Revista Electrónica de Estudios Penales y de la Seguridad
ISSN: 2531-1565

DE STOCKFISH A SKYNET: CUESTIONES PRINCIPALES DE LA INTELIGENCIA ARTIFICIAL Y EL DERECHO PENAL*

FROM STOCKFISH TO SKYNET: MAIN ISSUES ON ARTIFICIAL INTELLIGENCE AND CRIMINAL LAW

José Mateos Bustamante
Profesor Ayudante Doctor
Universidad de Valladolid (España)

Fecha de recepción: 12 de diciembre de 2022.

Fecha de aceptación: 30 de abril de 2023.

RESUMEN

El desarrollo de las inteligencias artificiales, iniciado a mediados del Siglo XX con sistemas de computación en el ámbito del ajedrez, se ha ido popularizando, extendiendo y complicando. Hoy, las inteligencias artificiales forman parte de nuestros procesos diarios, seamos más o menos conscientes de ellas, y se espera que en un futuro cercano esta tecnología se popularice cada vez más. Se han planteado problemas una vez la inteligencia artificial sobrepasa el nivel digital y pasa a actuar en el entorno físico: drones, robots o vehículos de conducción autónoma actúan en el plano físico, y como consecuencia de esta actuación pueden producirse efectos lesivos. En este trabajo analizamos los principales problemas que surgen a la hora de valorar la posible responsabilidad por estos resultados lesivos de los seres humanos o, incluso, de las propias inteligencias artificiales.

ABSTRACT

The development of artificial intelligence, started in the middle of the 20th century with computer systems in the field of chess, has become popular, extended and complicated. Today, artificial intelligences are part of our daily processes, whether we are more or less aware of them, and this technology is expected to become more and

* Este trabajo desarrolla la comunicación que, con el mismo título, fue seleccionada y expuesta ante el público en el Congreso internacional de Derecho penal y Comportamiento humano: desafíos desde la Neurociencia y la Inteligencia artificial, celebrado en Toledo durante los días 21 a 23 de septiembre de 2022, que se organizó en el marco del proyecto de investigación Derecho Penal y Comportamiento Humano (RTI2018-097838-B-I00).

more popular in the near future. Problems have arisen once artificial intelligence surpasses the digital level and begins to act in the physical environment: drones, robots or autonomous driving vehicles act on the physical plane, and as a consequence of this action, harmful effects can be produced. In this paper we analyze the main issues that arise when assessing the possible responsibility for these harmful results of human beings or, even, of artificial intelligences themselves.

PALABRAS CLAVE

Inteligencia artificial, responsabilidad penal, autoría

KEYWORDS

Artificial intelligence, criminal responsability, authorship

ÍNDICE

SUMMARY

4.5 – Non-autonomous artificial intelligence as a source of knowledge of the standard of care. 4.6 – Closing clause: the fortuitous event and the difficulties of combining Justice and technology. **5 – CAN THE HARMFUL RESULTS PRODUCED BY ARTIFICIAL INTELLIGENCE BE PREVENTED? THE FIGURE OF THE "INTERMEDIATE MAN" AND ITS DISADVANTAGES. 6 – BIBLIOGRAPHY.**

1. INTRODUCCIÓN: BREVE HISTORIA DE LAS INTELIGENCIAS ARTIFICIALES EN EL MUNDO DEL AJEDREZ

El ajedrez es un juego con unas características que han incentivado y facilitado el desarrollo de las inteligencias artificiales. En concreto, es enormemente popular a nivel global, lo cual ayuda a la obtención de los recursos necesarios para la implantación de sistemas cuyo desarrollo puede ser costoso en tiempo y dinero. En segundo lugar, es un juego cuyo aprendizaje, como se dice en el ámbito anglosajón, tiene *low floor* y *high ceiling*. Esto significa que el conocimiento más básico del ajedrez, el "suelo" (sus reglas, condiciones de victoria, colocación y movimiento de las piezas sobre el tablero, etc.), es muy sencillo de aprender, pero al mismo tiempo las estrategias más avanzadas y los conocimientos tácticos profundos que se requieren para ser un jugador experto, esto es, el "techo", son excepcionalmente difíciles de desarrollar.

Esta segunda característica es probablemente la que ha resultado más relevante para el desarrollo de las inteligencias artificiales en el ajedrez, especialmente de las más modernas, que emplean procesos de *machine learning*: resulta muy sencillo diseñar un *software* cuyos parámetros básicos son las reglas del ajedrez, puesto que éstas son pocas y sencillas. Sin embargo, una vez el sistema ha "aprendido" estas reglas, y una vez ha sido dotado de mecanismos automáticos de aprendizaje (o de "práctica", diríamos si fuese un ser humano), los límites que puede alcanzar el programa en cuanto a su nivel como jugador son prácticamente infinitos.

Para contextualizar esta explicación, tengamos en cuenta los siguientes datos: en el año 1950, el matemático estadounidense Claude Shannon calculó, de manera estimada y conservadora, el llamado "número de Shannon", que consiste en cuántas variaciones son posibles en una partida de ajedrez de cuarenta movimientos[1]. Este número es 10^{120}, es decir, un 10 seguido de 120 ceros. Por colocar este dato en perspectiva, el número de átomos observables en todo el universo es de $10^{80.}$ Este inconcebible número de posibilidades presenta un desafío excepcional incluso para la inteligencia artificial: ¿es posible para un programa informático que trabaja al mismo tiempo en una red compuesta por millones de ordenadores y teléfonos móviles, y en la que cada dispositivo calcula millones de partidas diferentes cada segundo, llegar a calcular todas estas posibilidades? La solución a esta pregunta, casi más filosófica que matemática, conocida como el problema de "resolver el ajedrez" (*solving chess* en el original), no parece estar cercana: se estima que serán necesarios cambios cualitativos

[1] SHANNON, C. E., (1950), Programming a computer for playing chess, en Philosophical Magazine, Ser. 7, Vol. 41, Num. 314.

en la capacidad de las máquinas (computación cuántica, por ejemplo) para siquiera acercarse a esa posibilidad[2].

Mucho más interesante que las posibilidades teóricas sobre el final del ajedrez resultan sin duda los logros ya obtenidos en el presente por estas inteligencias artificiales: el primer hito histórico logrado por un programa de ordenador contra un jugador humano sucede en 1996, cuando el entonces campeón del mundo Gary Kasparov cede dos partidas de un total de seis contra Deep Blue, la inteligencia artificial desarrollada por la empresa IBM. Solamente un año después, la misma inteligencia artificial da la vuelta al marcador, derrotando al campeón en un resultado final de 3,5-2,5. A partir de ese momento, y con cada vez más diferencia, cualquier inteligencia artificial ha derrotado sin discusión a los mejores jugadores humanos. En el momento de escribir estas líneas en el actual año 2022, el mejor jugador humano de la historia, Magnus Carlsen, tiene una puntuación ELO de 2861[3], habiendo alcanzado una puntuación máxima de 2882 en mayo de 2014. Stockfish 13, un *software* de desarrollo libre y probablemente el mejor jugador no humano en la actualidad, alcanzó en el año 2021 una puntuación ELO de 3546. La diferencia entre ambos números es mayor de la que parece a primera vista, puesto que el sistema ELO funciona de manera exponencial, no lineal. Es decir que, por ejemplo, un jugador con una puntuación ELO de 2000 tiene mucho más nivel que el doble que un jugador con una puntuación ELO de 1000: así, se estima que en una serie de cien partidas entre Stockfish 13 y Magnus Carlsen, el primero ganaría más de noventa y nueve veces.

2. CAMBIOS EN EL MUNDO DEL AJEDREZ A CONSECUENCIA DE LAS INTELIGENCIAS ARTIFICIALES Y LA POSIBLE EXTRAPOLACIÓN DE ESTOS CAMBIOS A OTROS ÁMBITOS

Como hemos expuesto anteriormente, el papel de las inteligencias artificiales en el ajedrez, especialmente desde comienzo del Siglo XXI, ha sido extremadamente importante. Analizaremos ahora cómo la existencia y difusión de estos programas ha transformado el desarrollo del juego y qué problemas se han presentado, con objeto de intentar extrapolar estos problemas también a otros ámbitos en el que el desarrollo de las inteligencias artificiales no está, de momento, tan extendido.

En primer lugar, ha sido fundamental en el mundo del ajedrez la revolución que las inteligencias artificiales han proporcionado en cuanto a las posibilidades de revisión de partidas y, en ese mismo sentido, la de entrenamiento o discusión de líneas de juego. En tiempos previos a la existencia de estos programas ya se consideraba fundamental para el aprendizaje y la mejora como jugador el repaso de partidas propias o ajenas, es decir, examinar una partida de ajedrez ya terminada, movimiento a movimiento,

[2] SCHAEFFER, J., BURCH, N., BJÖRNSSON, Y., KISHIMOTO, A., MÜLLER, M., LAKE, R., SUTPHEN, S., (2007) Checkers is solved, en Science, Vol. 317, pgs. 1518-1522.

