ULPLR I ULP LAW REVIEW I REVISTA DE DIREITO DA ULP
VOL. 16 N. 1 E N. 2 (2022)
DOI: 10.46294/ULPLR-RDULP.V16

LA ATRIBUCIÓN DE RESPONSABILIDAD PENAL POR LOS HECHOS COMETIDOS POR SISTEMAS AUTÓNOMOS INTELIGENTES, ROBÓTICA Y TECNOLOGÍAS CONEXAS*

CARLOS MARÍA ROMEO CASABONA**

RESUMEN: El planteamiento que se propone en este trabajo es que, al menos a corto o medio plazo, el Derecho Penal dispone de instrumentos adecuados para imputar la responsabilidad penal a los seres humanos que han intervenido en el diseño, fabricación, distribución o utilización de estos sistemas inteligentes, incluso aunque se trata de sistemas autónomos. Ello sin perjuicio de que se puedan imponer también medidas directas contra el sistema inteligente con el fin de bloquear un riesgo objetivo de reiteración en el delito. Para fundamentar tal construcción se parte de varias categorías, como son el control humano significativo (CHS) y el criterio del *compliance*, utilizado como sustento de la responsabilidad penal de las personas jurídicas. En este estudio se presenta un bosquejo.

ABSTRACT: *The approach proposed in this paper is that, at least in the short to medium term, criminal law has adequate instruments to hold criminally liable those human beings who have been involved in the design, manufacture, distribution or use of these intelligent systems, even if they are autonomous. This is without prejudice to the fact that measures may also be imposed directly against the intelligent system in order to block an objective risk of recidivism. The grounds for such a construction is based on several categories, such as meaningful human control (MHC) and the criterion of compliance, used for the criminal liability of legal persons. This study is a preliminary outline.*

1. INTRODUCCIÓN: Ya en la actualidad, y más todavía a medida de que avance el futuro, es presumible que se están produciendo comportamientos delictivos en los entornos más desarrollados y amplios de la inteligencia artificial, por no repetir lo que es una referencia común en los ciberdelitos. En el conjunto de los delitos financieros y las organizaciones criminales transnacionales los datos disponibles parecen apuntar en esta dirección (ej., blanqueo o lavado de dinero). Por consiguiente, podemos asumir que se ha abierto un nuevo escenario criminal y criminológico a los penalistas (y no solo a nosotros, también a otros profesionales, no solo juristas).

El Parlamento Europeo ha declarado que la determinación de la responsabilidad como consecuencia de las decisiones tomadas por sistemas de IA y robots es una cuestión de gran interés y de extrema importancia a medida de que el perfeccionamiento de estas tecnologías va creciendo en la industria y empieza a impactar más directamente en múltiples aspectos de nuestra vida cotidiana[1]. Asimismo, ha anunciado que van a ser objeto de regulación legal-mediante un Reglamento de la Comisión Europea- diversos aspectos relativos a la creación, desarrollo y puesta en funcionamiento de robots y sistemas de IA en el marco ético y jurídico adecuados[2].

Inmediatamente surgen numerosas preguntas: Debemos reconocer que los robots y sistemas de IA pueden lesionar bienes jurídicos penalmente protegidos? No cabe duda de que desde un punto de vista externo la respuesta ha de ser afirmativa,

* El presente trabajo se realiza en el marco del Proyecto de investigación financiado por el Ministerio español de Ciencia e Innovación sobre "Ciberseguridad y Ciberdelitos" (RTI2018-099306-B-100). Y del apoyo a Grupos de Investigación del Sistema Universitario Vasco del Gobierno Vasco (IT 1541-22). Un trabajo preliminar publiqué bajo el título *Criminal responsibility of robots and autonomous artificial intelligent systems?*, en Comunicaciones en propiedad industrial y derecho de la competencia, nº 91, 2020, pp. 167-187.

** Dr. iur. Dr. med. Dr h.c. mult. Catedrático de Derecho Penal Universidad del País Vasco/EHU, Bilbao, España. carlosmaria.romeo@ehu.eus

1 Parlamento Europeo, *Informe con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica*, 27.01.2017, A8-0005/2017.

2 V. *la Resolución del Parlamento Europeo,* de 20 de octubre de 2020, *con recomendaciones destinadas a la Comisión sobre un marco de los aspectos éticos de la inteligencia artificial, la robótica y las tecnologías conexas* (2020/2012(INL)).

VOL. 16 N. 1 E N. 2 (2022)
DOI: 10.46294/ULPLR-RDULP.V16

tanto si los calificamos como meros instrumentos en manos de los humanos, lo que no supondría una gran novedad para el sistema penal actual, como si les reconocemos cierta autonomía en sus decisiones, no dependiente ya de los humanos, o no significativamente dependiente de ellos. Por consiguiente, las preguntas que podemos plantearnos han de ser otras e ir más lejos.

Deberán responder penalmente los seres humanos que mantienen alguna forma de supervisión y control sobre los mismos? Cumplirán los sistemas inteligentes, especialmente los llamados autónomos, siempre los requisitos asignados por la doctrina penal y la jurisprudencia a la teoría del delito y será por ello posible su imputación penal? O bastará con satisfacer civilmente los daños y perjuicios causados por un sistema de IA? Es decir, si deberemos conformarnos con la reparación del daño y con la indemnización de los perjuicios materiales y morales sufridos.

2. NECESITAMOS UN NUEVO SISTEMA PENAL CONSTRUIDO PARA LA IA Y LOS ROBOTS?

La pregunta más atrevida es si también la IA, los robots, los sistemas inteligentes autónomos y las tecnologías conexas deberán responder directamente de los delitos cometidos por sus hechos, con independencia ahora de cómo se puedan calificar penalmente estos hechos; es decir, si podremos imputarles penalmente por considerar que existe el fundamento suficiente para considerarles sujetos penalmente responsables o, por el contrario, deberán quedar al margen del Derecho Penal, por no poder satisfacer los requisitos dogmáticos que pudieran fundamentar su propia y directa responsabilidad penal.

