías: Comisión Europea. (2018). *Declaración sobre Inteligencia artificial, robótica y sistemas “autónomos”*.
- Haenlein, M. y Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, 61 (4), 5–14.  
[https://journals.sagepub.com/doi/abs/10.1177/0008125619864925?casa\\_token=WcMMJX4arvMAAAAA:KOQT8hjyzAaMpdXO334X0OaXH\\_reMycaurniAOrUt\\_TtO01JFKMoS6a945sj2yIjmWDQtqur\\_lng8g](https://journals.sagepub.com/doi/abs/10.1177/0008125619864925?casa_token=WcMMJX4arvMAAAAA:KOQT8hjyzAaMpdXO334X0OaXH_reMycaurniAOrUt_TtO01JFKMoS6a945sj2yIjmWDQtqur_lng8g).
- Hallevey, G. (2010). The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Legal Social Control. *Akron Intellectual Property Journal*, 4(2), 171-201.
- King, T., Aggarwal, N., Taddeo, M. y Floridi, L. (2019). Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Science and Engineering Ethics*, 26, 89–120. <https://link.springer.com/article/10.1007/s11948-018-00081-0#additional-information>.
- Lin, T. (2017). The New Market Manipulation. *Emory Law Journal*, 66 (6), 1253-1314.  
<https://scholarlycommons.law.emory.edu/elj/vol66/iss6/1/>.
- Llonín, B. (2022). Acerca de la relación entre inteligencia artificial y responsabilidad penal empresarial. *Revista Sistema Penal Crítico*, 3, 27-48.

- McAllister, A. (2017). Stranger than science fiction: The rise of AI interrogation in the dawn of autonomous robots and the need for an additional protocol to the UN convention against torture. *Minnesota Law Review*, 101, 2527–2573. <https://minnesotalawreview.org/article/stranger-than-science-fiction/>.
- McCarthy, J. (2007). What Is Artificial Intelligence? *Stanford University*. <http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html>.
- Ministerio de Economía, Comercio y Empresa. (2020). *Inteligencia Artificial*. <https://portal.mineco.gob.es/es-es/ministerio/areas-prioritarias/Paginas/inteligencia-artificial.aspx>.
- Miró, F. (2018). Inteligencia Artificial y Justicia Penal: Más allá de los resultados lesivos causados por robots. *Revista de Derecho Penal y Criminología*, 20, 87-130.
- Morales, A. (2021). Inteligencia Artificial y Derecho Penal: Primeras Aproximaciones. *Revista jurídica de Castilla y León*, 53, 177-202. <https://dialnet.unirioja.es/servlet/articulo?codigo=7788274>.
- Morillas, D. (2023). Implicaciones de la inteligencia artificial en el ámbito del Derecho Penal. En J.M. Peris (Ed.), *Derecho Penal, Inteligencia Artificial y Neurociencias* (59-91). Roma Tre-Press. <https://dialnet.unirioja.es/servlet/articulo?codigo=8825016>.
- Muñoz, J.M. (2022). Inteligencia Artificial y responsabilidad penal. *Derecho Digital e Innovación. Digital Law and Innovation Review*, 11.
- Neff, G. y Nagi, P. (2016). Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication*, 10, 4915–4931. <https://ijoc.org/index.php/ijoc/article/view/6277>.
- Nilsson, J. (1998). *Inteligencia Artificial: Una nueva síntesis*. McGraw-Hill.

Obregón, A. y Gómez, J. (2023). *Derecho Penal. Parte General: Elementos Básicos de Teoría del Delito*. Editorial Tecnos.

Parlamento Europeo. (2017). *Resolución con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica*.

Parlamento Europeo. (2024). *La Eurocámara aprueba una ley histórica para regular la inteligencia artificial*. <https://www.europarl.europa.eu/news/es/press-room/20240308IPR19015/la-eurocamara-aprueba-una-ley-historica-para-regular-la-inteligencia-artificial>.

Peñaranda, E. y Basso, G. (2019). La pena: Nociones generales. En J.A., Lascuraín, (Ed.). *Manual de Introducción al Derecho Penal* (161-190). Agencia Estatal Boletín Oficial del Estado.

Quintero, G. (2017). La robótica ante el derecho penal: El vacío de respuesta jurídica a las desviaciones incontroladas. *Revista Electrónica de Estudios Penales y de la Seguridad*, 1.

Rainer Granados, J.J. y Rodríguez Baena, L. (2017). Perspectiva histórica y evolución de la inteligencia artificial. En Ministerio de Defensa (Ed.), *Documentos de Seguridad y Defensa 79: La inteligencia artificial, aplicada a la defensa* (17-37).

Real Academia Española. (s.f.). Artificial. En *Diccionario de la lengua española*. Recuperado el 20 de enero, 2024, de <https://dle.rae.es/artificial?m=form>.

Real Academia Española. (s.f.). Inteligencia. En *Diccionario de la lengua española*. Recuperado el 20 de enero, 2024, de <https://dle.rae.es/inteligencia>.

- Real Academia Española. (s.f.). Manipulación en el mercado de valores. En *Diccionario panhispánico del español jurídico*. Recuperado 2 de febrero, 2024, de <https://dpej.rae.es/lema/manipulación-en-el-mercado-de-valores>.
- Rich, E. (1985). Artificial intelligence and the humanities. *Computers and the Humanities*, 19 (2), 117-122.
- Rodríguez, L. (2010). *Compendio de Derecho Penal. Parte General*. Dykinson S.L.
- Samoili, S., López, M., Gómez, E., De Prato, G., Martínez-Plumed, F., and Delipetrev, B. (2020). *AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence*. Publications Office of the European Union
- Spatt, C. (2014). Security Market Manipulation. *Annual Review of Financial Economics*, 6, 405-418. <https://www.annualreviews.org/doi/10.1146/annurev-financial-110613-034232>.
- Stanford University. (2023). *Stanford AI Index Report 2023*. <https://aiindex.stanford.edu/report/>.
- Universidad de Montreal. (2018). Declaración de Montreal para un desarrollo responsable de la inteligencia artificial.
- Valls, J. (2022). Sobre la responsabilidad penal por la utilización de sistemas inteligentes. *Revista Electrónica de Ciencia Penal y Criminología*, 24. <https://dialnet.unirioja.es/servlet/articulo?codigo=8587635>.