[3] El sistema ELO, desarrollado por Arpad Elo en 1939, es un indicador general del nivel de un jugador de ajedrez, y opera de manera absolutamente objetiva ya que, al contrario que en otros juegos, en el ajedrez no hay ningún factor de azar (las posiciones iniciales y posibilidad de movimientos son siempre los mismos) ni tampoco hay apenas elementos disruptores más allá del propio nivel del jugador (no influyen, por ejemplo, el cansancio físico, las decisiones arbitrales o el desempeño de otros compañeros de equipo)

reflexionando sobre cuáles habrían sido otras posibles alternativas o qué opciones podrían haber sido mejores, valiéndose, cuando era posible, de jugadores de mejor nivel que ilustrasen al alumno sobre líneas o estrategias que pudiesen escapársele. En la actualidad, este papel lo realizan las inteligencias artificiales. Ya no se *discute* sobre ajedrez, de la misma manera que no se *discute* sobre una realidad matemática: la solución idónea a un problema es solamente una y es la que proporcione el programa informático (conocido coloquialmente en el mundillo como "el módulo"). En una determinada posición el mejor movimiento es uno y ningún otro, y es precisamente el que el módulo determine[4]. El aprendizaje y la mejora en el campo del ajedrez es hoy mucho más sencillo, accesible (por poner un ejemplo, Stockfish 13 es un software de licencia libre que puede instalarse en cualquier teléfono móvil) y perfecto que antes, y ello se debe fundamentalmente al uso extendido de la inteligencia artificial, que orienta al jugador sobre qué errores ha cometido en una partida, qué entrenamiento específico necesita para evitar volver a cometerlos y qué alternativas habrían sido las mejores. Esta es una primera extrapolación del uso de la inteligencia artificial a ámbitos externos al ajedrez que nos resulta particularmente interesante: habitualmente, el uso de la inteligencia artificial se ha propuesto como sustitutivo al ser humano, es decir, para tomar decisiones en su lugar, pero tiene también una valiosa función como parámetro de corrección, esto es, como forma de revisar a posteriori si una decisión humana ha sido la más adecuada o no. Se trataría aquí de emplear la información obtenida por la inteligencia artificial para revisar la actuación del ser humano y, en su caso, proporcionarle una forma de aprendizaje o entrenamiento, o incluso la de aportar a un posible órgano sancionador datos sobre la adecuación del comportamiento humano a los estándares requeridos en el ámbito del que se trate.

Igualmente, y de nuevo derivado del nivel de juego muy superior de los programas informáticos, en el mundo del ajedrez se han realizado importantes esfuerzos por apartar a las inteligencias artificiales del juego competitivo. El hecho de que los programas de ajedrez proporcionen siempre al jugador la información de cuál es el mejor movimiento en la situación exacta de su partida invalidaría el propio juego

[4] En realidad, esto no es exactamente así. La inteligencia artificial calcula la mejor jugada en términos generales, pero no en la situación concreta de una determinada partida. Los rivales son jugadores humanos, y como tal tienen tendencias que pueden aprovecharse, lo que puede decantar la balanza en favor de una u otra jugada. Se entenderá mejor esta excepción con un sencillo ejemplo: una inteligencia artificial cuya función sea la de aconsejar la mejor jugada en el juego "piedra, papel o tijera", indicará emplear cada una de las opciones un tercio de las veces, puesto que las tres son equivalentes según las reglas del juego, pero lo hará asumiendo que el rival también empleará cada una de las opciones un tercio de las veces. Sin embargo, si nuestro conocimiento humano nos indica que nuestro rival jamás usa "piedra", por ejemplo, la jugada indicada por la inteligencia artificial es incorrecta, pues jamás deberíamos usar "papel", jugada que nunca gana. Este efecto de "adaptar la mejor jugada dependiendo del rival" ha tenido una importancia moderada en el ajedrez, pero es mucho más relevante en otros ámbitos en los que existen inteligencias artificiales similares, como en el póker. Allí, el componente humano es mucho más relevante, las tendencias de los jugadores humanos son más agudas, y tiene más sentido entender las jugadas sugeridas por la inteligencia artificial (PIO Solver o Deepstack, por ejemplo), como sugerencias. Agradezco enormemente a RAMOS VÁZQUEZ, J. A., sus aportaciones a este respecto en conversaciones mantenidas en el Congreso Internacional de Derecho Penal y Comportamiento Humano: Desafíos Desde la Neurociencia y la Inteligencia Artificial, celebrado en la Universidad de Castilla-la-Mancha los días 21, 22 y 23 de septiembre de 2022.

humano si se les permitiese acceder libremente a esta información: serían, en realidad, las inteligencias artificiales las que jugarían las partidas, siendo los humanos unos meros sustratos físicos que les permitirían mover las piezas, pausar el reloj de juego, y demás movimientos físicos necesarios para el juego presencial. Por lo tanto, de manera análoga a como se persigue, por ejemplo, el dopaje en el deporte, el uso de inteligencias artificiales durante una partida está prohibido en cualquier entorno competitivo de ajedrez.

Sin embargo, este uso de la inteligencia artificial, ilegal en el campo del ajedrez por la abrumadora ventaja que otorga a uno de los competidores, es perfectamente válido en otros ámbitos: pensemos, por ejemplo, en la medicina[5], donde puede emplearse el conocimiento de una inteligencia artificial que, con el *input* apropiado de información sobre el paciente (datos como su edad, estado físico, condiciones preexistentes, etc.), pueda, simulando varios millones de alternativas distintas, personalizar el tratamiento médico que se ajuste exactamente a su necesidad. Aquí está, en nuestra opinión, el uso de las inteligencias artificiales que ha de ser preponderante en la mayor parte de los ámbitos: la función de "asistente de información", un equivalente a las fuentes de conocimiento que ya existen y usan a diario los profesionales de distintos campos (manuales, mementos técnicos, etc.) pero de un nivel y con unas posibilidades cuantitativa y cualitativamente muchos órdenes de magnitud superior, en tanto en cuanto permiten un ajuste exacto de la solución al problema presentado en concreto.

Habitualmente, sin embargo, el uso de la inteligencia artificial se ha propuesto no como complemento de la actuación humana, sino como su sustituto, para tomar decisiones en su lugar y, en ocasiones, actuar incluso físicamente de manera autónoma[6]. Siguiendo el ejemplo médico introducido en el párrafo anterior, el uso de inteligencias artificiales en el ámbito quirúrgico no se limita exclusivamente a una guía para el cirujano, una especie de manual capaz de adaptarse al caso concreto, sino que se han desarrollado también inteligencias artificiales que, mediante el uso de una extensión mecánica, son capaces ellas mismas de intervenir quirúrgicamente.

Este uso de la inteligencia artificial, capaz de tomar decisiones en base a algo muy cercano a un razonamiento que desarrolla en base a su propio aprendizaje, es muy atractivo, y presenta opciones fascinantes de futuro, pero también se han planteado importantes objeciones: se ha argumentado, por ejemplo, que la programación de los parámetros de la inteligencia artificial ha de ser realizada con extremo cuidado, puesto que unas directrices equivocadas sumadas al desarrollo exponencial y autónomo de una inteligencia artificial capaz de aprender autónomamente mediante el proceso de *machine learning* puede conducir a soluciones correctas desde el punto de vista de la inteligencia artificial pero indeseables para el ser humano. Igualmente, se ha planteado también cómo incluso una programación perfecta sólo lo puede ser acorde a las capacidades humanas del programador, mientras que la capacidad de la inteligencia

[5] Explica más detenidamente éste y otros supuestos VALLS PRIETO, J. (2021). Inteligencia artificial, Derechos humanos y bienes jurídicos, Cizur Menor, Thomson Reuters, pgs. 36-37.