La cuestión es averiguar si estas categorías fundamentales, concebidas para el ser humano y aplicadas exclusivamente al mismo durante más de dos siglos, son plenamente trasladables a los robots y en especial a los sistemas autónomos inteligentes y a otras tecnologías conexas; o si debemos renunciar a ellas y, en consecuencia, al sistema penal como respuesta a las conductas típicas "cometidas" por los sistemas inteligentes; o si, finalmente, debemos adaptar la estructura conceptual –dogmática- del delito a las características de estos sistemas o, incluso, aprovechar esta oportunidad para renovar y/o modernizar el Derecho Penal[3]. Pero, en qué dirección?

Una de las mayores y permanentes preocupaciones de los penalistas ha sido asegurar que el Derecho Penal se mantenga lo más cerca posible de los aspectos materiales del delito, huyendo de ficciones y formalismos (a salvo de los derivados principalmente del principio de legalidad) propios de otros sectores del ordenamiento jurídico. No obstante, no podemos olvidar que el Derecho, y también el Derecho Penal, es una creación humana, un instrumento normativo del que se vale, entre otros, para asegurar la convivencia social. Es decir, que esta poderosa y sofisticada herramienta jurídica puesta al servicio de la sociedad puede ser supeditada a las necesidades humanas de cada momento y por ello adaptarla y modificarla según demanden aquéllas[4].

No habría problema en revisar las características actuales de la teoría del delito y adaptarlas para dar cabida a la responsabilidad penal de los sistemas y productos de inteligencia artificial. Sin embargo, no es menos cierto que esta modificación, orientada a satisfacer unos objetivos específicos, podría "contaminar" la concepción y los requisitos de la teoría del delito aplicable a seres humanos, con el probable efecto de relajar algunos de ellos –de los requisitos-, o incluso prescindir de los mismos, en detrimento de las garantías que comportan y que en suma descansan en la concepción subjetiva de la responsabilidad penal a partir del elemento objetivo irrenunciable del hecho –(necesidad de una acción u omisión típicas), o principio del hecho.

En efecto, el dolo implica una voluntad de realización de una acción tipificada (descrita en la figura delictiva), además de la conciencia de la misma[5]. Y la culpabilidad presupone una capacidad de obrar de otro modo (o de que el autor sea motivado por la norma, defenderían algunos autores) y ser susceptible de recibir el reproche jurídico por haber obrado antijurídicamente, cuando pudo haberse actuado conforme a derecho; en suma, de otro modo a como se actuó en la situación concreta[6].

Una cuestión que debe ser discutida es, en consecuencia, si estas características

[3] Apuntan también a la conveniencia de revisar a fondo la categoría de la culpabilidad en particular Monika Simmler / Nora Markwalder, *Roboter in der Verantwortung? – Zur Neuauflage der Debatte um den funktionalen Schuldbegriff*, en ZSTW 2017, 129 (1), pp. 20–47 (p. 47).

[4] Como muy bien apunta Javier Valls Prieto, *Sobre la responsabilidad penal por la utilización de sistemas inteligentes*, Revista Electrónica de Ciencia Penal y Criminología, ISSN 1695-0194 RECPC 24-27 (2022), se trata de un asunto metodológico con diversas fases (ej., sí o no a la responsabilidad penal de los sistemas de IA; quién o qué será responsable; bajo qué presupuestos y condiciones, etc. Y, en consecuencia, cómo podría afectar cada una de estas respuestas al sistema penal. Este estudio apunta a la primera fase y a sus posibles incidencias para el sistema penal.

[5] José Cerezo Mir, *Derecho Penal Español, Parte General,* BdF, Buenos Aires, 2008, pp. 430 y s.

[6] Asier Urruela Mora, *La culpabilidad,* en CM Romeo Casabona / E Sola Reche / MA Boldova Pasamar (Coords.), "Derecho Penal, Parte General. Introducción y Teoría Jurídica del Delito", 2ª ed., Comares, Granada, 2016, pp. 257 y ss.

VOL. 16 N. 1 E N. 2 (2022)
DOI: 10.46294/ULPLR-RDULP.V16

subjetivas fundamentales del delito –o elementos del delito- concurren o pueden concurrir en los sistemas inteligentes actuales o que pueden estar disponibles a medio plazo; o si, al no ser posible encontrarlas en la actualidad, debemos renunciar a ellas cuando se trate de imputar penalmente a un sistema inteligente, asumiendo así la responsabilidad penal de estos sistemas tecnológicos. Por ello habría que valorar también el coste que podría comportar para la estabilidad del sistema penal introducir modificaciones profundas en la estructura del delito con el fin de dar cabida a la responsabilidad penal de los sistemas inteligentes.

Y al final, deberíamos preguntarnos también si no será mejor buscar otras vías de imputación penal que se dirijan contra seres humanos, si es que reúnen todos los presupuestos de imputación penal –objetiva y subjetiva-, sin perjuicio de que los sistemas de IA relacionados con el delito también puedan verse afectados por las consecuencias jurídicas del mismo.

3. VALIDEZ DE LOS ESTÁNDARES LEGALES PARA RESOLVER PROBLEMAS DE RESPONSABILIDAD DE SISTEMAS DE IA POR LESIONES A DETERMINADOS BIENES JURÍDICOS ? RESPONSABILIDAD PENAL POR IMPRUDENCIA

Si el resultado producido forma parte del tipo de un delito determinado y se halla como tal tipificado como un delito imprudente, habrá que comprobar entonces si concurren los elementos del tipo de lo injusto de este delito, en particular si se produjo o no la infracción del cuidado debido –objetivo- en la actividad realizada[7].

Un sistema inteligente podría cometer un daño ilícito a un bien jurídico penalmente protegido debido a un error o fallo en el momento de valorar la forma de llevar a cabo esa acción. Por ejemplo, debido a la falta de información suficiente en el sistema para tomar la decisión más correcta, no pudiendo por ello prever las consecuencias lesivas o los riesgos implícitos en su actuación, carecería de capacidad para preverlas en la situación concreta. Esta falta de previsión podría deberse a un diseño defectuoso de éste, que no le permitiría derivar una decisión o una propuesta de actuación sin disponer previamente de todos o los suficientes elementos de valoración de la situación.

