PREMIO SUSANA HUERTA DE DERECHO PENAL

# El sesgo de automatización en la supervisión humana de sistemas de inteligencia artificial: el riesgo permitido en el nuevo marco normativo europeo

CARLOS TRINCADO CASTÁN
Doctorando en Derecho. Universidad de La Laguna

## RESUMEN

*La supervisión humana es una de las medidas más relevantes para el control de los riesgos derivados de la introducción de la inteligencia artificial (IA) en la vida social. Uno de los principales desafíos para que la supervisión humana pueda ser efectiva es el sesgo de automatización, la tendencia a confiar de forma rutinaria o por complacencia en el funcionamiento de los sistemas que automatizan un proceso. Cuando se producen daños como consecuencia de la materialización de este riesgo, la responsabilidad penal del supervisor humano vendrá determinada por la configuración de su posición de garante y los límites del riesgo permitido en el ejercicio de sus funciones, presentándose diversos escenarios en función de la posibilidad de aplicar la nueva regulación europea sobre IA.*

Palabras clave: *sesgo de automatización, delitos imprudentes, riesgo permitido, diligencia debida, inteligencia artificial, Reglamento de IA.*

## ABSTRACT

*Human oversight is one of the most relevant measures to control the risks arising from the introduction of artificial intelligence (AI) into social life. One of the main challenges to effective human oversight is automation bias, the tendency to routinely or complacently rely on the operation of systems that automate a process. When harm occurs as a result of the materialisation of this risk, the criminal liability of the human supervisor will be determined by the configuration of his or her position of the supervisor´s position of guarantor, the content of the objective duty of care and the thresholds of permissible risk in the exer-*

*cise of its functions, with different scenarios arising depending on the possibility of application of the new European regulation on AI.*

Key words: *automation bias, negligence crimes, allowed risk, due diligence, artificial intelligence, AI Act*



## I. INTRODUCCIÓN

La inteligencia artificial (en adelante IA) como disciplina científica vio impulsado su desarrollo en la década de 2010 como consecuencia de la aplicación y perfeccionamiento de las técnicas de aprendizaje automatizado (*machine learning*) (1), que se vieron especialmente beneficiadas por un contexto caracterizado por la existencia de una cada vez mayor potencia de computación de los sistemas informáticos y la disponibilidad de datos a gran escala (*big data*) (2). El desarrollo tecnológico de la IA ha dado su siguiente paso en la década de 2020 con los sistemas de IA de uso general, que pueden ser aplicados en una amplia variedad de finalidades y tareas (en lugar de estar

(1) A pesar de que estos avances se han producido en la última década, se basan en teorías y fundamentos desarrollados por la IA como disciplina científica desde mediados del siglo pasado. En particular, los fundamentos teóricos sobre los que se han desarrollado las redes neuronales artificiales datan de los años 60, así como los trabajos sobre la retropropagación y las estrategias de aprendizaje no supervisado hicieron lo propio en los años 80, sin que pudiesen alcanzar estas técnicas su potencial como consecuencia de las limitaciones tecnológicas de la época.

(2) KAUTZ, H. A., «The third AI summer: AAAI Robert S. Engelmore memorial lecture», *AI Magazine*, vol. 43, núm. 1, 2022, p. 111.

confinados a un uso o aplicación específico) (3), de los cuales los modelos fundacionales, en particular los grandes modelos de lenguaje [*Large Language Models*, aunque no limitados a los mismos (4)], han permitido el desarrollo de los sistemas de IA generativa, cuya introducción en el mercado y la vida social ha supuesto una de las mayores irrupciones tecnológicas de las últimas décadas. Sin embargo, está aún por ver si estos avances terminan por cumplir las enormes expectativas que se han generado alrededor de esta tecnología, o si, por el contrario, la disciplina puede volver a estancarse en un nuevo «invierno de la IA» (5).

El vertiginoso desarrollo de la IA en los últimos años tiene asociados potenciales beneficios económicos, medioambientales y sociales en múltiples sectores económicos y actividades de la vida social (6). Al mismo tiempo, se ha venido advirtiendo que su utilización puede generar riesgos y menoscabar intereses públicos y derechos fundamentales (7). La literatura pone de manifiesto la existencia de diversos riesgos asociados al uso de la IA, como los sesgos discriminatorios en su funcionamiento (8) (como resultado del proceso de aprendizaje automatizado, de los datos utilizados para su entrenamiento o por el propio diseño del sistema); la imprevisibilidad en la toma de decisiones derivada de su autonomía (9); la opacidad y la falta de transparen-

(3) No debe confundirse este término con los de inteligencia artificial general (IAG) o IA fuerte (*strong AI*) que hacen referencia a una IA con capacidades de razonamiento general equiparables o que incluso superan a las de los seres humanos. Si bien el desarrollo de sistemas de IA de uso general es uno de los últimos avances de la IA como disciplina y es un paso en la dirección para alcanzar la IAG, existe cierto consenso a la hora de considerar que esta meta no tiene un horizonte temporal definido, ni cercano.

(4) Sobre una taxonomía de las distintas estrategias y aproximaciones técnicas para el desarrollo de sistemas de IA de uso general: *vid.* Triguero, I., *et al.*, «General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, societal implications and responsible governance», *Information Fusion*, vol. 103, 2024, pp. 6-9.

(5) Lazcoz Moratinos, G., *Sistemas de inteligencia artificial en la asistencia sanitaria: cómo garantizar la supervisión humana desde la normativa de protección de datos*, AEPD, Madrid, 2022, p. 2.

(6) Considerando 4 del Reglamento (UE) 2024/1689 de Inteligencia Artificial.

(7) Considerando 5 del Reglamento (UE) 2024/1689 de Inteligencia Artificial.

(8) Valls Prieto, J., «Sobre la responsabilidad penal por la utilización de sistemas inteligentes», *Revista Electrónica de Ciencia Penal y Criminología*, núm. 24-27, 2022, p. 4.

(9) Palma Herrera, J. M., «Inteligencia artificial y neurociencia. Algunas reflexiones sobre las aportaciones que pueden hacer al Derecho Penal», Peris Riera, J. M., Massaro, A. (coords.), *Derecho penal, inteligencia artificial y neurociencias*, Roma Tre-Press, Roma, 2023, p. 252.

cia de la operativa del sistema (10), o la dificultad para el seguimiento y trazabilidad del proceso de desarrollo y de su funcionamiento, siendo un desafío la detección de errores y la imputación de los mismos a un actor concreto en el ciclo de vida del sistema de IA (conocido este como el problema de las múltiples manos o *many hands problem*) (11), entre otros.

No todos estos riesgos son específicos o exclusivos del ámbito de la IA. Muchos de ellos son riesgos comunes al desarrollo tecnológico en el contexto de la sociedad del riesgo, como la relevancia de los fallos técnicos y su ubicación como riesgos permitidos o penalmente relevantes, la importancia de las decisiones de distribución de riesgos o la producción de daños a largo plazo en los que existe incertidumbre sobre la relación causa-efecto entre el fallo y el resultado (12). Con todo, también existen riesgos específicamente asociados al desarrollo y uso de sistemas de IA derivados de sus características particulares, concretamente los ya señalados en relación con la autonomía en su funcionamiento y sus capacidades de aprendizaje que les permiten fijar nuevos objetivos, distintos de aquellos para los que fueron inicialmente diseñados y programados (13).