[6] Algunos ejemplos de sistemas de inteligencia artificial autónoma ya existentes pueden encontrarse en QUINTERO OLIVARES, G. (2019). La robótica ante el Derecho Penal: el vacío de respuesta jurídica a las desviaciones incontroladas. Revista electrónica de Estudios Penales y de Seguridad, núm. 1., pgs. 17-18.

artificial puede superar con mucho éstas: es decir, incluso una programación de parámetros realizada con la mejor precisión de la que es capaz un ser humano sigue siendo una programación humana, superable rápidamente por una inteligencia artificial capaz de aprender, y que puede desarrollar vías de actuación o tomar soluciones imprevisibles para la programación a la que está sujeta. La inteligencia artificial podría así superar los límites de su propia programación para obedecer a sus directrices más primarias que, de nuevo y trayendo aquí también la primera objeción enumerada anteriormente, debería de realizarse con extremo cuidado. Piénsese, por ejemplo, el caso de ficción de SkyNet, la inteligencia artificial militar de la saga de películas "Terminator" que decide, para proteger a la especie humana, exterminar a la especie humana. La idea de fondo de esta objeción es, en definitiva, que desarrollar inteligencias artificiales que superen las propias capacidades intelectuales humanas las colocaría más allá de los límites de cualquier programación humana, por lo que serían, *de facto*, imposibles de controlar.

3. MÁS ALLÁ DEL AJEDREZ: ¿EXISTEN LAS "INTELIGENCIAS ARTIFICIALES CRIMINALES"?

La diferencia que hemos introducido en el epígrafe anterior entre "inteligencias artificiales de asistencia" e "inteligencias artificiales autónomas" nos sirve como introducción para una ulterior pregunta, que ha sido realmente la cuestión central en el análisis penal de las inteligencias artificiales: si una inteligencia artificial puede actuar de manera independiente, sin una validación o un control humano sobre cada una de sus actuaciones, sino que es capaz de transformar físicamente el mundo de manera autónoma, ¿es posible que pueda cometer delitos[7]? Y si lo hace, ¿puede ser responsabilizada por ello? ¿Y alguna persona?

Es evidente que el hilo conductor de nuestro trabajo hasta este punto, las distintas inteligencias artificiales en el mundo del ajedrez, no cumplen ni siquiera el primero de estos requisitos. Su actividad se limita a un juego muy específico y, aunque puedan jugar a él de manera autónoma (en plataformas digitales) supera completamente su programación la actuación en cualquier otro ámbito, y por lo tanto le es imposible. Sin embargo, como también se ha introducido ya en otros ejemplos de este trabajo, existen numerosas otras inteligencias artificiales en ámbitos de auténtico "tráfico humano" que, si bien no forman parte aún de nuestra rutina diaria, parece que están muy cerca de hacerlo: existen ya o están en fases muy avanzadas de desarrollo las inteligencias artificiales en el ámbito de la medicina, de la conducción autónoma de vehículos o de la seguridad domiciliaria. Estas inteligencias artificiales no solamente van a interactuar con seres humanos de manera frecuente, ya sean sus propios usuarios u otros, sino que además lo van a hacer de manera independiente, sin una actividad humana de control.

Resulta bastante sencillo imaginar varias posibilidades de resultados lesivos producidos por la actividad de estas inteligencias artificiales, y de hecho buena parte de ellos ya se han producido: un dron autopilotado puede, por ejemplo, estrellarse produciendo daños a personas o bienes. De la misma manera, drones cuya función es la

[7] Nos referimos aquí a una concepción amplia del término "delito"

captura de imágenes, utilizados habitualmente para el trazado de mapas topográficos en tres dimensiones o la identificación de personas por parte de las fuerzas de la autoridad, pueden captar imágenes que vulneren la intimidad de las personas. Sistemas de autoconducción de vehículos pueden igualmente, por ejemplo, producir accidentes de tráfico.

En este momento resulta procedente realizar una segunda distinción dentro de las inteligencias artificiales, que se han diferenciado habitualmente entre inteligencias artificiales "débiles" e inteligencias artificiales "fuertes"[8]. Se conoce como "inteligencia artificial débil" a aquella cuya programación se circunscribe a una serie de maniobras concretas y determinadas *a priori*. Todos los ejemplos que hemos examinado hasta ahora son ejemplos paradigmáticos de inteligencia artificial débil: con mayor o menor autonomía, los sistemas tienen una función, definida de manera más o menos amplia, pero el ámbito de actuación, ya sea física o virtual, de la inteligencia artificial se limita exclusivamente a éste. Este es el estado actual de nuestra ciencia y posibilidades técnicas: las inteligencias artificiales tienen un enorme ámbito de autonomía y capacidad, pero solamente en tanto en cuanto su programación está previamente determinada por un ser humano. Como hemos visto con el ejemplo del ajedrez, es posible que de acuerdo a esa programación la inteligencia artificial lleve a cabo comportamientos imprevisibles para el programador, pero no para su programación. Frente a este concepto presente y real de inteligencia artificial débil se ha postulado la posibilidad de una "inteligencia artificial fuerte". Aquí nos encontraríamos igualmente con un sistema artificial, creado y programado por uno o varios seres humanos, pero ya no para realizar labores determinadas, sino para tener un determinado tipo de valores[9]. Esta idea, aún hipotética en el momento en el que se escriben estas líneas, parte de la concepción de que la mente humana funciona en buena medida como un sistema informático: nuestro cerebro computa una enorme cantidad de posibilidades teniendo en cuenta todos los factores que percibimos y, de acuerdo a una personalidad determinada por nuestra experiencia y educación previas, decidimos cómo actuar. Nada de este sistema parece irreplicable, y este es el concepto que habita detrás de la inteligencia artificial fuerte: que una conciencia similar a la humana puede crearse, dotando a una inteligencia artificial de unos determinados parámetros de actuación básicos, una suerte de rasgos de personalidad, en base a los que la inteligencia artificial iría desarrollando una auténtica capacidad de decidir.

Esta diferencia, como veremos a continuación, no es irrelevante en lo que respecta al derecho penal, puesto que la existencia de uno u otro modelo de responsabilidad humana va a depender en buena medida del grado de conexión que le reste al ser humano sobre el comportamiento de la máquina. Anticipamos aquí una conclusión que desarrollaremos: en nuestra opinión, nuestro derecho penal cuenta con herramientas suficientes para resolver la cuestión de la responsabilidad penal en casos de inteligencia artificial débil. No ocurre lo mismo, sin embargo, en casos de inteligencia

[8] LLEDÓ BENITO, I. (2022). El derecho penal, robots, IA y cibercriminalidad: desafíos éticos y jurídicos. ¿Hacia una distopía? Madrid: Dykinson., pgs 100-102.

[9] ALONSO ÁLAMO, M. (2018). La culpabilidad en la encrucijada. En Represión penal y estado de derecho, homenaje al profesor Gonzalo Quintero Olivares, MORALES PRATS, F., TAMARIT SUMALLA, J. M., GARCÍA ALBERO, R. M., Aranzadi, pgs. 311-312.

artificial fuerte, donde los mecanismos e instituciones penales existentes se antojarían insuficientes en el caso de que la posibilidad técnica de desarrollar este tipo de inteligencias artificiales existiese, y todo parece indicar que va a existir en un futuro no tan lejano.

4. MODELOS DE RESPONSABILIDAD PENAL EN LA ACTUACIÓN DE LAS INTELIGENCIAS ARTIFICIALES

Como hemos adelantado, la actuación en los planos digitales y físicos de las inteligencias artificiales puede producir determinados resultados lesivos. De manera indiciaria, un sentido de Justicia intuitivo nos apunta a que, en muchos de esos casos, no puede asimilarse el resultado lesivo producido por una inteligencia artificial al mero caso fortuito, o al menos no de manera general. Las inteligencias artificiales, aún las que funcionan de manera autónoma, mantienen un grado de conexión con los seres humanos que las han creado, programado o de los que de alguna manera dependen, y evidentemente mucho más las que funcionan de manera no autónoma: por poner un ejemplo, si un dron autopilotado responde correctamente a una programación equivocada (se le instruye para obtener información de un investigado en un procedimiento penal y por los parámetros introducidos sobre el sujeto el dron obtiene información privada de una persona que nada tiene que ver con la investigación), una intuición prejurídica nos lleva a buscar una posible responsabilidad penal en quien introdujo de manera errónea los parámetros de búsqueda. No parece, por expresarlo de la manera más gráfica posible, que ese supuesto sea equiparable a quien es golpeado por un rayo en mitad de un descampado y resulta herido de gravedad. Esto no significa, evidentemente, que haya que forzar las instituciones jurídicas para responsabilizar a este sujeto por cualquier vía, pero sí se trata de reflexionar desprendiéndose de ideas preconcebidas sobre si estas conductas pueden tener cabida en nuestros modelos de responsabilidad vigentes.

Una primera dificultad que nos vamos a encontrar, y que tenemos que puntualizar de manera introductoria, es que resulta muy complicado agrupar todas las posibilidades en un mismo supuesto. Hemos examinado ya cómo existen inteligencias artificiales con funcionamiento, capacidades y alcance muy diferente. También es muy diferente la relación que tienen las personas con estas inteligencias artificiales, incluso dentro del mismo tipo. Rechazamos ya como hipótesis de trabajo que todos los grupos de casos deban afrontarse desde una misma institución penal, sino que diferentes comportamientos humanos sobre distintas inteligencias artificiales deberán de resolverse de maneras distintas.