Lo realmente preocupante en relación con el sistema automatizado inteligente es que aunque podríamos asumir que aquél ha cometido un error que puede producir una lesión a un bien jurídico (p. ej., de una persona), pero no parece ser capaz de conocer no sólo el contenido del cuidado debido, es decir, de un elemento normativo, sino también de valorarlo como tal, para lo que sólo el ser humano es apto; o, incluso, no parece que fuera capaz de actuar de forma contraria a la norma, si su programación funciona de acuerdo a como fue diseñada. Se ha recordado que para que se pueda imputar un hecho imprudente a un delincuente, ya sea un ser humano o un sistema de IA, es necesario que ambos sean conscientes de los aspectos fácticos de su conducta[8]. Si un daño producido figura como tal en el tipo de un delito imprudente, entonces será necesario comprobar si se cumplen los elementos del tipo de ese delito, en particular si se produjo o no la infracción del cuidado debido en la actividad realizada. En este punto también es necesario recordar la importancia del llamado "riesgo permitido" en la modulación del cuidado debido exigible en la situación concreta.

Los criterios dogmáticos desarrollados sobre este concepto serían plenamente aplicables a estas situaciones, siendo especialmente relevante para establecer el límite máximo del "riesgo permitido"[9]. El Derecho Penal debe cumplir su función preventiva, lo cual es especialmente aconsejable y factible en relación con las tecnologías avanzadas. Es probable que en el futuro se implemente una fuerte intervención de las autoridades administrativas, reforzada por normas que coordinen y/o regulen el funcionamiento permitido de los sistemas de IA según sus características específicas respectivas, exigiendo la reducción del riesgo del sistema de IA a límites tolerables.

Esto significa que el marco en el que se permite el margen de riesgo en relación con las decisiones e intervenciones basadas en la IA debería construirse con una estructura cercana a la idea de "cumplimiento" (*compliance*), que constituye uno de los fundamentos de la responsabilidad penal de las personas jurídicas para la mayoría de los estudiosos del Derecho Penal[10]. directrices externas e internas, que son objetivos importantes de la empresa. directrices internas, que son objetivos corporativos importantes.

La puesta en el mercado del producto tecnológico estaría precedida por un examen general y minucioso del sistema por una agencia externa acreditada[11], formada por técnicos expertos autorizados e independientes. Ellos valorarían la adecuación técnica del robot o del sistema de IA para el servicio o actividades que deberían prestar, el grado de error predecible, su básica inocuidad para los

[7] Carlos M. Romeo Casabona, *El tipo del delito de acción imprudente,* en CM Romeo Casabona / E Sola Reche / MA Boldova Pasamar (Coords.), "Derecho Penal, Parte General. Introducción y Teoría Jurídica del Delito", 2ª ed., cit., pp. 133 y ss

[8] G. Hallevy, *Liability for Crimes Involving Artificial Intelligence Systems,* Springer, 2015, p. 125..

[9] V. sobre criterios para determinar el límite máximo de riesgo permitido Carlos M. Romeo Casabona, *El tipo del delito de acción imprudente,* cit., p. 138.

[10] V. con más detalle, Thomas Rotsch, *Criminal Compliance,* InDret, 1/2012. p. 8.

[11] Perin, *Prudenza, dovere di conoscenza e colpa penale. Proposta per un metodo di giudizio,* Ed Scientifica, Napoli, 2020, p. 407.

bienes jurídicos, etc. Habría que determinar previamente qué sistemas debería pasar por estos controles o revisiones, dado el nivel de riesgo que pueden presentar. La valoración positiva o acreditación oficial daría lugar a la autorización de la utilización pública o privada del sistema de IA, indicando las revisiones y validaciones que deberían realizar los usuarios, o que deberían establecer sus empresas u organizaciones, sobre las conclusiones y propuestas de aquél; y señalando también las funciones específicas para las que hubiera sido acreditado[12]. Estos dos elementos ayudarían a delimitar el terreno del riesgo permitido. La dificultad de ofrecer criterios practicables empieza precisamente aquí, al presentarse en la realidad variables de difícil subsunción en algún delito imprudente. Ello nos devuelve a los criterios normativos propios de la imprudencia, como son el criterio del hombre sensato y responsable y el principio de confianza, principalmente[13]. Sin poder entrar ahora mucho más allá en el contenido de estos criterios[14], podemos apuntar algunas observaciones que relativizan o al menos cuestionan su traslado al escenario de la IA.

De este modo se reincorpora a este escenario otro elemento normativo más. Sin embargo, lo que en realidad preocupa en relación con el sistema automatizado inteligente es que éste cometa un error que pueda producir la lesión de un bien jurídico, pero no parece que aquél sea capaz de conocer no sólo el contenido del cuidado debido, es decir, de un elemento normativo, sino, además, de valorarlo como tal, para lo que solo son aptos los seres humanos; o, incluso, tampoco parece que sea capaz de actuar de forma contraria a la norma, si su programación funciona conforme a cómo fue diseñada.

El principio de confianza indica que una persona que participa en una actividad de relación, en la vida social, puede actuar en la confianza de que los demás participantes en la misma actuarán a su vez cumpliendo sus propios deberes de cuidado, salvo que se tengan indicios de que algún participante está incumpliendo el cuidado debido que a él le incumbe[15].

La pregunta que surge inmediatamente es si con este planteamiento las grandes expectativas depositadas en la IA no se verán disminuidas, al sugerir una constante actividad de revisión y validación. Al enfrentarnos a tecnologías en constante desarrollo y perfeccionamiento, las máximas de experiencia desempeñarán una relevante función a la hora de señalar el cuidado debido, que apelará fundamentalmente a la figura ideal del hombre sensato y responsable y ayudará a determinar el margen de automatización asumible en cada situación y la relajación progresiva de esta prevención a medida de que se refuerce el perfeccionamiento del sistema de IA en cuestión y sea más *confiable*.