## II. LA REGULACIÓN DE LA IA EN LA UNIÓN EUROPEA

Como respuesta al intenso desarrollo de la IA en la última década, con el objetivo de aprovechar adecuadamente sus beneficios y establecer garantías frente a los riesgos específicos asociados su uso, se han

(10) Giannini, A., Kwik, J., «Negligence failures and negligence fixes. A comparative analysis of criminal regulation of AI and autonomous vehicles», *Criminal Law Forum*, núm. 34, 2023, p. 54.

(11) Salvadori, I., «Agentes artificiales, opacidad tecnológica y distribución de la responsabilidad penal», *Cuadernos de política criminal*, n. 133, mayo 2021, pp. 144-145 y 160.

(12) Silva Sánchez, J. M., *La expansión del Derecho penal: Aspectos de la Política criminal en las sociedades postindustriales*, 2.ª edición, Editorial Bdef, Montevideo-Buenos Aires, 2006, pp. 15-17.

(13) Debe tenerse en cuenta, no obstante, que estas características no están presentes en todos los sistemas de IA, así como que tampoco en todos aquellos en las concurren lo hacen con la misma intensidad. Los estudios relacionados con las implicaciones jurídicas derivadas del uso de sistemas de IA tienden a enfocar sus análisis en torno a una categoría concreta de esta tecnología: los sistemas de IA basados en técnicas de aprendizaje automatizado (*machine learning*), que son a los que habitualmente se asocian las características de opacidad, complejidad, autonomía e imprevisibilidad en su funcionamiento, mientras que otras categorías de sistemas de IA (como los sistemas basados en conocimientos o en modelos lógicos) no presentan estas características o, al menos, no con la misma intensidad y relevancia.

ido sucediendo iniciativas por parte de diversos actores a nivel estatal y supranacional para regular el desarrollo y uso de esta tecnología. Han sido especialmente notables los esfuerzos realizados por distintas entidades supranacionales en el último lustro, que han cristalizado a nivel europeo en el Reglamento de Inteligencia Artificial (Reglamento 2024/1689, del Parlamento Europeo y del Consejo, de 13 de junio de 2024, en adelante el RIA), entre otras iniciativas (14).

El RIA ha sido la norma cuya tramitación ha tenido mayor protagonismo y la que mayor expectación ha generado hasta su aprobación. Si bien este Reglamento no es la primera norma aplicable a sistemas de IA (15), sí que se ha configurado como la primera regulación de carácter transversal aplicable al desarrollo y utilización de la IA, mediante la que se establecen definiciones comunes, un catálogo de obligaciones para los distintos actores de la cadena de valor de los sistemas de IA, así como un sistema institucionalizado de gobernanza y supervisión de su cumplimiento (16). Si bien el RIA entró en vigor el pasado 1 de agosto de 2024, la mayor parte de sus disposiciones no serán aplicables hasta que transcurran dos años desde esta fecha (17).

---

(14) Aunque no será objeto de análisis en este trabajo, cabe destacar también, en el contexto del Consejo de Europa, el Convenio Marco sobre inteligencia artificial y derechos humanos, democracia y Estado de Derecho de 5 de septiembre de 2024.

(15) En la medida que los sistemas de IA pueden ser considerados como sistemas de información, les resulta de aplicación la normativa que los regula, tanto aquella que presenta un carácter general (como el Convenio de Budapest sobre ciberderdelincuencia; o el artículo 22 del RGPD, en relación con el tratamiento automatizado de datos), así como aquella de carácter sectorial (como el Reglamento Europeo de Productos Sanitarios o el Reglamento Europeo de Servicios Digitales). En relación con la aplicación de las normas relativas a los sistemas de información a los sistemas de IA, *vid.* Picotti, L., «Traditional Criminal law categories and AI: crisis or palingenesis? General report», *Revue Internationale de Droit Pénal*, vol. 94, iss. 1, 2023, p. 15.

(16) El Convenio Marco del Consejo de Europa presenta un enfoque y contenido más limitado y menos exhaustivo que el del RIA, en la medida que no establece obligaciones específicas, más allá de compromisos de los Estados parte para aprobar regulaciones de acuerdo con lo dispuesto en el Convenio Marco. No obstante, este instrumento ha supuesto un importante avance en la regulación de la IA a nivel global, en la medida que países ajenos al contexto europeo como EE. UU. o Japón han participado en su elaboración y, previsiblemente, se sumarán como signatarios, fomentando la adopción de un enfoque armonizado a nivel internacional de esta tecnología que presenta potencial para producir efectos transfronterizos.

(17) Algunos capítulos del Reglamento serán de aplicación antes (a los seis meses, los relativos a los usos y prácticas prohibidas) y otros con posterioridad (a los tres años las disposiciones relativas a los requisitos y obligaciones de los operadores de sistemas de IA de alto riesgo, con el objetivo de dar tiempo a las distintas partes para realizar los preparativos y adaptaciones a la normativa necesarios), conforme a lo dispuesto en el artículo 113 RIA.

El RIA no tiene como objetivo regular la IA como tecnología, sino que, partiendo de un enfoque basado en el riesgo (*risk-based approach*) regula aquellos usos de los sistemas de IA que presentan mayores riesgos para la salud, la seguridad y los derechos fundamentales de las personas (art. 1 RIA), «adaptando el tipo y contenido de las normas a la intensidad y el alcance de los riesgos que puedan generar» (considerando 26 RIA). De este modo, el RIA establece obligaciones más restrictivas para aquellos usos que conllevan mayores riesgos, siendo menos restrictivo o incluso dejando fuera del ámbito de aplicación de la norma aquellas aplicaciones y ámbitos que presenten riesgos bajos o insignificantes, dando lugar a una clasificación cuatripartita: 1) usos y prácticas prohibidas, 2) sistemas de IA de riesgo alto, 3) sistemas de IA de riesgo limitado y 4) sistemas de IA de riesgo bajo o nulo.

El artículo 5 RIA establece un régimen de prohibiciones de desarrollo y despliegue de sistemas de IA para determinados usos y prácticas para los que se presume que suponen un riesgo inaceptable para la sociedad (como el *social scoring*, el reconocimiento biométrico en tiempo real, el uso de sistemas de policía predictiva basadas en el perfilado de personas, etc.)(18).

Respecto a los sistemas de alto riesgo, el artículo 6 RIA clasifica como tales aquellos sistemas de IA que deban ser sometidos a una evaluación de conformidad de acuerdo con determinadas normas comunitarias (recogidas en el Anexo I RIA (19)) o que formen parte de determinados ámbitos (Anexo III RIA (20)), salvo que se demuestre que el concreto sistema de IA no presenta riesgos para la vida, la salud o los derechos fundamentales de las personas, o que no influye en una toma de decisiones de forma significativa (art. 6.3 RIA). Para los sistemas de IA de alto riesgo se establece un conjunto de obliga-

(18) Sin perjuicio de que se prevean excepciones a tales prohibiciones. Por ejemplo, se permite el uso de sistemas de identificación biométrica remota en tiempo real, bajo determinadas condiciones, para determinados fines de «garantía de cumplimiento del Derecho» (la traducción que se ha utilizado en el RIA para el término anglosajón *law enforcement*).

(19) Destacan aquí los Reglamentos (UE) 2017/745 y 2017/746, sobre productos sanitarios y productos sanitarios *in vitro* respectivamente, la Directiva 2006/42/CE relativa a máquinas.