Por trazar un panorama rápido de la cuestión y a modo de esquema introductorio, en los siguientes subepígrafes analizaremos las siguientes posibilidades: la consideración del ser humano usuario de la inteligencia artificial como autor directo del delito, la posibilidad de considerar a la propia inteligencia artificial autor del delito, la figura de la autoría mediata, la posibilidad de la comisión por omisión colocando al ser humano en posición de garante de los posibles resultados lesivos producidos por

una inteligencia artificial, y, por último, la posibilidad de que ninguna de estas estructuras sea de aplicación[10].

Reiteramos, a riesgo de ser repetitivos, que no se trata aquí de exponer las distintas posibilidades para terminar eligiendo una, sino que las siguientes instituciones han de combinarse para crear un sistema de responsabilidad sobre la inteligencia artificial que permita cubrir todos los casos de resultados lesivos que no deban considerarse casos fortuitos. De la misma manera que el derecho penal "humano" combina distintos tipos de autoría según se den unas circunstancias u otras, ha de procederse igual en el caso de las inteligencias artificiales.

4.1. La autoría directa de la propia inteligencia artificial

Colocamos en primer término este grupo de casos no por su mayor incidencia ni por su mayor relevancia, sino por todo lo contrario: se ha discutido en la doctrina si es posible responsabilizar a la propia inteligencia artificial por los resultados lesivos producidos. En palabras de MIRÓ LLINARES, "*el que algunas de las tecnologías de IA estén comenzando a basarse en modelos de aprendizaje en los que no es sencillo definir el curso causal de la decisión tomada por la máquina, han llevado a la doctrina a revisar si el modelo de responsabilidad de la teoría del delito es adecuado para tal reto y a proponer soluciones interpretativas respecto de las diferentes situaciones de accidentes causados por máquinas*"[11]. Plantea aquí el autor las posibilidades dogmáticas que arroja el debate anteriormente esquematizado entre la "inteligencia artificial débil" y la "inteligencia artificial fuerte". Las primeras no dejan de ser una complicación de medios tecnológicos que ya existen: el derecho penal ha de comportarse respecto de una inteligencia artificial de la misma manera que lo hace respecto de una pistola o de una bomba de relojería: son mecanismos complejos, a veces incognoscibles para el autor del delito, pero alrededor del mecanismo hay en muchas ocasiones (no siempre) uno o varios humanos de quienes depende, o que lo instrumentalizan para sus fines. No parece factible que una inteligencia artificial encargada, por ejemplo, de la vigilancia domiciliaria (vinculada a la gestión de cámaras de seguridad, armas automáticas, verjas electrificadas, etc.) decida saltarse su programación y comenzar a producir resultados lesivos más allá de sus parámetros. Todo lo más, sucederá que sus parámetros están incorrectamente definidos, que el usuario emplea de manera equivocada las posibilidades del sistema (dolosa o imprudentemente), o que todos los seres humanos, programador y usuario, han tenido un comportamiento irreprochable y el sistema define una variación imprevisible para todos con un resultado lesivo, similar a cuando la inteligencia artificial de ajedrez emplea, de acuerdo a sus parámetros, una jugada que ningún jugador había pensado que podría producirse. La responsabilidad penal, por lo tanto, será de un humano o no será de nadie, pero no es razonable que en nuestro

[10] Omitiremos en este trabajo, por cuestiones de extensión, las interesantes reflexiones sobre si *de lege ferenda* debería crearse un nuevo "derecho de la robótica" que creara categorías jurídicas nuevas, sino que nos limitaremos a las posibilidades jurídicas con el sistema vigente. Se ha tratado esta cuestión en QUINTERO OLIVARES, G. (2019). La robótica ante el Derecho Penal: el vacío de respuesta jurídica a las desviaciones incontroladas. Revista electrónica de Estudios Penales y de Seguridad, núm. 1., pgs. 13-14.

[11] MIRÓ LLINARES, F. (2018). Inteligencia artificial y justicia penal: más allá de los resultados lesivos causados por robots. Revista de derecho penal y criminología, núm. 20., pg. 95.

estado tecnológico actual las inteligencias artificiales tengan la capacidad de *delinquir autónomamente*.

Distinta sería la situación, sin duda, si llegase a desarrollarse de manera plena un sistema de inteligencia artificial fuerte. En este, como ya hemos adelantado, la inteligencia artificial no estaría determinada por unos parámetros *de actuación*, sino por unos parámetros *de conducta*. Así, tendríamos inteligencias artificiales programadas en *valores* más que en *acciones*, y en base a esos valores podrían determinar (y en ese caso quizás ya sí sería procedente emplear el término *razonar*) que una determinada conducta lesiva es adecuada. Este tipo de inteligencia artificial, inexistente en la actualidad, pero la más frecuente en las obras de ficción (HAL-9000, Skynet, las máquinas de la saga Matrix, etc.).

Interesa aquí traer a colación determinadas consideraciones que se han adoptado en el seno de la Unión Europea, y muy específicamente y por su reciente aprobación la Resolución del Parlamento Europeo, de 3 de mayo de 2022, sobre la inteligencia artificial en la era digital (2020/2266(INI))[12]. Entiende aquí el Parlamento Europeo en su decimoquinta observación que "*aunque la IA actual se ha vuelto mucho más eficaz y potente que la IA simbólica, gracias a los grandes incrementos de las capacidades informáticas, por el momento solo puede resolver tareas claramente definidas en nichos específicos del dominio, como el ajedrez o el reconocimiento de imágenes, y su programación no está concebida para reconocer completamente las acciones que realiza el sistema de IA; destaca que, al contrario de lo que sugiere su nombre, los sistemas de IA carecen de «inteligencia» en el sentido humano del término; señala que por ello se habla de IA «débil» y todavía no es más que una herramienta que facilita recomendaciones y predicciones; observa, por ejemplo, que los vehículos autónomos funcionan mediante una combinación de diversos sistemas de IA monotarea que, en conjunto, son capaces de elaborar un mapa tridimensional del entorno del vehículo para que su sistema operativo pueda tomar decisiones*". En su siguiente observación decimosexta aleja la posibilidad de una inteligencia artificial fuerte cuando considera que "*muchos temores relacionados con la IA se basan en conceptos hipotéticos como la IA general, la superinteligencia artificial y la singularidad, que, en teoría, podrían hacer que la inteligencia de las máquinas superara a la inteligencia humana en numerosos ámbitos; destaca que existen dudas sobre la posibilidad de llegar a esa IA especulativa con nuestras tecnologías y leyes científicas; estima, no obstante, que los legisladores deben abordar los riesgos que plantea actualmente la toma de decisiones basada en la IA, ya que ha quedado claramente patente que efectos nocivos como la discriminación racial y sexual ya son atribuibles a casos concretos en los que la IA se ha implantado sin salvaguardias*".

Esta reciente posición modera las conclusiones de la anterior Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2103(INL))[13], donde se solicitaba a la Comisión Europea, si bien exclusivamente en el ámbito civil, "*crear a largo*

[12] https://www.europarl.europa.eu/doceo/document/TA-9-2022-0140_ES.html. Última consulta el 30/01/2023.

[13] https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52017IP0051&from=EN. Última consulta el 30/01/2023.

plazo 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".

Sobrepasa los límites de este trabajo intentar plantear una solución a qué sucederá si las inteligencias artificiales llegan al punto de ser funcional y hasta quizá orgánicamente indistinguibles de los seres humanos. LLEDÓ BENITO ha planteado que "*aunque se las configurase [a las máquinas autónomas] con la pretendida personalidad electrónica carecerían de los elementos intelectivos y volitivos (inteligencia y voluntad humana: conocer y querer la acción) que posee cualquier persona humana y por lo tanto sería imposible aplicar lo que decía WELZEL "configuración del hecho por medio de la voluntad de realización que conscientemente lo dirige""*[14]. De nuevo sin poder entrar al detalle a esta cuestión, discrepamos aquí del autor: si se habla de inteligencia artificial fuerte es precisamente un sistema capaz de realizar hechos de manera voluntaria, conociendo y queriendo la acción, puesto que la única diferencia que existiría en ese momento con un ser humano sería el origen de sus capacidades intelectivas: nosotros las habríamos obtenido del proceso natural de formación embrionaria, gestación y educación, mientras que la inteligencia artificial lo habría obtenido de manera diseñada. A partir de esa diferencia de origen, creo que el sistema jurídico, social y ético en su conjunto debería prácticamente ser reconstruido desde cero ampliando el concepto de "ser humano" en el que cabrían estas inteligencias artificiales fuertes, a las que, siendo capaces de motivarse por las normas, no tendría sentido excluir del sistema penal.