En resumen, no descartamos por completo que el principio de confianza pueda desempeñar alguna función para determinar si el ser humano ha infringido -o no- el cuidado objetivamente debido como elemento del tipo del delito imprudente en relación con el funcionamiento de un sistema de IA causante de un resultado de lesión de un bien jurídico penalmente protegido. Sin embargo, este principio no aporta ninguna respuesta en relación con la perspectiva que estamos analizando en este estudio: considerar la IA como posible actor de una conducta imprudente punible, pues no parece que a corto o medio plazo un sistema pueda valorar con la precisión necesaria los comportamientos ajenos en el mismo entorno. En conclusión, el principio de confianza no aporta, al menos por el momento, argumentos decisivos que ayuden a delimitar a quién o a qué se atribuye la imputación por imprudencia y con qué fundamento dogmático. No cabe señalar en el estado actual de la tecnología pautas más precisas que den mayor seguridad al profesional actuante ni al juzgador sentenciador en un proceso judicial.

4. LESIONES DE BIENES JURÍDICOS PRODUCIDAS DOLOSAMENTE

4.1. EL ROBOT O EL SISTEMA DE IA, COMO MEROS INSTRUMENTOS DE LA ACCIÓN HUMANA

Si el sistema inteligente ha sido utilizado por una persona humana ésta podría ser responsable penalmente como *autora directa* o inmediata (p. ej., delitos dolosos de homicidio, de lesiones corporales o de daños, según el caso), pues él realiza materialmente la acción típica en la medida en que el sistema hubiera sido utilizado como un instrumento o herramienta para cometer el delito. En la actualidad puede servir de ejemplo de esta hipótesis el recurso a las ciberarmas; el sujeto al que imputar el resultado dañoso será el ser humano (p. ej., militar) que activó el arma o el sistema que determina su activación, incluso aunque, siendo "inteligentes", modifiquen su trayectoria u objetivo, el total de la descarga destructiva o el momento de activación. Después vendrá la comprobación de la concurrencia de los demás elementos de imputación penal.

No procede calificar estas hipótesis entonces como de autoría mediata. Como es

[12] C.M. Romeo Casabona / G. Lazcoz Moratinos, *Inteligencia artificial aplicada a la salud: ¿Qué marco jurídico?*, Rev Der Gen H, 52, 2020, 161 y ss.

[13] V. A. Perin, Estandarización y automatización en medicina: El deber de cuidado del profesional entre la legítima confianza y la debida prudencia, *Revista Chilena de Derecho y Tecnología*, vol. 8, nº 1, 2019, pp. 3-28 (14 y s.).

[14] Sobre ellos v. Romeo Casabona, *El tipo del delito de acción imprudente*, cit., pp. 141 y s.

[15] V. Romeo Casabona, *El tipo del delito de acción imprudente*, cit., p. 141.

VOL. 16 N. 1 E N. 2 (2022)
DOI: 10.46294/ULPLR-RDULP.V16

sabido, ésta consiste en que el autor no ejecuta directamente el hecho delictivo, sino que se vale de otro (un ser humano), que es quien ejecuta el hecho típico, sobre el que tiene el dominio del hecho, valiéndose para ello de conseguir el dominio de la voluntad del sujeto instrumento. Varias son las reflexiones que se pueden apuntar sobre esta hipótesis: en realidad no hay un ser humano que ejecute la acción –el sujeto instrumento- interpuesto entre el supuesto autor mediato y el sujeto pasivo del delito.

Yendo más allá, evaluando la posibilidad de que el sistema inteligente actuase como instrumento tampoco parece factible que aquél pudiera ser algo más que un instrumento material, de modo que pudiera actuar, bien bajo error o coacción, o valiéndose de un menor o de un inimputable, o actuando en el contexto de una organización social jerarquizada (aparatos organizados de poder)[16], pues probablemente no nos encontraríamos entonces ante un sistema realmente autónomo en el que pudieran darse las circunstancias mencionadas. Esta propuesta tiene como punto de partida la tesis que defendemos aquí y que se va extendiendo en el plano normativo supranacional o internacional, del Control Humano Significativo, al que deberían estar sometidos y revisados determinados sistemas inteligentes y en particular los sistemas inteligentes autónomos.

4.2. SISTEMAS AUTÓNOMOS INTELIGENTES: AGENTES CRIMINALES?

Partimos aquí de que ni el propietario ni el creador o diseñador del sistema inteligente tienen la capacidad de un control previo sobre las decisiones que pueda adoptar el sistema inteligente; éste las toma autónomamente.

Por consiguiente, el tratamiento jurídico de las decisiones de esta naturaleza parece, a primera vista, que debería dirigirse al sistema de IA en cuanto tal y analizar la posibilidad de atribución de un delito doloso a dicho sistema autónomo. La adopción de medidas preventivas de aseguramiento del sistema parece necesaria previamente a la introducción en el mercado o a la incorporación funcional efectiva de estos sistemas.

Nos encontramos, pues, ante una situación similar a la que genera la responsabilidad penal de las personas jurídicas, pues aquí también parece necesario atribuir un contenido específico y diferente a los elementos subjetivos del tipo y a la culpabilidad.

El fundamento se basa en la existencia o no de estructuras organizativas que impidan conductas ilícitas en su seno (*compliance*). Habría que trasladar al régimen de los sistemas inteligentes procedimientos de validación de su inocuidad antes de introducirlos en el mercado o de ponerlos a disposición de la persona que los encargó (probable usuario final). Una cuestión crucial en este contexto de validación debería ser asimismo mantener el sistema o robot bajo un control ''significativo" (o relevante) por parte de los humanos (CHS, el usuario del sistema o robot)[17] .

La duda es si el incumplimiento o cumplimiento defectuoso de estos procedimientos enfrentaría la responsabilidad del diseñador, el fabricante, el homologador o el usuario del sistema, o si habría que hacer responsable directamente al ente inteligente.

Siendo de entrada poco favorable a que se haya introducido en el Derecho Penal de los Estados de la Unión Europea el sistema de responsabilidad penal de las personas jurídicas, por innecesario político-criminalmente (existen otras vías jurídicas no estrictamente penales), por comportar una quiebra de la unidad de perspectiva de la teoría del delito, y por exigir un abandono de dos elementos fundamentales de la teoría del delito, considero también, aclarado lo anterior, que hay algunas claves de esta responsabilidad que pueden ser útiles para abrir el sistema a otros posibles sujetos/entes no humanos a los que podría atribuirse responsabilidad penal sin pasar por la exigencia de requisitos subjetivos emocionales y morales que los penalistas hemos venido considerando esenciales.