(20) Destacan aquí los ámbitos de infraestructuras críticas, biometría, garantía del cumplimiento del Derecho y administración de justicia y procesos democráticos. Otros ámbitos considerados de alto riesgo son los de migración, educación, empleo y acceso a servicios privados y públicos esenciales. Como se puede apreciar, a excepción del de infraestructuras críticas, no son ámbitos en los que se vayan a producir lesiones o daños contra la salud o la integridad física, sino que están más bien vinculados a tratos discriminatorios (aunque, de forma indirecta, pueden dar lugar a vulneraciones de la integridad psíquica y moral de las personas).

ciones dirigidas a los distintos actores de la cadena de valor de los sistemas de IA (proveedores, responsables del despliegue, importadores, representantes autorizados, etc.) para su desarrollo, comercialización y despliegue (artículos 8 a 27 RIA).

En cuanto a los sistemas de IA de riesgo limitado (art. 50 RIA), se consideran como tales aquellos diseñados para interactuar con personas físicas (art. 50.1 y.3 RIA) y para la generación o manipulación de contenidos (art. 50.2 y.4 RIA), estableciéndose obligaciones de transparencia dirigidas a garantizar que las personas que interactúen con el sistema sean conscientes de que se trata de una IA, así como el marcado o señalización de que los contenidos han sido generados o manipulados por un sistema de IA, respectivamente.

Todo aquel sistema de IA que no se pueda subsumir en ninguna de las categorías mencionadas se podrá considerar como un sistema de IA de riesgo bajo o insignificante, al menos en el ámbito del RIA. Formalmente podrán ser considerados como sistemas de IA de acuerdo con la definición prevista en el artículo 3 (1) RIA, pero su desarrollo y despliegue no estarán materialmente sujetos a las obligaciones previstas en el RIA, quedando en manos de los distintos actores privados y públicos la creación de códigos de conducta para el sometimiento voluntario a las normas y disposiciones del Reglamento (art. 95 RIA)(21).

## 1. **Supervisión humana como medida de gobernanza y control de riesgos**

Uno de los principios que se han venido considerando como esenciales desde que se comenzó a debatir acerca de cómo regular la IA es el de su control y supervisión humana, esto es, la necesidad de que los sistemas de IA estén al servicio y bajo el control de, al menos, una persona física(22). Existen diversas formas en las que este control se puede hacer efectivo: desde mecanismos de gobernanza, el establecimiento de requisitos previos a la introducción del sistema de IA en el mercado, obligaciones de seguimiento posterior a la comercialización, etc., pero la principal forma en la que este control se ha considerado

(21) El RIA establece una categoría adicional para los modelos y sistemas de IA de uso general, para los que prevé una regulación específica (artículos 51 a 56 RIA). En esta categoría se pueden incluir los sistemas de IA generativa que se han popularizado en los últimos años, como *ChatGPT*, *Gemini* o *Copilot*.

(22) Los primeros trabajos de la UE relacionados con la regulación de la IA ya hacían referencia a la necesidad de que pudiese ser «controlada y vigilada adecuadamente por seres humanos». *Vid.* Grupo de expertos de alto nivel sobre inteligencia artificial de la Comisión Europea, *Directrices éticas para una IA fiable*, 2019, p. 20.

que se puede hacer efectivo es mediante la supervisión directa del funcionamiento del sistema por una persona o grupo de personas.

Se han propuesto diversos modelos y configuraciones sobre la manera en la que el ser humano puede desarrollar tal supervisión: *human in the loop*, en el que para funcionar el sistema de IA requiere que el humano confirme, acepte o ejecute la acción o recomendación propuesta por el sistema; *human on the loop,* si el sistema puede funcionar y ejecutar las tareas que le sean encomendadas sin intervención de una persona, pero existiendo la posibilidad de que ésta intervenga cancelando o modificando su funcionamiento, y *human out of the loop,* cuando el sistema de IA funciona sin intervención de un ser humano, limitándose la labor del supervisor a realizar una evaluación posterior de su rendimiento y funcionamiento (23).

La relevancia de la supervisión humana como medida de control de riesgos tiene su reflejo en el RIA, siendo uno de los requisitos específicamente previstos para los sistemas de IA de alto riesgo, que deberán estar desarrollados de tal forma que sea posible su supervisión por personas físicas (art. 14.1 RIA), así como deberán estar encomendadas tales labores de supervisión a personas físicas con la competencia, formación y autoridad necesarias para el ejercicio de sus funciones (art. 26.2 RIA). No obstante, esta importancia de la supervisión humana no debe ser utilizada como una justificación para convertir a los sujetos que realicen tales tareas en chivos expiatorios a los que responsabilizar de resultados lesivos e indeseados (24). Por mucho que sean los sujetos que habitualmente estarán más próximos a los riesgos, no siempre serán los que tengan control sobre los mismos o una posibilidad real de controlarlos (25). En este contexto, el Derecho penal puede aportar instrumentos para delimitar cuándo se puede considerar realmente responsable a la persona que realiza las labores de supervisión humana.

El objetivo de la supervisión humana es prevenir o reducir al mínimo los riesgos para la salud, la seguridad o los derechos fundamentales que pueden surgir cuando se utiliza un sistema de IA (así se reconoce expresa-

(23) Harbers, M., Peeters, M. M. M., Neerincx, M. A., «Perceived Autonomy of Robots: Effects of Appearance and Context», en Aldinhas Ferreira, M., Silva Sequeira, J., Tokhi, M., E. Kadar, E., Virk, G. (eds.) *A World with Robots. Intelligent Systems, Control and Automation: Science and Engineering*, Springer, Cham, 2017, pp. 19-33.

(24) Beck, J., Burri, T., «From «Human Control» in International Law to «Human Oversight» in the New EU Act on Artificial Intelligence», *Forthcoming* en Amoroso, D., Santoni de Sio, F. (eds.*), Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*, Elgar, 2023, p. 17. Disponible en: https://ssrn.com/abstract=4236554 (recuperado el 15 de septiembre de 2024).

(25) Paredes Castañón, J. M., *El riesgo permitido en Derecho Penal*, Ministerio de Justicia e Interior, Madrid, 1995, p. 377

mente en el artículo 14.2 RIA). Sin embargo, en determinados contextos no será posible el desempeño de una supervisión humana efectiva como, por ejemplo, cuando el tiempo disponible para supervisar la decisión no pueda ser asumido por un ser humano (es decir, el tiempo necesario para dar una respuesta es menor que el tiempo que requeriría una persona diligente para realizar el comportamiento necesario (26)) o en los que la información disponible para el operador del sistema sea reducida o esté limitada a la que el propio sistema de IA provee, sin que se pueda acudir a otros instrumentos, herramientas o fuentes de información alternativas (27). Deben tenerse en cuenta, adicionalmente, otros factores que afecten a contextos o sectores específicos. En el ámbito clínico, por ejemplo, se han señalado a este respecto las prácticas de medicina defensiva (28). En aquellos contextos en los que la supervisión humana tenga un papel más limitado, deberán concurrir con mayor intensidad y exigencia las demás verificaciones y medidas de gobernanza cuanto menor sea el nivel de supervisión que se pueda ejercer (29).