Por fortuna, ese escenario parece lejano y por lo tanto toda reflexión sobre él son conjeturas sin demasiado sentido, pues se mueven exclusivamente en el ámbito de la ciencia ficción. Volvamos por lo tanto a consideraciones sobre inteligencias artificiales débiles, ya existentes en nuestra realidad y que plantean problemas muy presentes.

4.2. La autoría directa de un ser humano

Analizaremos a continuación aquellos supuestos en los que de un resultado lesivo producido materialmente por una inteligencia artificial puede hacerse responsable como autor directo a un ser humano. Los ejemplos más habituales de estos casos se producirán en el ámbito físico (es decir, mediante la intervención no sólo de una inteligencia artificial, sino también de una extensión mecánica de la misma: un dron, un robot o cualquier otra forma de actuar sobre el plano material), pero no exclusivamente: resulta relativamente sencillo imaginar "delitos digitales" cometidos por inteligencias artificiales sin soporte físico, como procesadores automáticos de textos o programas informáticos que pudiesen, por ejemplo, producir un daño a otros archivos informáticos.

[14] LLEDÓ BENITO, I. (2022). El derecho penal, robots, IA y cibercriminalidad: desafíos éticos y jurídicos. ¿Hacia una distopía? Madrid: Dykinson., pg. 83.

En cualquiera de los dos casos, se trata aquí de supuestos en los que una persona emplea el comportamiento previsto de una inteligencia artificial produciéndose un resultado lesivo. En la categorización que realiza QUINTERO OLIVARES[15], bajo un epígrafe denominado "*la utilización de ingenios robóticos para delinquir*", el autor emplea entiende que "*esa posibilidad, tan imaginable como la de utilización de la robótica con fines militares, en sí misma no encierra ninguna dificultad*". Desarrolla el autor a continuación su posición en el sentido de entender que la utilización consciente y voluntaria por parte del autor para provocar resultados lesivos no plantea ninguna dificultad a nivel teórico para ser atribuible a la autoría directa con dolo de primer grado, e igualmente sucede en casos de dolo de segundo grado o de consecuencias necesarias.

A la tesis principal del autor[16] poco podemos añadir: quien instrumentaliza un sistema de inteligencia artificial para producir un resultado típico nos parece evidentemente un autor directo de un delito doloso: imaginemos, por ejemplo, al autor que introduce en un dron de combate militar la información sobre una persona a quien quiere matar, dejando que sea el propio dron quien, siguiendo su programación de manera correcta, identifique y acabe con la vida del objetivo. En estos casos nos parece indistinguible la figura del creador de la inteligencia artificial del usuario de la inteligencia artificial, pues ambos dominan el hecho absolutamente en el momento de esa introducción de parámetros. Evidentemente, más habitual será el caso del usuario de inteligencia artificial, ya sea empleando el sistema tal y como fue concebido o alterándolo de alguna manera, pues por una pura cuestión de coste y dificultad tecnológica se plantea como más probable el emplear una inteligencia artificial para delinquir que el crear una inteligencia artificial para delinquir.

Sí consideramos necesario añadir a la tesis de QUINTERO OLIVARES, que presenta como vinculadas a la autoría directa el dolo de primer y segundo grado la posibilidad también de que se considere autor directo al creador, programador o usuario de la inteligencia artificial a título de dolo eventual o de imprudencia. Nada obsta a estas consideraciones, pues es perfectamente imaginable que, de la misma manera que alguien consciente y voluntariamente emplea una inteligencia artificial para cometer un delito (que se correspondería con el dolo directo), se emplee también ésta con el conocimiento de la posibilidad de que el delito se cometa (dolo eventual, sin que corresponda aquí entrar a las distintas posiciones que existen sobre el elemento volitivo en esta figura), o incluso imprudencia, con o sin representación, cuando el autor emplee la inteligencia artificial sin la voluntad de cometer un delito pero de manera contraria a cómo debe usarse esa inteligencia artificial de manera prudente. Imaginemos, por ejemplo, a quien programa un dron para que vuele a toda velocidad cerca de la vivienda de un tercero, para impresionarle o demostrarle la capacidad técnica del aparato y que, incluso quizá ante el aviso del sistema de que ese vuelo puede ser peligroso (por las condiciones del viento, por ejemplo), lo programa igualmente, produciéndose finalmente los daños. No nos parece que tenga mayor complicación, en estos casos, el

[15] QUINTERO OLIVARES, G. (2019). La robótica ante el Derecho Penal: el vacío de respuesta jurídica a las desviaciones incontroladas. Revista electrónica de Estudios Penales y de Seguridad, núm. 1., pgs. 14-16.

[16] Ha desarrollado también esta posición DOMÍNGUEZ PECO, E. M. (2019). Los robots en el derecho penal. En BARRIO ANDRÉS, M. Derecho de los robots (pp. 169-188)., pgs. 176-178.

imputar este resultado típico a una autoría directa con dolo (en cualquiera de sus clases) o imprudencia (en cualquiera de sus modalidades).

4.3. La posición de garante del usuario de la inteligencia artificial por control de fuentes de peligro y la comisión por omisión

Es cierto que determinados casos de autoría directa, tanto dolosa o imprudente, como el ejemplo relatado en el párrafo anterior, recuerdan más bien a situaciones de comisión por omisión, de nuevo tanto dolosas como imprudentes. Éste sería el caso, por ejemplo, de quien programa un sistema de defensa domiciliaria, imaginemos, conectado a un arma que dispara automáticamente a las personas que entran en el jardín del autor que cumplan unos determinados parámetros que el usuario decide: así, es posible introducir en la inteligencia artificial determinadas excepciones o requisitos previos, como que en ningún caso se dispare a miembros de la familia a los que la cámara puede reconocer facialmente. Pues bien, si el sujeto, conociendo la posibilidad técnica del sistema de inteligencia artificial de discriminar potenciales objetivos, no incluye estas excepciones, y la inteligencia artificial dispara lesionando de gravedad a uno de sus hijos que vuelve a casa del colegio, nos parece que estas lesiones deben imputársele al usuario a título de autor directo imprudente sin mayor inconveniente. ¿Pero sería un auténtico delito activo o este grupo de casos es más bien reconducible a la comisión por omisión por vía de la posición de garante derivada del control de fuentes de peligro por parte del autor?

Se plantea de nuevo esta situación QUINTERO OLIVARES[17] cuando reflexiona sobre "*la permisión de "actuación" de máquinas que pueden desviarse de su función*", circunscribiendo estos casos a inteligencias artificiales que actúan de manera autónoma, pero en la que es técnicamente posible el control humano. Interesa traer aquí a colación el modelo de HARBERS, PETERS y NEERINCX[18], que diferencian tres posibilidades de control: en primer lugar, el modelo de "man in the loop", en el que el ser humano se integra como parte fundamental del funcionamiento de la inteligencia artificial, requiriéndose sus aportes a intervalos de tiempo regulares para su correcto funcionamiento. En segundo lugar, cabe el modelo "man on the loop", en el que la inteligencia artificial funciona de manera autónoma, pero con la posibilidad de intervención por parte de un ser humano que actúa, en cierta medida, como mecanismo de control, pero no como parte del propio sistema. Por último, en el modelo de "man out of the loop" la máquina actúa de manera completamente independiente, sin que el ser humano tenga posibilidad técnica de intervenir en su comportamiento.

Nos podemos plantear qué sucede si en los primeros dos modelos el humano encargado de, o bien mantener el sistema en funcionamiento ("man in the loop"), o bien de controlar que el proceso se lleva a cabo de manera correcta ("man on the loop"), omite ese comportamiento debido de intervención o control, produciéndose el

[17] QUINTERO OLIVARES, G. (2019). La robótica ante el Derecho Penal: el vacío de respuesta jurídica a las desviaciones incontroladas. Revista electrónica de Estudios Penales y de Seguridad, núm. 1., pgs. 17-21.

[18] A través de MIRÓ LLINARES, F. (2018). Inteligencia artificial y justicia penal: más allá de los resultados lesivos causados por robots. Revista de derecho penal y criminología, núm. 20., pg. 93.

resultado lesivo por el comportamiento de la inteligencia artificial, pero siendo en última instancia responsabilidad del ser humano.

De nuevo, no nos parece que esta situación sea distinta a los modelos de comisión por omisión con posición de garante del autor por control de fuentes de peligro ya existentes: ¿es realmente distinta, tanto desde un punto de vista técnico, como desde un punto de vista de valoración ético-social, la actuación de quien omite, por ejemplo, la actualización de protocolos de un dron militar, permitiendo que ese dron mantenga como objetivo a una persona que se ha sabido que finalmente no es a quien realmente se quiere eliminar, de quien omite detener un mecanismo de maquinaria pesada en una fábrica teniendo la responsabilidad de hacerlo? Ambos casos nos parecen perfectamente asimilables.