Por el momento, parece más razonable y eficaz, al tiempo que menos perturbador, optar por la primera vía.

En el caso de que un sistema artificial inteligente pudiera causar una lesión tipificada tal vez la respuesta podría ser bloquearlo o destruirlo, sin calificar por el momento la naturaleza jurídica (penal) de estas respuestas. La cuestión es si sólo habría que actuar de este modo sobre ese producto o sistema en concreto o habría que hacerlo también sobre todos los modelos de la misma serie de un sistema inteligente que presentaran un potencial similar riesgo objetivo de lesión; éstos deberían ser, bien bloqueados y sometidos a revisión, o bien destruidos como último recurso, en todo caso irrenunciable en hipótesis extremas.

En efecto, si el modelo inteligente en cuanto tal ha producido la lesión de un bien jurídico penalmente protegido sin haber sido sometido a ninguna variación o modificación de su programación, todos los modelos producidos y puestos en funcionamiento en el mercado adolecen del mismo *riesgo objetivo* de producir la lesión de bienes jurídicos, por lo que deben recibir la misma respuesta jurídica. Es en ese riesgo objetivo donde radica el fundamento jurídico de controlar, bloquear o eliminar el sistema inteligente (o el robot). Podríamos decir que el riesgo de "reincidencia" a partir del hecho típico cometido por un modelo se traslada a toda la producción de ese modelo.

[16] V. Fernando G Sánchez Lázaro, *Autoría y participación*, en CM Romeo Casabona / E Sola Reche / MA Boldova Pasamar (Coords.), "Derecho Penal, Parte General. Introducción y Teoría Jurídica del Delito", 2ª ed., cit., pp. 172 y s.

[17] European Group on Ethics in Science and New Technologies (EGE), Opinion on *Artificial Intelligence, Robotics and 'Autonomous' Systems*, Brussels, 2018, 9 y s.

VOL. 16 N. 1 E N. 2 (2022)
DOI: 10.46294/ULPLR-RDULP.V16

5. ASPECTOS ESPECÍFICOS RELATIVOS A LA RESPONSABILIDAD PENAL DE SISTEMAS AUTÓNOMOS INTELIGENTES

Aunque hemos indicado diversas vías jurídico-penales para algunas situaciones en las que el sistema de IA o el robot produjeran lesiones en las personas o daños en las cosas, incluso por vía penal, todavía no hemos entrado en el fondo de la cuestión: el fundamento de la responsabilidad penal de estos entes capaces de actuar de forma completamente autónoma. (quiere decirse, cuando ello sea técnicamente posible y pueda ponerse en ejecución).

En este punto los juristas y otros especialistas están tomando diversas posiciones[18] , pero en realidad podrían reducirse a dos: quienes sostienen que el Derecho Penal actual no está concebido para ser aplicado a entes diferentes de los seres humanos, a salvo de lo que prevén numerosos ordenamientos jurídicos respecto a la responsabilidad penal de las personas jurídicas. Por consiguiente, conforme a la concepción y a la estructura actual de las leyes penales no sería posible atribuir responsabilidad penal a los sistemas inteligentes ni a los robots. Y hay quienes consideran, por el contrario, que estos sistemas inteligentes podrían llegar a cumplir las exigencias estructurales del Derecho Penal y en particular de la Teoría del Delito, sin perjuicio de revisar todos sus elementos configuradores e identificadores (en especial los subjetivos), incluso aunque hubiera que proceder a una adaptación o modificación del contenido de los mismos.

Expongamos entonces algunos argumentos en ambas direcciones, con varias de las objeciones que han recibido. Necesariamente, esta parte será breve.

5.1. ARGUMENTOS A FAVOR DE ATRIBUIR RESPONSABILIDAD PENAL A LOS ROBOTS, SISTEMAS INTELIGENTES Y TECNOLOGÍAS CONEXAS AUTÓNOMAS

Para quienes son partidarios de atribuir responsabilidad penal directa a los sistemas inteligentes, un nivel elevado de inteligencia y de actuar autónomo podría satisfacer en el futuro las exigencias del concepto del delito, aunque hubiera que proceder a alguna adaptación legal, probablemente profunda para algunas situaciones. Y ello sin perjuicio de que es una potencialidad que todavía se encuentra lejos de la disponibilidad de los robots y sistemas de IA actualmente conocidos.

Antes de ponderar los argumentos favorables a la imputación penal de estos sistemas, hay que asumir que la aceptación, como se apuntaba más arriba, de la responsabilidad penal de las personas jurídicas parece que podría facilitar, al menos de forma aparente, el camino hacia la responsabilidad de los entes de IA. Pero también es cierto que en ambos casos se parte de presupuestos diferentes que no hacen tan sencillo el traslado en bloque de las teorías sobre la responsabilidad de aquellas para fundamentar la imputación penal de éstos. Pero vayamos por partes.

En cuanto a la capacidad de realización de una acción jurídico-penalmente relevante, los conceptos actuales de acción que suele manejar la doctrina son lo suficientemente sencillos y prácticos como para que justamente puedan acoger los demás elementos del delito, sin aspirar a construir o mantener grandes categorías conceptuales sobre la misma, como se produjo durante los dos primeros tercios del siglo XX, lo que fue muy enriquecedor para la construcción de un sistema Penal casi completo, coherente y seguro (al menos estos eran los objetivos) y la proliferación de una elevada y rica discusión doctrinal en gran parte del territorio europeo y americano, principalmente.

Por lo que se refiere a los efectos de la pena en la conciencia moral del ente artificial, es cierto que podrían no ejercer ninguna influencia, si centramos nuestra atención en cómo es previsible que estén diseñados estos sistemas respecto a los elementos volitivos y emocionales que caracterizan al comportamiento humano. Algunos estudiosos sugieren que podríamos imaginar que el sistema de IA llegue a ser capaz de percibir o entender su mala conducta[19]. De verdade ? respecto a qué valores sociales? Si esto pudiera ser así, tampoco habría que desdeñar que el sistema inteligente pudiera interiorizar el rechazo oral –o similar- de su mala conducta. Tan prematuras algunas previsiones…

Tampoco parece previsible que se pudiera generar un supuesto sentimiento de castigo con base en la respuesta orientada a la retribución establecida por la ley. Por lo demás, la retribución como fundamento principal o exclusivo de la pena había sido casi abandonada en una buena parte de los sistemas jurídico-penales europeos.