Por otra parte, la supervisión del sistema de IA no implica una reducción unilateral de los riesgos, sino que da lugar a la emergencia de otros distintos (30), algunos derivados de la interacción del supervisor con el sistema como consecuencia de defectos en el diseño, desarrollo y entrenamiento del sistema, como la opacidad y la falta de explicabilidad de su funcionamiento; otros, derivados de la ineficacia de los seres humanos como supervisores (31), de entre los cuales destaca el sesgo de automatización, siendo uno de los riesgos que particularmente supone un desafío en el ámbito de la supervisión humana.

## 2. **El sesgo de automatización como riesgo de la supervisión humana**

El sesgo de automatización se define como el riesgo de que las personas confíen excesivamente en los resultados, propuestas e información que generan los sistemas que permiten la automatización de

(26) Giannini, A., Kwik, J., *op. cit.,* p. 58.

(27) Lazcoz Moratinos, G., *op. cit.*, pp. 26-27.

(28) Verdicchio, M., Perin, A., «When Doctors and AI Interact: on Human Responsibility for Artificial Risks», *Philosophy and Technology*, 2022, vol.35, núm. 11, p. 16.

(29) Grupo de expertos de alto nivel sobre inteligencia artificial de la Comisión Europea, *op. cit.*, 2019, p. 20.

(30) Skitka, L. J., Mosier, K. L., Burdick, M., «Does automation bias decision-making?», *International Journal of Human-Computer Studies*, vol. 51, núm. 5, 1999, p. 992.

(31) Giannini, A., Kwik, J., *op. cit.*, p. 58.

tareas y procesos (32). Esta tendencia de las personas a confiar en el funcionamiento de sistemas automatizados se produce incluso en contextos en los que existen otros indicadores o sistemas de ayuda no automatizados, pudiendo llegar las personas a tomar decisiones y actuar en contra de su propio criterio y experiencia (33).

El sesgo de automatización no es un problema exclusivo del contexto de uso de los sistemas de IA, sino que es uno más de los riesgos derivados de la automatización de procesos y viene siendo analizado desde hace décadas, tanto desde una perspectiva regulatoria (34), como en el ámbito de sectores y actividades en los que se ha producido una progresiva automatización de procesos (35) (como, por ejemplo, la aviación, el diagnóstico médico o el uso militar de sistemas de armamento autónomos). La cuestión es que la introducción de sistemas de IA en cada vez más contextos, ámbitos y actividades genera un incremento de los espacios en los que el sesgo de automatización puede tener impacto y relevancia. Sin embargo, son escasas las investigaciones en relación con las implicaciones penales específicas del sesgo de automatización (36), limitándose habitualmente su cita como uno de los factores de riesgo relacionados con el uso de sistemas de IA.

Las causas del sesgo de automatización son diversas: desde las limitaciones cognitivas de los seres humanos, que tendemos a optimizar esfuerzos en la toma de decisiones (*cognitive laziness*); el «acomodamiento social» (*social loafing) y* la «difusión de responsabilidad» *(diffusion of responsibility)* que, de forma similar a cuando las personas trabajan en equipo con otras, tienden a reducir su diligencia y a repartir la carga de la responsabilidad con el sistema automatizado; y, por último, la denominada «autoridad de los computadores» (*authority of computers)*, la tendencia de las personas a percibir los sistemas informáticos como entidades con mayores capacidades que las suyas propias (37).

Este conjunto de causas puede dar lugar, de manera individual o concurriendo de forma simultánea, a errores por parte del sujeto que

(32) Coco, A., «Exploring the Impact of Automation Bias and Complacency on Individual Criminal Responsibility for War Crimes», *Journal of International Criminal Justice*, vol. 21, núm. 5, 2023, p. 1078.

(33) Skitka, L. J., Mosier, K. L., Burdick, M., *op. cit.*, p. 992

(34) Ya en 1992 declaraba la Comisión Europea que existía una preocupación por que el abuso de la informática pudiese llevar a una dejación de responsabilidades por parte de quien implementase sistemas de tratamiento de datos automatizados. *Vid.* Lazcoz Moratinos, G., *op. cit.*, p. 27.

(35) Verdicchio, M., Perin, A., *op. cit.*, p. 23.

(36) Coco, A., *op. cit.*, pp. 1085 y ss.

(37) Skitka, L. J., Mosier, K. L., Burdick, M., *op. cit.*, p. 993.

utiliza o supervisa el sistema de IA tanto en forma de omisiones (es decir, la no realización de una conducta debida, por ejemplo, la interrupción del funcionamiento o la reversión del resultado generado por la IA), como de conductas activas (realizar una acción que debería haberse omitido o desarrollar el comportamiento de forma indebida) (38). Algunos estudios sugieren que las personas pueden llegar incluso a cometer más errores y omisiones en entornos automatizados, como consecuencia de la tendencia a ser menos vigilantes, que en aquellos en los que no existen ayudas o asistencia automatizada (39). Esto se produce independientemente de si las personas comprenden (o declaran comprender) las limitaciones del sistema automatizado (40).

En consecuencia, el debate en torno al sesgo de automatización no se desarrolla en términos de «riesgo cero», sino de «riesgo permitido» (41). La introducción de sistemas automatizados en procesos y actividades tiene como uno de sus objetivos la reducción de riesgos derivados de errores humanos, aunque estos sean sustituidos o hagan emerger otros riesgos distintos (como el sesgo de automatización), cuya existencia es socialmente aceptada con el fin de aprovechar las ventajas en eficiencia, ahorro en términos de costes y de tiempo, o la reducción de la probabilidad de que se produzcan accidentes (42).

## III. LA RESPONSABILIDAD PENAL DERIVADA DEL SESGO DE AUTOMATIZACIÓN: DEBER OBJETIVO DE CUIDADO Y RIESGO PERMITIDO EN LA SUPERVISIÓN HUMANA DE SISTEMAS DE IA

Desde una perspectiva jurídico-penal, la relevancia del sesgo de automatización se ubica en el ámbito del delito imprudente. En particular, la valoración de la adecuación de la conducta del supervisor

(38) Goddard, K., Roudsari, A., Wyatt, J. C., «Automation bias: a systematic review of frequency, effect mediators, and mitigators», *Journal of the American Medical Informatics Association*, *núm. 19,* 2012, p. 121.

(39) Skitka, L. J., Mosier, K. L., Burdick, M., *op. cit.*, p. 999.

(40) *Idem*.

(41) Salvadori, I., *op. cit.*, p. 164.

(42) Por ejemplo, en el contexto de los vehículos de conducción automatizada el objetivo es reducir o eliminar el factor del error humano en la conducción, causa principal o interviniente en el 90% de los accidentes en carretera. *Vid.* considerando (23) del Reglamento (UE) 2019/2144 del Parlamento Europeo y del Consejo, de 27 de noviembre de 2019, relativo a los requisitos de homologación de tipo de los vehículos de motor.

humano al deber objetivo de cuidado y cómo afecta a tal valoración el hecho de que el ejercicio de sus labores se haya visto afectado por el sesgo de automatización, así como la delimitación de los niveles de riesgo permitido en el desarrollo de las actividades de supervisión y uso de sistemas de IA.

El cuidado objetivamente debido se determina partiendo del hecho de que el resultado típico sea objetivamente previsible, quedando prohibidas las acciones de las que se abstendría una persona inteligente y prudente en la que se encontraba el autor (43). Debe señalarse que no resulta posible realizar de forma general una delimitación *ex ante* del contenido del deber objetivo de cuidado y de los umbrales del riesgo permitido en el uso de sistemas de IA, en la medida que los límites concretos vendrán determinados por la normativa aplicable y las circunstancias de cada caso de uso concreto (44), habiendo quedado este principio así plasmado en el enfoque basado en los riesgos que asume la normativa europea (considerando 26 RIA).