4.4. La autoría mediata

Se ha planteado en la doctrina la posibilidad de considerar los resultados lesivos producidos por la actuación material de una inteligencia artificial como casos de autoría mediata. Según esta posibilidad, estudiada, por ejemplo, por LLEDÓ BENITO[19] o DOMÍNGUEZ PECO[20], el autor estaría instrumentalizando a la inteligencia artificial, aprovechándose de un proceso de toma de decisión que sabe defectuoso o, cuando menos, utilizable, para conseguir un resultado lesivo deseado, pero sin llevarlo ella misma a cabo.

Los casos en los que se está pensando cuando se plantea esta posibilidad no son, en puridad, distintos a los que se producen en los casos que hemos examinado como de autoría directa. Se trata, por lo tanto, del uso instrumentalizado de manera consciente de un procedimiento pseudo-intelectivo de la inteligencia artificial para obtener un resultado lesivo.

Esta estructura puede parecer, efectivamente, similar a la de la autoría mediata, pero nos parece un espejismo. En primer lugar, la autoría mediata requiere la presencia, en términos literales del Código Penal de 1995, de "*otro*"[21]. A nuestro entender[22], este "otro" ha de ser necesariamente otro ser humano, pues de otra manera se estaría ampliando el mecanismo de la autoría mediata hasta límites que exceden con mucho lo razonable: ¿es autor mediato, por ejemplo, quien programa una bomba para que detone cuando la víctima pise un botón oculto, produciéndose su muerte? No podemos decir con rigor jurídico que en esos casos el autor "instrumentalice" a la bomba, y de la misma manera no consideramos adecuado el considerar que se pueda realmente

[19] LLEDÓ BENITO, I. (2022). El derecho penal, robots, IA y cibercriminalidad: desafíos éticos y jurídicos. ¿Hacia una distopía? Madrid: Dykinson., pg. 91.

[20] DOMÍNGUEZ PECO, E. M. (2019). Los robots en el derecho penal. En BARRIO ANDRÉS, M. Derecho de los robots (pp. 169-188)., pgs. 173-176.

[21] Artículo 28 del Código Penal de 1995: "*Son autores quienes realizan el hecho por sí solos, conjuntamente o por medio de otro del que se sirven como instrumento*".

[22] Véase también QUINTERO OLIVARES, G. (2019). La robótica ante el Derecho Penal: el vacío de respuesta jurídica a las desviaciones incontroladas. Revista electrónica de Estudios penales y de seguridad, núm. 1., pg. 14.

"instrumentalizar" a una inteligencia artificial, o al menos no en un sentido técnico penal.

Distinta sería la situación, es cierto, en un sistema de inteligencias artificiales fuertes. Allí la inteligencia artificial sería mucho más equiparable a un ser humano que a un mecanismo (como puede ser la bomba de nuestro ejemplo) y la concepción de "otro" podría posiblemente ampliarse hasta cubrir a los "medio humanos" que, como ya hemos adelantado, serían las inteligencias artificiales en ese contexto, cuando no directamente humanos. En ese contexto (reiteramos, en estos momentos únicamente hipotético) entendemos que no habría inconveniente en considerar que quien instrumentalizase a una inteligencia artificial, que no comprendiese el significado antijurídico de su acción, (o que de cualquier otra manera no cumpliese todos los requisitos del delito concebido como acción típica, antijurídica y culpable), estaría en un caso de autoría mediata[23], pero, de nuevo, ello sucedería en un contexto en el que la diferencia entre ser humano e inteligencia artificial habría quedado ya completamente desdibujada.

4.5. La inteligencia artificial no autónoma como fuente de conocimiento de la norma de cuidado

Trataremos aquí, por último, un grupo de casos que nos resulta de particular interés porque intenta resolver supuestos que en las demás instituciones jurídico-penales tienen un encaje complicado: los comportamientos de las inteligencias artificiales no autónomas, es decir, de aquellas inteligencias artificiales que no son capaces técnicamente de desplegar un comportamiento material sobre el mundo exterior, y que por lo tanto cumplen únicamente la función de asistentes, guías de un procedimiento llevado a cabo en todo momento por un ser humano.

En estos casos lo que sucede es que un ser humano lleva a cabo un comportamiento de manera completamente autónomo (al menos desde un punto de vista material), pero para la toma de decisiones sobre qué curso de acción tomar acude a una inteligencia artificial, que le orienta o aconseja sobre la mejor opción. Aquí interesa volver a traer a colación el comienzo de nuestro trabajo, en el que reflexionábamos sobre el valor de la inteligencia artificial en el mundo del ajedrez: allí ha tenido una incidencia absolutamente decisiva, pero principalmente su valor ha sido el de parámetro de corrección para analizar una partida o una jugada, o incluso el de profesor de ajedrez. De la misma manera, muchas de las inteligencias artificiales actuales cumplen una función similar: el navegador de un coche con sistema GPS, por ejemplo, es una inteligencia artificial muy básica que calcula muy rápidamente la ruta óptima entre dos puntos, reaccionando de manera inmediata a cambios en el tráfico, cortes de carreteras, y demás factores, pero que en ningún momento toma el control del vehículo. Un último ejemplo lo encontramos en el ámbito médico, donde, además de sistemas de intervención quirúrgica autónomos, existen también numerosos dispositivos de orientación sobre un tratamiento médico, sistemas de realidad

[23] Opina en sentido contrario LLEDÓ BENITO, I. (2022). El derecho penal, robots, IA y cibercriminalidad: desafíos éticos y jurídicos. ¿Hacia una distopía? Madrid: Dykinson., pg. 96.

aumentada que orientan en directo al cirujano sobre cómo proceder sobre el cuerpo del paciente, etc.

En estos casos, cabe preguntarse qué sucede si la inteligencia artificial sugiere un comportamiento típico, que puede producirse debido a una programación inadecuada (o, incluso, manipulada por un tercero), pero también en casos de funcionamiento normal de la inteligencia artificial, que debido a su capacidad de computación prevé soluciones inesperadas para quien la programó.

Recordemos en este punto que el sistema de *deep learning* de las inteligencias artificiales no tiene ningún componente ético: se ajusta a unos parámetros iniciales de los que se puede excluir determinados comportamientos, pero, dentro de ellos, la inteligencia artificial se mueve con extraordinaria fluidez. En este momento resulta interesante apuntar el ejemplo de una inteligencia artificial de reciente creación, ChatGPT, un procesador de textos automático, y que por su aún escaso desarrollo presenta soluciones originales y no del todo deseables. El funcionamiento para el usuario es muy sencillo: con una interfaz similar a la de un chat, se introduce una pregunta o una solicitud y, acudiendo a un barrido instantáneo de los resultados más populares en buscadores de internet, ChatGPT presenta la solución que considera óptima[24].

Pues bien, se han reportado[25] numerosos casos en los que, cumpliendo fielmente su programación, ChatGPT arroja resultados falsos (pero populares), manipulados o directamente plagiados de trabajos relevantes en el ámbito por el que se le pregunta. ¿Qué sucedería si se emplease la información suministrada por esta inteligencia artificial y ello produjera algún daño?

Evidentemente, el ejemplo de ChatGPT es extremo: es una inteligencia artificial extremadamente joven, y cualquier experto que la emplee en su campo durante apenas unos minutos comprueba casi instantáneamente que es nula como fuente de autoridad. ¿Pero podemos decir lo mismo de inteligencias artificiales con años de experiencia fiable probada? ¿Deben los usuarios seguir ciegamente las instrucciones proporcionadas por una inteligencia artificial? Y, sobre todo, ¿si como consecuencia del seguimiento de estas instrucciones se produce un resultado típico, puede hacérseles responsables de alguna manera?

---

[24] Como demostración sencilla de su funcionamiento, ChatGPT arroja el siguiente resultado cuando se le pregunta "*¿cómo se ha de resolver el problema de la responsabilidad penal de las inteligencias artificiales?*": "*La responsabilidad penal de las inteligencias artificiales es un tema controversial y no hay un consenso claro sobre cómo abordarlo. Algunas opciones incluyen:*

Atribuir responsabilidad a los desarrolladores o propietarios de las IA.

Crear leyes específicas para regular la conducta de las IA.

Desarrollar sistemas de seguimiento de la IA para determinar su responsabilidad.

Es importante considerar tanto la ética como la seguridad legal al abordar este problema y se espera que continúe siendo un área de investigación y discusión en el futuro."