La pena en este caso se fundamentaría de otro modo: en la gravedad del hecho delictivo realizado y perseguiría fines exclusivos de prevención especial. Este fin preventivo-especial consiste en una advertencia de que no se vuelva a delinquir para no ser castigado de nuevo con una pena, quizás más grave que la impuesta anteriormente por la previa infracción cometida. Estos fines estarían orientados

[18] V. Eric Hilgendorf, Carsten Kushe, Brian Valerius, *Computer- und Internetstrafrecht. Ein Grundriss,* ISBN: 978-3-662-59446-, Springer, 2022, pp. 13 y ss.

[19] M.B. Magro, *Robot, Cyborg, e Intelligenze Artificiali,* in A. Cadoppi, S. Canestrari et a.t. (Dirs.), Cybercrime, UTET, Milano, 2019, p. 1204.

VOL. 16 N. 1 E N. 2 (2022)
DOI: 10.46294/ULPLR-RDULP.V16

fundamentalmente a la separación y aislamiento de la sociedad del perpetrador del hecho delictivo –haciéndole inocuo o "neutralizando" su capacidad de producir la lesión de un bien jurídico penal-, es decir, sobre los sistemas inteligentes que probada y objetivamente constituyeran una fuente de riesgo delictivo, o dicho en términos menos conflictivos, que comportaran una peligrosidad criminal, de recidiva delictiva.

Aquí los fines preventivo-especiales de advertencia, enmienda y rehabilitación social del ente inteligente en la práctica no serían posibles, salvo si aceptamos como reacción preventivo-especial la reprogramación del sistema inteligente, pues en sentido estricto no sería necesario satisfacerlos ni tendrían sentido. Como tampoco tendría sentido alguno la prevención general, sea negativa o positiva.

Algunos autores barajan la posibilidad de readaptar la teoría del delito a la fenomenología criminal de la IA. Por ejemplo, reconducir la culpabilidad a terrenos más objetivos u objetivables, lo que, alegan, contribuiría al mismo tiempo a resolver algunos problemas actuales de imputación –subjetiva- de los seres humanos.

5.2. ARGUMENTOS EN CONTRA

En primer lugar, los autores que se manifiestan contrarios a admitir que los sistemas inteligentes, incluidos los autónomos, puedan ser penalmente responsables, ponen en duda que estos entes artificiales puedan realizar una acción humana abarcable por el concepto de delito: actuar bajo la dirección de la voluntad dirigida a la consecución de un fin, sea éste constitutivo de delito o no. Podría calificarse al sistema inteligente como sujeto activo de la acción correspondiente dirigida por la voluntad, típificada por la ley penal? o su decisión estará causalmente determinada como consecuencia del procesamiento necesario de un conjunto de algoritmos, aunque sea de una forma no prevista inicialmente en la fase de creación y programación del sistema?

Aun aceptando que los entes artificiales inteligentes fueran capaces de actuar de forma completamente autónoma orientados a un fin, son seres sin conciencia moral a los que la imposición de una pena no podría hacerles sentir culpa moral por el delito cometido, pero tampoco culpa jurídica respecto al hecho concreto. En consecuencia, la pena no tendría ningún efecto retributivo sobre ellos ni sería capaz de generar ningún sentimiento de castigo. ¿Qué sentido debería tener entonces la pena?

Es dudoso que el fin preventivo de la pena, tanto de prevención general como especial, pudiera tener algún efecto relevante en el ente inteligente, salvo que en su programación se hubiera introducido algún mecanismo eficiente de abstención de cometer un hecho prescrito como delito por la ley y tal mecanismo no pudiera ser alterado por el propio sistema. Sería un modo de aplicación del CHS.

En cualquier caso, la pregunta principal sigue quedando en pie, al menos en el momento actual, a la vista de las capacidades reales o alcanzables a medio plazo por los sistemas inteligentes.

En primer lugar, ha de tenerse en cuenta que los sistemas inteligentes actuales están especializados en una sola tarea o en una pluralidad de ellas, pero limitadas y próximas entre si, y en esa actividad pueden demostrar una "inteligencia" superior incluso a la humana, pero aquéllos no poseen unas cualidades intelectuales múltiples, como le ocurre al ser humano; no disponemos todavía de una generación de sistemas "superinteligentes".

El lado emocional desempeña asimismo un papel decisivo en el funcionamiento de la inteligencia humana, en concreto en la toma de decisiones y en las deliberaciones previas a éstas. Aunque se está trabajando también en los aspectos emocionales de los sistemas inteligentes, puede sostenerse en la actualidad que esta cualidad forma parte esencial de los sistemas inteligentes que conocemos? Podemos aceptar que la respuesta debería ser negativa en la actualidad.

La culpabilidad, como elemento también esencial del delito, consistente en un reproche jurídico al sujeto activo del delito por haber obrado de manera antijurídica, tampoco se vería satisfecha en relación con estos entes, pues este juicio de reproche no produciría ningún efecto sobre su actitud respecto al delito cometido, es decir, una vez realizado un hecho respecto al cual el sistema inteligente "conoce" que es contrario a la ley penal.

La culpabilidad supone al mismo tiempo como presupuesto, aunque haya sido tradicionalmente objeto de profundas y a veces irreconciliables discusiones entre los penalistas, el libre albedrío del ser humano. Podrá predicarse lo mismo de sistemas inteligentes y autónomos con este máximo nivel de autonomía o con el que podría desarrollarse en un futuro delimitable? En el momento actual las decisiones que toman los sistemas inteligentes son el resultado de un proceso algorítmico fundamentalmente causal, lejos de interferencias de naturaleza axiológica.