### 1. **La regulación extrapenal**

A la hora de determinar el contenido del deber objetivo de cuidado del supervisor de un sistema de IA, se deberá tener en cuenta el contenido de los requisitos y obligaciones que específicamente se establecen en la regulación para su desarrollo y despliegue (en concreto el RIA, tal como se ha analizado *supra* en el apartado 2). En la medida que una actividad esté regulada por el Derecho administrativo (como ocurre, por ejemplo, en el contexto del RIA) la infracción del deber objetivo de cuidado se puede comprobar a partir de la constatación de la infracción de esta normativa (45). Esto no debe interpretase como una relación de necesidad entre la infracción de la disposición administrativa y la existencia de una imprudencia jurídico-penal, así como el cumplimiento de la normativa administrativa tampoco excluye una imprudencia penalmente relevante (aunque, por lo general, la infrac-

(43) Romeo Casabona, C. M., «El tipo del delito de acción imprudente», en Romeo Casabona, C. M., Sola Reche, E., Boldova Pasamar, M. A., *Derecho Penal, Parte General*, 2.ª ed., Comares, Granada, 2016, p. 135.

(44) Romeo Casabona, C. M., *Conducta peligrosa e imprudencia en la sociedad de riesgo*, Comares, Granada, 2005, p. 16; PAREDES CASTAÑÓN, J. M., *op. cit.*, p. 341.

(45) Romeo Casabona, C. M., *Conducta peligrosa e imprudencia …*, *op. cit.*, p. 210.

ción de la normativa administrativa conllevará la infracción del deber de cuidado penal (46)).

En contraposición con la tradicional tendencia a que la regulación esté integrada por estándares legales abiertos que necesitan ser concretados en cada caso concreto por el juez (por ejemplo, en el ámbito societario, la diligencia de un «buen comerciante» o la del «administrador leal»), en la actualidad se tiende a sustituir estos estándares genéricos por reglas más precisas contenidas en leyes, reglamentos (47) e incluso mediante remisiones a normas de carácter privado (48). Este fenómeno resulta especialmente problemático en sectores como las nuevas tecnologías, en los que la producción formal del Derecho no es suficientemente ágil (siendo la IA un ejemplo más de este fenómeno) por lo que se recurre a fuentes informales, más cercanas a las bases de la adecuación social (49).

En el contexto de la IA esta situación se ha visto corregida (hasta cierto punto) con la reciente aprobación del RIA, que sirve como un elemento de concreción de los límites del riesgo permitido en el ámbito del desarrollo y despliegue de sistemas de IA. A pesar de que el escenario tras la aprobación de esta norma se presenta como uno dotado de mayor seguridad jurídica, al positivizarse un conjunto de obligaciones, requisitos y prohibiciones en el desarrollo e implementación de sistemas de IA, la existencia de normativa administrativa que regula un determinado ámbito o actividad no está exenta de ambigüedades, no siendo el RIA una excepción, habiendo sido ya criticada la vaguedad e indeterminación de su contenido y de los requisitos y obligaciones que establece (50).

Por otra parte, el ámbito de aplicación material del RIA es limitado y no abarca todas las posibles actividades y aplicaciones a las que se pueden destinar los sistemas de IA, quedando restringida la aplicación de la mayor parte de las obligaciones y requisitos previstas en

---

(46) Silva Sánchez, J. M., «Mandato de determinación e imprudencia», *Indret*, núm. 2, 2012, p. 2.

(47) Silva Sánchez, J. M., *El riesgo permitido en Derecho penal económico*, Atelier, Barcelona, 2022, p. 24.

(48) Contreras Caimovich, L., «Reglas extrajurídicas y creaciones de riesgos toleradas o desaprobadas en los delitos culposos de homicidio y lesiones», *Política criminal*, núm. 13, 2018, pp. 396-397; Pastor Muñoz, N., *Riesgo permitido y principio de legalidad: la remisión a los estándares sociales de conducta en la construcción de la norma jurídico-penal*, Atelier, Barcelona, pp. 61 y ss.

(49) Silva Sánchez, J. M., *El riesgo permitido…*, *op. cit.*, pp. 34-35.

(50) Hacker, P., «The European AI liability directives, Critique of a half-hearted approach and lessons for the future», *Computer law and security review*, núm. 51, 2023, p. 34.

esta norma a los sistemas de IA de alto riesgo, en particular las relativas a su supervisión humana. De esta forma, habrá sistemas de IA para los que será preceptiva la presencia de un supervisor humano, mientras que en otros la designación de este sujeto dependerá, bien de la existencia de una regulación sectorial que así lo establezca, bien de la decisión del proveedor de establecer en las instrucciones de uso del sistema la supervisión humana como una medida de seguridad, o bien por la decisión del responsable del despliegue del sistema de nombrar un supervisor humano. Esta decisión puede ser adoptada por iniciativa propia del operador o por su adhesión al correspondiente código de conducta (*ex* artículo 95 RIA, ver *supra* apartado 2). Esto no tiene que ser, en realidad, un problema, en la medida que las obligaciones y requisitos están establecidas por el RIA en función del riesgo que presenta el uso de sistemas de IA en ámbitos concretos, por lo que aquellos sistemas que no sean considerados como de alto riesgo y, por tanto, no les resulte de aplicación el Reglamento, no deberían presentar (al menos sobre el papel) riesgos significativos para la vida, la salud u otros derechos fundamentales.

Otra de las limitaciones que presenta el RIA es que los requisitos y obligaciones que se prevén en la norma relacionados con la supervisión humana de los sistemas de IA no están destinados a las personas que realizan tales tareas de supervisión, sino a los proveedores y los responsables del despliegue de los sistemas de IA. El Reglamento no aporta criterios que determinen o, al menos, orienten cómo se deben desempeñar las tareas de supervisión. Sin embargo, esto tampoco tiene por qué ser algo necesariamente negativo, en la medida que el contenido de estas obligaciones de supervisión podrá ser determinado acudiendo a otras fuentes, como la normativa sectorial que resulte de aplicación en el ámbito de uso o actividad concreto en el que se despliegue el sistema como, por ejemplo, en el ámbito de la asistencia médica (el Reglamento de Productos Sanitarios o la normativa administrativa de ordenación de la actividad sanitaria (51)) o en el contexto de la conducción de vehículos automatizados (52); las reglas técnicas o *leges artis* que correspondan en el ámbito de actividad concreto en

(51) Lazcoz Moratinos, G., *op. cit.*, pp. 23-24.