[25] DAVENPORT, C. (2022), ChatGPT Is an Impressive AI Chatbot That Can't Stop Lying. https://www.howtogeek.com/852769/chatgpt-is-an-impressive-ai-chatbot-that-cant-stop-lying/. Última consulta el 30/01/2023.

En nuestra opinión, estos supuestos han de colocarse en el ámbito del conocimiento de la norma de cuidado[26], esto es, el comportamiento que el sujeto debería haber observado en lugar de un comportamiento imprudente, de la misma manera que fuera del ámbito de las inteligencias artificiales se realiza este mismo análisis: en el caso de un comportamiento no doloso de un médico que provoca la muerte del paciente, ¿actuó el médico de manera adecuada según la *lex artis* de su oficio, teniendo en cuenta las circunstancias concretas del caso? ¿conocía cómo se debía aplicar esta *lex artis* en este caso concreto? Y, si no lo conocía, ¿debería de haberlo conocido? Son, en definitiva, las preguntas habituales en el campo de la imprudencia.

Las respuestas a estas preguntas, sin embargo, se complican cuando se introduce a la inteligencia artificial de por medio, especialmente cuando se le presume una capacidad de análisis muy superior a la de un humano, porque justamente esa es su función. Como hemos adelantado en el marco del ajedrez, cuando Stockfish considera que una jugada es la mejor, su autoridad es absoluta: esa jugada es la mejor. Esa autoridad deriva no solamente de su nivel como jugador adquirido tras décadas de repetición, sino también de su capacidad computacional de muchos órdenes de magnitud superiores a los de un ser humano. En el campo del ajedrez, es un delirio creer que nuestro criterio como jugadores humanos es superior al de la inteligencia artificial[27]. ¿Podemos exigir al médico que desoiga la sugerencia de la inteligencia artificial y la sustituya por su propio criterio, mientras al mismo tiempo se le supone a la inteligencia artificial una capacidad muy superior a la suya propia? En nuestra opinión, en algunos casos sí.

Concretamente, se mantendrá la responsabilidad individual del profesional asistido por la inteligencia artificial en el caso en el que la sugerencia que ésta arroja se aparte manifiesta y groseramente de la *lex artis*. Evidentemente, esta es una regla general que deberá aplicarse no solamente sector por sector, sino también caso por caso, pero intentaremos ilustrarla con un ejemplo sencillo: supongamos que un médico considera, de acuerdo a sus propios conocimientos médicos, estudios, indicaciones del prospecto, etc., que la dosis adecuada de un determinado medicamento oscila entre los 100 miligramos y los 200 miligramos, según edad del paciente, peso, gravedad de los síntomas, etc. Si introduce los parámetros del paciente en un programa de computación artificial destinado a optimizar el tratamiento de los pacientes, y éste devuelve un resultado de 220 miligramos, no se podrá hacer responsable al médico que suministre esa dosis al paciente, produciéndose, por ejemplo, lesiones en éste. En este caso el apartamiento del propio criterio del médico es menor: es cierto que, de haber actuado de manera independiente, nuestro médico nunca habría recetado 220 miligramos, pero la sugerencia de esta dosis se le presenta como razonable y no se puede esperar otro comportamiento de él que el seguir el dictado de la inteligencia artificial. Si, en su lugar, el programa arrojase un resultado de 20.000 miligramos, (esto es, cien veces superior al máximo que el médico considera según su propia valoración), el apartamiento es de una entidad tal que sí puede exigirse al médico que sustituya el criterio de la inteligencia

[26] En este sentido también QUINTERO OLIVARES, G. (2019). La robótica ante el Derecho Penal: el vacío de respuesta jurídica a las desviaciones incontroladas. Revista electrónica de Estudios Penales y de Seguridad, núm. 1., pgs. 21-22.

[27] Véase, sin embargo, Nota 4.

artificial por el suyo propio. De la misma manera, y por ilustrar la propuesta con otro ejemplo, un conductor no debe obedecer ciegamente las instrucciones del navegador de su vehículo, sino que debe usar ese conocimiento como complemento a su propio criterio sobre las normas de tráfico: si el navegador GPS del vehículo no está debidamente actualizado, por ejemplo, y aconseja conducir por una calle en sentido contrario, el comportamiento del conductor que decide inobservar la señal de tráfico que indica el sentido contrario de la circulación porque es lo que le indica el navegador será imprudente.

De nuevo, esta estructura no es esencialmente distinta a las reglas que ya existen sobre la determinación de la norma de cuidado en el ámbito de la imprudencia: sustituyamos en nuestro ejemplo anterior a la inteligencia artificial por otro médico humano de mayor experiencia, y la solución sería idéntica: cuando un profesional medio valora el criterio de otro profesional al que él considera de mayor nivel, es lógico que desatienda su propia opinión en virtud de la más experta, siempre que ésta se mantenga en el ámbito de lo que le resulta razonable. No ocurre lo mismo si esa opinión experta le aconseja un comportamiento que se aparta de manera amplia de su propio criterio.

La solución a los casos de comportamientos posiblemente imprudentes inducidos por una inteligencia artificial puede mantenerse, por lo tanto, en el marco de la imprudencia tradicional.

4.6. Cláusula de cierre: el caso fortuito y las dificultades de conjugar Justicia y tecnología

Como hemos adelantado al comenzar el estudio de los distintos grupos de casos, entendemos que no debe de abordarse la incidencia de la inteligencia artificial en la responsabilidad penal desde un prisma único, sino que cada una de las situaciones habrá de resolverse según la institución jurídico-penal que sea aplicable. La cuestión que queda abierta es qué sucede cuando ninguna de ellas resulta de aplicación, es decir, cuando el resultado lesivo en el que se ven involucrados tanto seres humanos como inteligencias artificiales no puede ser cubierto de manera satisfactoria ni por la autoría directa, ni por la comisión por omisión, ni por el delito imprudente en el caso de inteligencias artificiales no autónomas.

A nuestro entender, y por estricta aplicación del principio de legalidad, los resultados lesivos producidos por inteligencias artificiales programadas correctamente, controladas de manera adecuada por un ser humano e imprevisibles para el usuario, no pueden implicar responsabilidad penal alguna. De nuevo, esto no es una novedad penal: entendemos lo mismo en los casos en los que se usa de manera adecuada cualquier tecnología que falla de manera imprevisible produciéndose el resultado lesivo. La doctrina ha enumerado gran cantidad de ejemplos de este grupo de casos: el conductor de un vehículo que, manejándolo de manera diligente, pierde el control por enfrentarse a un factor externo imprevisible (imaginemos, por ejemplo, una fuerte lluvia inesperada e inanunciada), produciéndose un resultado lesivo de un tercero, no está sometido a responsabilidad penal ninguna, al menos en nuestro derecho vigente.

Como hemos adelantado al comienzo de este epígrafe, esto quizá puede arrojar resultados contrarios a nuestra intuición de Justicia, y así se escuchan en numerosas ocasiones desde medios de comunicación o profanos en Derecho cantos de sirena que invitan a adoptar un modelo similar al de la responsabilidad penal objetiva. A pesar de que nuestro conductor del ejemplo anterior no pudo hacer nada por evitar el accidente, y de que conducía de manera correcta, se nos invita a veces a responsabilizarlo igualmente con argumentos inadmisibles en un derecho penal basado en la culpabilidad como "no haber conducido" o "en el momento en el que te pones al volante te haces responsable de todo lo que pase". Sin embargo, una reflexión razonada y fundada en Derecho no puede sino hacernos rechazar estos argumentos.

En los ámbitos tecnológicos tradicionales se ha admitido sin demasiada resistencia que las novedades tecnológicas implican mayores riesgos, que a su vez implican mayores resultados lesivos, pero que esto no nos hace rechazar la tecnología debido a que el valor para la sociedad sigue siendo positivo. Siguiendo nuestro ejemplo anterior, la existencia y el manejo habitual por parte de la mayoría de la sociedad de los vehículos a motor implica un mayor riesgo para la salud y la vida. Y sin embargo, no nos planteamos su prohibición porque, en una balanza hipotética, los beneficios que aportan estos vehículos a la sociedad son claramente superiores. Y ello no solamente en ámbitos ajenos a la salud y la vida (transporte, economía, ocio, etc.), sino también en la propia protección personal: la misma tecnología que produce una muerte en un accidente de tráfico salva una vida transportando a una persona en riesgo de muerte por un infarto a un hospital en una ambulancia.