Estos y otros argumentos semejantes no implican que haya que renunciar a cualquier forma de reacción jurídico-penal frente a los hechos tipificados como delito cometidos por sistemas de IA. Estos habría que imputarlos a los seres humanos a los que de un modo u otro corresponde una esfera de dominio o control sobre aquéllos y fallaron en el cumplimiento de los deberes relacionados con dicho dominio o control (*compliance*), bien en el conjunto de todas sus actividades, bien al menos en su origen.

Esta conclusión no significa abandonar en estos casos el principio de culpabilidad

VOL. 16 N. 1 E N. 2 (2022)
DOI: 10.46294/ULPLR-RDULP.V16

en relación con los humanos a los que se imputaría penalmente. En ellos es también necesaria la concurrencia de dolo o imprudencia, tener capacidad de culpabilidad y poder realizar sobre ellos un juicio de reproche jurídico por haber actuado en contra del Derecho.

Como apuntaba más arriba, la extensión en numerosos ámbitos sociales de la IA exigirá ir estableciendo un aparato normativo que los regule en multitud de aspectos antes de su puesta a disposición de los usuarios, tanto sean grandes corporaciones y los poderes públicos como los ciudadanos de modo individual. Dentro de este marco regulativo, fijado en general por iniciativa de las administraciones públicas, es previsible que se establezcan obligaciones de supervisión y control a cargo de los diseñadores, fabricantes, distribuidores y usuarios finales.

De este modo estos delitos podrían constituirse como delitos de infracción de deberes especiales cometidos por personas físicas –y, en su caso, por personas jurídicas-. Para satisfacer las exigencias del principio de legalidad de los delitos, podría el legislador introducir una cláusula general de cobertura de modo semejante a lo que suele hacerse con los delitos de comisión por omisión.

Frente a los robots y a los sistemas inteligentes en cuanto tales, bastaría con aplicar las llamadas consecuencias accesorias oportunas, introduciendo, si fuera preciso, alguna Frente a los robots y a los sistemas inteligentes en cuanto tales, bastaría con aplicar las llamadas consecuencias accesorias oportunas, introduciendo, si fuera preciso, alguna específica adaptada a las necesidades de respuesta a los sistemas inteligentes involucrados en la comisión de un delito. El presupuesto sería la comprobación de la existencia de peligrosidad objetiva en el sistema, es decir, un riesgo objetivo de volver a cometer un delito pro o a través del sistema inteligente, autónomo o no. Las medidas a tomar por el juez encargado del proceso penal contra los seres humanos acusados girarían en torno a la eliminación del foco de peligro que constituye sistema inteligente, aplicando una o varias de las siguientes: bloqueo del software, retirada del mercado, reprogramación, destrucción del sistema inteligente o robot involucrados en el delito o destrucción en su conjunto del modelo diseñado. Todo ello con independencia de las penas aplicables a los seres humanos condenados por el delito de que se trate.

Otros casos podrían resolverse por medio de la normativa sancionadora a cargo de las administraciones públicas. Estas podrían consistir en el bloqueo temporal o definitivo del sistema, o mejor en su reprogramación con el fin de prevenir la reiteración de algunas decisiones relacionadas con el hecho delictivo. Y al ser humano infractor se le podría imponer una multa y/o la retirada temporal o definitiva de la licencia de diseño, fabricación, venta y distribución o uso.

6. LA ATRIBUCIÓN DE RESPONSABILIDAD PENAL A LOS ROBOTS AUTÓNOMOS Y A LOS SISTEMAS INTELIGENTES: EL CONTROL HUMANO DE LOS SISTEMAS DE IA COMO ENFOQUE ALTERNATIVO PREVIO

Debemos ser conscientes de que desde el momento en que se introdujo la responsabilidad penal de las personas jurídicas en los Estados miembros de la Unión Europea, también sería posible asumir –sin olvidar las reformas necesarias, con el fin de respetar el principio de legalidad- la atribución de responsabilidad penal a los sistemas de IA, prescindiendo de los elementos de un Derecho Penal concebido para atribuir responsabilidad sólo a los seres humanos. Sin embargo, tal y como se argumenta en este trabajo, no parece que sea necesario a corto o medio plazo, ni está en consonancia con el planteamiento general de que los humanos deben mantener el dominio sobre las tecnologías autónomas emergentes, como los robots y otros sistemas de IA y tecnologías conexas, en particular cuando lleguen a operar de forma autónoma.

Es necesario continuar el debate sobre una realidad que ya tenemos a las puertas y a la que tendremos que dar respuestas, también jurídicas, y tenerlas preparadas al menos en sus aspectos más básicos.

Se está extendiendo la posición, que partió de los debates en torno a los Sistemas de Armas Autónomas Letales (LAWS)[20] y a los sistemas de vehículos autónomos, de que el Control Humano Significativo (CHS, MHC)[21] es esencial para atribuir la responsabilidad moral. Esto significa que los seres humanos -y no los sistemas computacionales y sus algoritmos- deben mantener el control final, pudiendo así ser responsables moralmente y quizás también legalmente de él.

El CHS debería convertirse en un estándar de lo que debería ser la actitud del ser humano hacia cualquier tecnología que pueda aumentar su autonomía respecto a los humanos que la crearon o utilizaron y puedan comportar riesgos más elevados de lesión de bienes jurídico-penalmente protegidos. Este es nuestro deber como especie viviente que de alguna manera domina el planeta y todos los seres vivos que lo habitan, a pesar de que desgraciadamente a menudo abusa de él.

El Reglamento General de Protección de Datos (RGPD) de la Unión Europea[22] ha adoptado un enfoque similar. Excluye que las decisiones en el contexto de los sistemas de IA puedan tomarse exclusivamente de forma automatizada. Esta

[20] V. con más detalle United Nations Institute for Disarmament Research (UNIDIR), *The Weaponisation of Increasingly Autonomous Technologies: Considering how Meaningful Human Control might move the discussion forward,* p. 5 y s.

[21] Control Humano Significativo *(Meaningful Human Control)* es una expresión creada por el colectivo Article 36, Killer Robots: *UK Government Policy on Fully Autonomous Weapons,* April 2013 p. 4: http://www.article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf.

[22] Reglamento (UE) 2016/679 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos.

disposición considera que una decisión automatizada es un proceso de toma de decisiones basado únicamente en el tratamiento automatizado de datos[23].