(52) Por ejemplo, en Francia, para la circulación con vehículos de conducción automatizada se exige la presencia de una persona física en «estado y condiciones de retomar la conducción del vehículo» cuando sea necesario o requerido: Miró Llinares, F., «General Report», *Revue Internationale de Droit Pénal,* vol. 95, iss. 1, 2024, p. 61; Giuca, M., «Disciplinare l'intelligenza artificiale. La riforma francese sulla responsabilità penale da uso di auto a guida autónoma», *Archivio Penale*, núm. 2, 2022, p. 23; Giannini, A., Kwik, J., *op. cit.*, p. 28.

el que se desarrolle la supervisión del sistema de IA (53), así como por las obligaciones que específicamente se prevean en el seno de la relación contractual con el responsable del despliegue, ya sea esta de naturaleza laboral o mercantil o, incluso, de carácter no profesional (54) (aunque los sistemas de IA en este último caso quedan fuera del ámbito de aplicación del Reglamento *ex* artículo 2.10 RIA). Además de los supuestos en los que, sin exigirlo la normativa, se designe a un supervisor por decisión del responsable del despliegue o del proveedor del sistema, puede haber casos en los que un sujeto realice la asunción voluntaria de estas tareas (por ejemplo, en el contexto de actividades no profesionales) de tal forma que se asuma una posición de garante sin necesidad de que exista formalmente un contrato o una delegación de responsabilidades o funciones (55).

Sin perjuicio de lo hasta aquí expuesto, a pesar de que el RIA no establezca obligaciones directamente aplicables a las personas que realicen las funciones de supervisión humana, sí que pueden utilizarse los requisitos que deben cumplir los proveedores de sistemas de IA de alto riesgo relativos al diseño de forma que permita su supervisión efectiva que, de forma indirecta, pueden servir como criterios para valorar la conducta del supervisor (56).

Asumiendo que el proveedor ha diseñado el sistema de IA de forma que se cumple lo dispuesto en el artículo 14.4 RIA, el supervisor humano deberá poder: a) ser consciente de las limitaciones del sistema; b) ser consciente del sesgo de automatización; c) interpretar correctamente los resultados generados; d) descartar, invalidar o revertir los resultados de salida generados, y e) intervenir o interrumpir el funcionamiento del sistema. Podemos distinguir aquí dos grupos de obligaciones: por un lado, las relativas a *comprender el funcionamiento del sistema*, tanto su operativa y sus resultados, como sus capacidades y limitaciones (a, b y c); y, por otro, las relativas a la *posibilidad de intervenir en el funcionamiento del sistema* (d y e).

Resultarían de especial relevancia, por ejemplo, en el contexto de un resultado lesivo causado porque el supervisor se ha visto afectado por el sesgo de automatización, el primer grupo de requisitos (a, b y c), tanto en relación con la posibilidad de «entender adecuadamente» las capacidades y limitaciones del sistema (de forma que el supervisor

(53) Silva Sánchez, J. M., «Mandato de determinación …», *op. cit.,* p. 2.

(54) Feijoo Sánchez, B. J., *Homicidio y lesiones imprudentes: requisitos y límites materiales*, Ediciones Olejnik, Santiago de Chile, 2021, p. 37.

(55) Jakobs, G., *Derecho Penal: Parte general*, 1997, Marcial Pons, Madrid, p. 987.

(56) Beck, J., Burri, T., *op. cit.*, p. 31.

pueda detectar anomalías y problemas de funcionamiento), como las de «ser consciente del sesgo de automatización» y «entender los resultados del sistema». Los problemas en este punto están vinculados con lo abiertos, ambiguos y poco precisos que son los términos utilizados por la normativa, como «entender adecuadamente», la «capacidad de entender» o la «capacidad de ser consciente» de las limitaciones, sesgos y riesgos derivados de su funcionamiento. Esta ambigüedad es consecuencia de que el RIA sea una norma de carácter general y transversal (57).

En cuanto al requisito de que el supervisor pueda «ser consciente del sesgo de automatización» (art. 14.4.b RIA), si en el caso concreto resultase probado que las medidas establecidas por el proveedor del sistema permitían al supervisor ser consciente de este riesgo, será un indicio relevante de que el supervisor tenía la posibilidad de conocer que su decisión estaba viéndose afectada por este sesgo. Si bien este indicio será relevante, no tiene que suponer, de por sí, una infracción del deber objetivo de cuidado, en la medida que la posibilidad de conocer, en abstracto, las limitaciones del sistema y de verse afectado por el sesgo de automatización, no implica en todo caso que, primero, fuese objetivamente previsible que se iba a producir el fallo o el resultado lesivo (58) [debiendo estos supuestos reconducirse a la figura del caso fortuito (59)]; y, segundo, que fuese posible evitar resultado (60) (analizaremos esta cuestión en mayor profundidad *infra* en el apartado 6, en relación con los comportamientos alternativos conforme a Derecho). Por ello, aunque el sistema esté adecuadamente diseñado para que se pueda ejercer una supervisión humana eficaz, de forma que haga que el sujeto sea consciente de la posibilidad de verse afectado por el sesgo de automatización, se deberá analizar el resto de las circunstancias que expliquen por qué, a pesar de tener la posibilidad de ser consciente de las limitaciones del sistema y de la tendencia a asumir o confiar indebidamente en los resultados que este genera (sesgo de automatización), no se pudo desarrollar la supervisión de forma efectiva.

---

(57) Enqvist, L., «Human oversight' in the EU artificial intelligence act: what, when and by whom?», *Law, Innovation and Technology*, vol. 15, iss.2, 2023, p. 525.

(58) Valls Prieto, J., *op. cit.,* p. 27; Salvadori, I., *op. cit.,* pp. 143-144.

(59) De forma general: Paredes Castañón, J. M., *op. cit.*, p. 515; en relación con los sistemas de IA y los robots específicamente: Quintero Olivares, G., «La robótica ante el derecho penal: el vacío de respuesta jurídica a las desviaciones incontrolada», *Revista Electrónica de Estudios Penales y de la Seguridad*, núm. 1, 2017, pp. 16-17.

(60) Salvadori, I., *op. cit.,* pp. 157-158.

La otra categoría de sujetos para los que se establecen obligaciones específicas relativas a la supervisión humana de sistemas de IA es la del «responsable del despliegue», esto es, las persona física o jurídica que utilice el sistema de IA bajo su propia autoridad (art. 3.4 RIA). La única obligación atinente a estos sujetos que puede tener incidencia en el ámbito de la supervisión humana del sistema de IA es la relativa al deber de nombrar a una persona con la competencia, autoridad y formación necesarias para desempeñar las labores de supervisión humana (art. 26.2 RIA). Los incumplimientos que, por tanto, se podrían producir en relación con este requisito, se darían en supuestos en los que la persona designada como supervisor no tuviese la formación, competencia o autoridad necesarias. No obstante, incluso en el caso en el que se demostrase la falta de uno de estos elementos (formación, competencia o autoridad), ello no determinaría automáticamente la ausencia de responsabilidad por parte del supervisor humano, en la medida que, si éste fuese consciente (o debiese ser consciente) de sus limitaciones, estaríamos ante una imprudencia por asunción, en la que la acción cuidadosa impone la abstención de actuar (61). En resumidas cuentas, una posible culpa *in eligendo* del responsable del despliegue podría suponer un supuesto de concurrencia de culpas, pero no implicaría la compensación total de una culpa con la otra en todo caso (62).

## 2. La *lex artis* en el contexto de la IA

En defecto de regulación (o junto con la misma), se deberá tener en cuenta la existencia de estándares de conducta o normas profesionales (*lex artis)* en el concreto ámbito en el que se despliegue un sistema de IA a la hora de evaluar si la conducta del supervisor se ajusta al cuidado debido (63). Estos estándares técnicos no determinan lo que está o no permitido por la sociedad, «sino la mejor manera de

(61) Feijoo Sánchez, B. J., *Homicidio y lesiones imprudentes..., op. cit.*, p. 140.