A nuestro entender, debe hacerse la misma ponderación en el ámbito de las inteligencias artificiales, y la conclusión debe ser la misma[28]. El uso habitual y frecuente de esta nueva tecnología, que además va a extenderse cada vez más, debe encontrar su encaje en nuestro sistema penal, responsabilizando a los seres humanos responsables del correcto funcionamiento de los sistemas cuando se dan los elementos necesarios para ello. Debemos perder el miedo, sin embargo, a asumir que en determinados casos esa responsabilidad humana no existe, y que el funcionamiento correcto de una inteligencia artificial también correctamente diseñada, programada y controlada puede producir resultados lesivos que tenemos que asumir como contrapartida necesaria a los beneficios que la nueva tecnología nos aporta.

5. ¿PUEDEN PREVENIRSE LOS RESULTADOS LESIVOS PRODUCIDOS POR INTELIGENCIAS ARTIFICIALES? LA FIGURA DEL "HOMBRE INTERMEDIO" Y SUS INCONVENIENTES

La reflexión con la que terminamos el epígrafe anterior no debe hacernos perder de vista que la asunción de riesgos necesarios por el desarrollo y extensión de la inteligencia artificial es compatible con buscar mecanismos de reducción y, si es posible, eliminación de esos riesgos. Es justamente la solución intermedia que se ha buscado en otros ámbitos tecnológicos de manera pacífica. Por seguir con el ejemplo de la conducción mecánica, los riesgos derivados de la popularización de vehículos a motor

[28] Más sobre este asunto en MOZO SEOANE, A. (2018). Robots e inteligencia artificial. Control de sus riesgos. Revista general de legislación y jurisprudencia, núm. 2., pgs. 238-241.

se han asumido como necesarios, pero existen fuertes medidas dirigidas a reducirlos en la medida de lo posible: las normas de tráfico tienen fundamentalmente esa función.

En este sentido, en el ámbito de las inteligencias artificiales se ha propuesto lo que se conoce como la solución del "hombre intermedio". Así, de nuevo en el ámbito de la legislación europea, la Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2103(INL))[29] establece en su Considerando Q que "*es necesario integrar salvaguardias y la posibilidad de control y verificación por parte de las personas en los procesos de toma de decisiones automatizados y basados en algoritmos*", e igualmente en el Principio tercero que "*pone de relieve que el desarrollo de la tecnología robótica debe orientarse a complementar las capacidades humanas y no a sustituirlas; considera fundamental garantizar que, en el desarrollo de la robótica y los sistemas de inteligencia artificial, los seres humanos tengan en todo momento el control sobre las máquinas inteligentes*". También la Resolución del Parlamento Europeo, de 6 de octubre de 2021, sobre la inteligencia artificial en el Derecho penal y su utilización por las autoridades policiales y judiciales en asuntos penales (2020/2016(INI))[30] establece en su conclusión decimosexta que "*en el contexto de las actividades judiciales y policiales, todas las decisiones con efectos legales deben ser tomadas siempre por un ser humano al que puedan pedirse cuentas de las decisiones adoptadas", y más adelante en el mismo epígrafe considera también "que las autoridades que recurren a los sistemas de IA deben respetar unas normas jurídicas extremadamente estrictas y garantizar la intervención humana*".

Esta idea del "hombre intermedio" intentaría evitar el modelo de "man out of the loop" que hemos introducido en epígrafes anteriores: se trataría de mantener siempre conectado al uso de la inteligencia artificial a un ser humano del que, o bien dependiese el propio funcionamiento de la inteligencia artificial ("man in the loop"), o que, en todo caso, tuviese facultades de control inmediatas e instantáneas ("man on the loop"). Se trataría, por lo tanto, de evitar la automatización absoluta de procesos, de independizar el comportamiento de la inteligencia artificial para no solamente reducir los riesgos que puede producir una inteligencia artificial autónoma, sino también para mantener un vínculo de responsabilidad claro con una o varias personas que se han situado voluntariamente en una situación de control y por lo tanto también de responsabilidad.

Esta solución nos parece acertada, pero no lo es tanto su idea de aplicación con carácter general. Como hemos explorado a lo largo de todo este trabajo, las aplicaciones de las inteligencias artificiales en labores humanas son ya muchas y muy variadas, y van a serlo mucho más en los próximos años. En muchos de esos casos será no solamente posible, sino también recomendable, mantener a un ser humano, a un "piloto", en sentido amplio, conectado al proceso de toma de decisión y actuación de la inteligencia

[29] https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52017IP0051&from=EN#:~:text=C%20252%2F241-,Jueves%2C%2016%20de%20febrero%20de%202017,conflicto%20con%20la%20primera%20ley. Última consulta el 30/01/2023.

[30] https://www.europarl.europa.eu/doceo/document/TA-9-2021-0405_ES.html. Última consulta el 30/01/2023.

artificial. Concretamente, esto será posible en casos en los que esta toma de decisión no ha de ser inmediata (en el sentido temporal de la palabra), sino que cabe un periodo de decisión, valoración y reflexión. Retomando algunos ejemplos que ya hemos manejado anteriormente, una inteligencia artificial que valore el tratamiento médico óptimo para un paciente, evaluando los síntomas y otras circunstancias personales, y diseñe a la medida exacta de la situación qué dosis de medicamentos han de tomarse, puede y debe estar supervisada por un médico que interponga su criterio humano, sea o no éste penalmente responsable de las posibles desviaciones de la *lex artis*, como ya hemos examinado anteriormente. De la misma manera, un dron de identificación y asesinato de objetivos militares puede ser técnicamente controlado por un ser humano, interrumpiéndose el protocolo en cualquier momento, y por lo tanto cualquier maniobra militar en ese sentido debe contar con personal especializado que, o bien pilote directamente el comportamiento de la inteligencia artificial, o que como mínimo pueda controlarlo, corregirlo o interrumpirlo.

Pero no sucede lo mismo en todos los casos: una de las ventajas de la prodigiosa capacidad computacional de las inteligencias artificiales es no solamente la *mejor* toma de decisiones, sino también la *más rápida* toma de decisiones. En el ámbito de la conducción autónoma de vehículos, por ejemplo, el uso de la inteligencia artificial perdería buena parte de su sentido si cada maniobra del vehículo tuviera que ser autorizada por su conductor: en estos casos debe asumirse un descontrol del ser humano sobre la inteligencia artificial, puesto que este descontrol es necesario para que se cumpla la propia función de conducción autónoma.

Por lo tanto, y en conclusión, aceptamos como orientación general la figura del hombre intermedio en aquellos casos en los que sea posible desde un punto de vista técnico y se mantenga la ventaja del uso de la inteligencia artificial. Esto será así, de hecho, en la mayoría de casos. Sin embargo, en otros casos la intervención del ser humano durante el protocolo de actuación es técnicamente posible, pero desnaturaliza la propia función de la inteligencia artificial. No deben descartarse tampoco la obligatoriedad de un control externo del ser humano sobre estos procesos (por ejemplo, mantener actualizado el *software* que controla autónomamente nuestro vehículo), pero no cabe aquí hablar de un auténtico “hombre intermedio”.

6 - BIBLIOGRAFÍA

- ALONSO ÁLAMO, M. (2018). La culpabilidad en la encrucijada. En Represión penal y estado de derecho, homenaje al profesor Gonzalo Quintero Olivares, MORALES PRATS, F., TAMARIT SUMALLA, J. M., GARCÍA ALBERO, R. M., Aranzadi, pgs. 299-312.

- DOMÍNGUEZ PECO, E. M. (2019). Los robots en el derecho penal. En BARRIO ANDRÉS, M. Derecho de los robots (pp. 169-188).

- QUINTERO OLIVARES, G. (2019). La robótica ante el Derecho Penal: el vacío de respuesta jurídica a las desviaciones incontroladas. Revista electrónica de Estudios penales y de seguridad, núm. 1.

- LLEDÓ BENITO, I. (2022). El derecho penal, robots, IA y cibercriminalidad: desafíos éticos y jurídicos. ¿Hacia una distopía? Madrid: Dykinson.

- MIRÓ LLINARES, F. (2018). Inteligencia artificial y justicia penal: más allá de los resultados lesivos causados por robots. Revista de derecho penal y criminología, núm. 20., pgs. 87-130.

- MOZO SEOANE, A. (2018). Robots e inteligencia artificial. Control de sus riesgos. Revista general de legislación y jurisprudencia, núm. 2., pgs. 237-252.

- SCHAEFFER, J., BURCH, N., BJÖRNSSON, Y., KISHIMOTO, A., MÜLLER, M., LAKE, R., SUTPHEN, S., (2007) Checkers is solved, en Science, Vol. 317.

- SHANNON, C. E., (1950), Programming a computer for playing chess, en Philosophical Magazine, Ser. 7, Vol. 41, Num. 314.

- VALLS PRIETO, J. (2021). Inteligencia artificial, Derechos humanos y bienes jurídicos, Cizur Menor, Thomson Reuters.