El concepto de CHS (MHC) podría ir más allá de su origen en el control humano de los sistemas de armas letales autónomas y abarcar otras tecnologías cada vez más autónomas, en particular los sistemas autónomos inteligentes. De acuerdo con la UNICIDIR, el concepto de CHS descrito por la Organización del Artículo 36 se distingue de la caracterización tradicional de un ser humano "en o sobre el bucle" ("*in or on the loop*"), ya que ofrece más precisión (control frente al algo ambiguo concepto de "bucle" o el más pasivo de "juicio"); hace hincapié explícitamente en la calidad del control ("relevante") y concede implícitamente la responsabilidad a los agentes humanos por las decisiones relativas a cada acción individual[24].

No obstante, también es cierto que en el marco del Derecho Penal continúa siendo una expresión bastante imprecisa, en particular por el término "significativo" (o: "relevante"), que es, sin embargo, la palabra clave de este concepto. La seguridad jurídica, como derivación del principio de legalidad, que es una de las bases fundamentales del Derecho Penal contemporáneo occidental, exige que la descripción de las categorías jurídico-penales sea lo suficientemente concreta (principio de taxatividad). Por ello, y porque reconocemos el interés que la categoría del CHS puede tener para abordar los sistemas inteligentes desde el punto de vista jurídico-penal, sería necesario estudiarla en profundidad, lo que queda para otro momento.

Entonces, cómo se podría trasladar el CHS al Derecho Penal como categoría dogmática en relación con la IA? En primer lugar, podría hacerse mediante revisiones y validaciones de la decisión concreta que el sistema inteligente pueda proponer o ejecutar. En cualquier caso, deberían tener un carácter preventivo en el contexto del cumplimiento de la normativa, antes de introducir en el mercado o autorizar el funcionamiento del sistema autónomo inteligente con cierta capacidad -relevante- de causar daños.

7. MIRANDO AL FUTURO

Nos encontramos en una fase de discusión sobre la razón de ser de la responsabilidad penal de los robots y sistemas de IA que es todavía preliminar y necesariamente provisionales serán las conclusiones y propuestas que puedan hacerse en el debate actual. Por tal motivo, me he mostrado claro pero prudente en mis propias reflexiones.

El concepto de CHS o cualquier otro similar que pueda desarrollarse en el futuro puede trasladarse al ámbito del Derecho. Se trata de una premisa que puede ayudar a establecer la posición de dominio que corresponde a los seres humanos respecto a las crecientes tecnologías autónomas; es un requisito metodológico para establecer el correcto enfoque jurídico de la relación entre los seres humanos y las máquinas inteligentes. La dogmática penal europea viene aceptando de forma pacífica desde hace décadas el concepto de dominio del hecho que posee un ser humano, ejecute o no personalmente la acción y permite imputarle penalmente ese hecho (Claus Roxin)[25] . Un camino similar deberíamos ser capaces de desarrollar partiendo del CHS, para poder imputar a un ser humano.

A corto y medio plazo el sistema penal más extendido puede dar respuestas satisfactorias, centrando adecuadamente la atribución de responsabilidad penal a los seres humanos implicados, por la vía de las medidas de seguridad preventivas y no la de respuestas estrictamente penales (punitivas) para hacer frente a los riesgos –graves- de los sistemas inteligentes autónomos. Se trata de aplicar el modelo de cumplimiento con las correspondientes adaptaciones.

Estas medidas tendrían como requisito previo el riesgo objetivo e irrazonable que presenta el sistema inteligente de realizar un acto tipificado como delito por la ley penal.

Si el daño producido no constituyera delito, las medidas no serían penales ni ordenadas por los jueces, sino por la autoridad administrativa correspondiente, de acuerdo con lo que prevea la ley. En cualquier circunstancia, sin perjuicio de la indemnización por daños y perjuicios producidos de acuerdo con la normativa civil correspondiente[26] .

Como yo mismo me he propuesto en este trabajo y en otros de temática próxima, quiero evitar llegar a territorios que puedan rozar el llamado Derecho ficción, y no voy a plantearme la hipótesis (esta posición comporta también metodología) de que pudiera desarrollarse –legalmente- un sistema inteligente verdaderamente autónomo, pues probablemente si comportaría adaptaciones muy relevantes del Derecho Penal de los humanos o dar lugar a un paralelo Derecho Penal de

[23] Art. 22.1: "Todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar".

[24] United Nations Institute for Disarmament (UNIDIR), *The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control might move the discussion forward*, cit.

[25] V. referencia a Roxin sobre este punto en Sánchez Lázaro, *Autoría y participación*, cit. p. 172.

[26] En desacuerdo con Magro, *Robot, Cyborg, e Intelligenze Artificiali*, p. 1211.

VOL. 16 N. 1 E N. 2 (2022)
DOI: 10.46294/ULPLR-RDULP.V16

entes no humanos (personas jurídicas, TIC, incluida IA, seres vivos no humanos[27], seres extraterrestres, etc.). La velocidad actual de las tecnologías, los resultados encadenados a los que pueden dar lugar, hacen ociosa cualquier predicción más allá de los años más cercanos (tampoco deberíamos señalar una fecha fija, convencido del error seguro) y obligan a los juristas (y a otros especialistas) a seguir trabajando en paralelo con el avance científico, de modo semejante a cómo ocurrió hace unas décadas con el estudio y exploración del genoma humano, la "revolución genética"[28].

[27] Sobre la inevitabilidad de adaptaciones del Derecho en esta hipótesis. Hilgendorf, Kushe, Valerius, *Computer- und Internetstrafrecht. Ein Grundriss,* cit., pp. 300 y ss.

[28] V. sobre las nuevas reflexiones y metodológicas que suscitó la puesta en marcha y conclusión del llamado Proyecto Genoma Humano, Carlos María Romeo Casabona, *Los genes y sus leyes. El Derecho ante el genoma humano,* Cátedra de Derecho y Genoma Humano, Fundación BBVA – Fundación Foral de Bizkaia, Universidad de Deusto y Universidad del País Vasco/EHU y Editorial Comares, Bilbao-Granada, 2002, pássim (pp. 11 y ss.).