(62) *Ibidem*, p. 173. Por lo demás, véase Sánchez Lázaro, F. G., «Argumentos consecuencialistas, concurrencia de imprudencias y evaluación de soluciones técnico jurídicas», *Estudios Penales y Criminológicos*, vol. XXX, 2010, pp. 447 y ss.; El Mismo: «Strafbarkeit nicht voll zurechenbarer Rechtsgutsverletzungen nach Versuchsgrundsätzen im Fahrlässigkeitsbereich», *Goltdammer´s Archiv für Strafrecht*, vol. 152, 2005, pp. 700 y ss.

(63) Romeo Casabona, C. M., *Conducta peligrosa e imprudencia...*, op. cit., p. 209.

enfrentarse a un riesgo o controlarlo», sin que implique la tolerancia por el ordenamiento jurídico de dicho riesgo (64).

Los sujetos que hagan uso de sistemas de IA estarán sometidos a los estándares profesionales y técnicos existentes en su respectivo sector del mismo modo que si utilizasen cualquier otro dispositivo o artilugio. De esta forma, si el supervisor de un sistema de IA, por ejemplo, en el ámbito sanitario, se ve afectado por el sesgo de automatización y realiza un diagnóstico basado en una aplicación rutinaria de los resultados que genere el sistema o priorizando el criterio de la IA por encima del suyo propio, se podrán considerar tales conductas como comportamientos contrarios a la *lex artis* y, en estos términos, una infracción del deber objetivo de cuidado, en la medida que no se podría justificar tales desviaciones de la *lex artis* en la libertad de método ni considerarse como el comportamiento que desarrollaría un profesional prudente (65).

### 3. **Sesgo de automatización y principio de confianza**

El principio de confianza tiene relevancia a la hora de delimitar los límites del riesgo permitido en el contexto de supervisión de los sistemas de IA debido a dos circunstancias que caracterizan el ecosistema de desarrollo y utilización de esta tecnología: 1) la existencia de procedimientos de certificación y evaluaciones de conformidad; y 2) el reparto de roles y tareas en el contexto del trabajo en equipo.

Respecto de la certificación y evaluación de los sistemas de IA, puede relacionarse con la vertiente del principio de confianza vinculada con la que podríamos denominar *confianza institucional*. El supervisor que utiliza un sistema de IA que ha sido certificado, conforme a lo dispuesto en el RIA (u otra norma al efecto), puede asumir un margen de confianza legítimo en que el sistema funcionará de manera adecuada y segura en relación con su finalidad prevista. En el contexto del RIA, esta certificación se materializa en las evaluaciones de conformidad que demuestren el cumplimiento de los requisitos establecidos en el propio Reglamento (art. 43 RIA), o bien con lo dispuesto en aquellas normas que establezcan la obligatoriedad de la realización de una evaluación de conformidad (Anexo I, en relación con el artículo 6.1 RIA), teniendo en cualquier caso especial relevancia al ser tales evaluaciones preceptivas para que el sistema de IA pueda ser

(64) Feijoo Sánchez, B. J., *Homicidio y lesiones imprudentes…*, *op. cit.*, p. 147.

(65) Romeo Casabona, C. M., *Conducta peligrosa e imprudencia…*, *op. cit.*, pp. 210-211.

introducido en el mercado (66). No obstante, el hecho de que el sistema esté certificado no exime automáticamente de responsabilidad al supervisor, el cual no deberá desempeñar sus funciones asumiendo una confianza ciega en el funcionamiento del sistema por el hecho de estar certificado o haber obtenido una evaluación de conformidad favorable (67) (lo que, además, podría calificarse de por sí como un comportamiento afectado por el sesgo de automatización).

El principio de confianza también se utiliza en el contexto del riesgo permitido para describir las situaciones de trabajo en equipo, reparto de roles y la división horizontal y vertical del trabajo. Este principio autoriza a confiar en que otros sujetos con los que se interactúa, o con los que se desarrolla una actividad en común, van a realizar la conducta que se espera de ellos de forma correcta (68). En el contexto de la supervisión de sistemas de IA se ha planteado la posibilidad de asemejar la interacción humano-máquina, *mutatis mutandis*, como una relación de este tipo, en la que el ser humano confiaría en que el sistema de IA va a funcionar adecuadamente, del mismo modo que confiaría en una persona con la que trabaja en equipo o con la que se reparte una tarea (69). No resulta esta idea descabellada, en la medida que existen estudios que asemejan a nivel psicológico la conducta de los sujetos que comparten procesos con sistemas que automatizan tareas con aquellas en las que se desarrolla un trabajo en equipo con otras personas (70).

El problema de esta idea reside en que, al menos tal y como se concibe por la dogmática el principio de confianza en relación con el

(66) La evaluación externa de la fiabilidad de los sistemas de IA venía siendo demandada por la doctrina: *Vid.* Perin, A., *Prudenza, dovere di conoscenza e colpa penale. Proposta per un método di giduzio,* Editoriale Scientifica, Nápoles, 2020, p. 407; Romeo Casabona, C. M., «La discusión sobre la atribución de responsabilidad penal a sistemas de inteligencia artificial, en particular a sistemas autónomos», en Romeo Casabona, C. M., Rueda Martín, M. Á. (eds.), *Derecho Penal, ciberseguridad, ciberdelitos e inteligencia artificial, volumen II: Inteligencia artificial y responsabilidad penal*, Granada, Comares, 2023, p. 64; Valls, J., *op. cit.*, pp. 23-24. No obstante, se debe tener en cuenta que no todos los procedimientos de evaluación implican la intervención de una entidad externa en el contexto del RIA, existiendo la posibilidad de que la evaluación de conformidad se realice basados en controles internos desarrollados por el propio proveedor del sistema de IA (arts. 43.1 y.2 RIA).

(67) Íñigo Corroza, M. E., *La responsabilidad penal del fabricante por defectos de sus productos*, J. M. Bosch Editor, Barcelona, 2001, p. 220.

(68) Jakobs, G., *op. cit.*, p. 254.

(69) Mongillo, V., «Responsabilidad penal y extrapenal de las personas jurídicas por delitos relacionados con la inteligencia artificial: vías de intervención legal y principales obstáculos», *Revista Electrónica de Responsabilidad Penal de Personas Jurídicas y Compliance*, núm. 4, 2024, p. 10; Verdicchio, M., Perin, *op. cit*., p. 16.

(70) Skitka, L. J., Mosier, K. L., Burdick, M., *op. cit.*, p. 993.

reparto de roles, se enfrenta a dos obstáculos: la relevancia de las obligaciones de actuación frente a riesgos y la posibilidad de «confiar» en un sistema de IA. Respecto del primero, la existencia de una norma que establezca la obligación de actuar y evitar resultados lesivos a un sujeto imposibilita que este pueda escudarse en el principio de confianza para justificar su actuación (71), al menos respecto de aquellos concretos supuestos en los que esté obligado a actuar. En cuanto al segundo, incluso en aquellos casos en los que no exista tal regla u obligación (por ejemplo, cuando no sea obligatorio designar a un supervisor humano para el sistema de IA), el sujeto en el que se confía deberá actuar de acuerdo con su deber objetivo de cuidado, resultando debatible que se pueda considerar a un sistema de IA como participante en una actividad del que se puedan tener tales expectativas (72). El fundamento del principio de confianza se encuentra, precisamente, en que cada sujeto es responsable de actuar en su ámbito de organización, por lo que este principio decae cuando el sujeto en el que se pretende confiar carece del conocimiento o la capacidad de conocer las reglas (73), lo que podemos afirmar que ocurre con un sistema de IA, por muy autónomo que sea (74).

Por último, en relación con el específico riesgo del sesgo de automatización, se entiende que para que un sujeto pueda acogerse al principio de confianza deberá estar actuando de conformidad con el deber objetivo de cuidado, siendo debatible, como se ha venido señalando, que una aplicación rutinaria de los resultados, decisiones o recomendaciones generados por el sistema de IA, o sin realizar las comprobaciones necesarias para contrastarlos, se pueda considerar como una actuación diligente y cuidadosa (75).

---

(71) Jakobs, G., *op. cit.,* p. 255.

(72) Feijoo Sánchez, B. J., *Homicidio y lesiones imprudentes…*, *op. cit.*, p. 159.

(73) Jakobs, G., *op. cit.*, pp. 254-255.

(74) Favorable a considerar que resulta posible confiar en el funcionamiento del sistema de IA en la medida que «su empleo se justifica precisamente por el objetivo de favorecer mayores niveles de seguridad […] tratando de disminuir la tasa de errores humanos evitables»: Perin, A., «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, núm. 1, 2019, p. 18.

(75) Romeo Casabona, C. M., «El tipo del delito…», *op. cit.*, p. 141. En contra Feijoo Sánchez, que considera esta limitación inaceptable, ya que abre la puerta a que el sujeto que se comporta imprudentemente responda por las imprudencias de terceros. *Vid.* Feijoo Sánchez, *Homicidio y lesiones imprudentes…*, *op. cit.*, pp. 157-158.

## IV. SESGO DE AUTOMATIZACIÓN Y REPRESENTACIÓN DEL RESULTADO LESIVO

En caso de que se pueda considerar que la conducta del supervisor del sistema de IA ha vulnerado el deber objetivo de cuidado, surge la duda de cuáles serían los criterios para diferenciar cuándo la conducta del sujeto afectado por el sesgo de automatización se puede calificar como una imprudencia consciente o inconsciente (76). De forma general, se considera que nos encontramos ante una imprudencia consciente cuando el sujeto ha previsto la posibilidad de producción del resultado y ha actuado en la confianza de que no se produciría, mientras que la imprudencia será inconsciente si, aun siendo el resultado objetivamente previsible, el sujeto no ha previsto tal posibilidad (77).

En principio, los casos en los que el supervisor de un sistema de IA se ha visto afectado por el sesgo de automatización encajarían como supuestos de imprudencia inconsciente, en los que este sujeto aplica de forma rutinaria y sin revisar la información de salida generada por el sistema sin ser consciente (sin representarse la posibilidad) de que el sistema pueda cometer errores ni de que pueda estar aceptando o utilizando decisiones erróneas. Sin embargo, en aquellos supuestos en los que el sistema de IA sea considerado de alto riesgo, si cumple con los requisitos establecidos en el artículo 14 del RIA, el supervisor debería poder ser consciente de «la posible tendencia a confiar automáticamente o en exceso en los resultados de salida generados por un sistema de IA de alto riesgo». En tanto en cuanto el sistema esté diseñado de forma que el supervisor humano pueda ser consciente del riesgo de verse afectado por el sesgo de automatización, si tales medidas son realmente capaces de hacer al supervisor consciente (o habría sido razonable que el supervisor fuese consciente) de este sesgo, será más difícil argumentar que la persona que supervisa no podía prever la posibilidad de que se produjese un error del sistema y que, por lo tanto, en realidad conocía tal riesgo y actuó en la confianza de que no ocurriese, lo que encajaría en la descripción

(76) Coco, A., *op. cit.*, p. 1082.

(77) Romeo Casabona, C. M., «el tipo del delito…», *op. cit.*, p. 138; Mir Puig, S., *Derecho Penal, Parte General*, Reppertor, Barcelona, 2006, p.285; Sánchez-Vera Gómez-Trelles, J., «Nuevas tendencias normativistas en el concepto y la prueba del dolo», *Derecho Penal y Criminología*, vol. 26, núm. 79, 2005, p. 104.

de la culpa consciente (78), salvo que concurran circunstancias que hayan impedido o dificultado que tuviese tal conocimiento (79).

Por otra parte, en el caso de que el resultado lesivo se haya producido como consecuencia de la infracción del deber de conocer unas circunstancias fácticas determinadas, la conducta del sujeto será considerada como una imprudencia consciente, en la modalidad de imprudencia por asunción. Por lo general, una conducta respetuosa con el deber de cuidado pasará por abstenerse de actuar en determinadas actividades peligrosas sin disponer de los conocimientos, capacidades o la experiencia necesarios, debiendo quien actúa ser consciente de sus cualidades y del límite de sus posibilidades para desarrollarlas. Esto ocurre, por ejemplo, cuando un sujeto no utiliza instrumental adecuado o cuando, teniendo conocimientos generales suficientes, no se informa sobre las circunstancias del caso concreto, no pudiendo considerarse dentro del riesgo permitido una actividad que, si se hubiese llevado a cabo estando dotado el sujeto de capacidades suficientes, sí lo sería (80). Trasladando estas reflexiones al ámbito de la supervisión de sistemas de IA, el uso del sistema de forma rutinaria o sin realizar las comprobaciones necesarias para confirmar que los resultados que genera son correctos puede ser interpretado en estos términos, es decir, como el desarrollo de una actividad sin las capacidades o conocimientos suficientes para el caso concreto, estando las primeras mermadas o limitadas por el sesgo de automatización, y como consecuencia de la falta de recopilación de la información necesaria, en los segundos.

En aquellos casos en los que se pueda considerar que el sujeto se había representado el riesgo asociado a su actuación (o estaba obligado a tener previsto tal riesgo), se abre la puerta a valorar cuándo la conducta del supervisor afectado por el sesgo de automatización se puede considerar que rebasa el umbral de la imprudencia consciente para ubicarse en el terreno del dolo eventual. Hay autores que consideran que la diferencia entre dolo e imprudencia depende del grado en el que se produce la representación de la probabilidad del resultado y las circunstancias que determinan el hecho (81). Otros consideran que en la imprudencia consciente el autor conoce el peligro abstracto de una acti-

(78) Quintero Olivares, G., *op. cit.*, p. 19.

(79) Ragués I Vallès, R., *La ignorancia deliberada en Derecho Penal*, Atelier, Barcelona, 2007, p. 157.

(80) Feijoo Sánchez, B. J., *Homicidio y lesiones imprudentes…*, *op. cit.*, pp. 140-141.

(81) Struensse, E., «Consideraciones sobre el dolo eventual», *InDret*, 2009, pp. 11-12.

vidad, pero no la dimensión del riesgo que se está creando en el caso concreto (82), es decir, que desarrolla la actividad con el convencimiento de que no va a materializarse el resultado, concurriendo dolo si existe un pronóstico concreto de lo que puede ocurrir en el caso concreto (83). De esta forma, de acuerdo con estos autores, únicamente debería tenerse en cuenta el elemento cognitivo a la hora de diferenciar entre dolo e imprudencia, sin tener en cuenta el elemento volitivo (los criterios de actuar «tomando en serio» o «contando con el resultado») (84), siendo lo relevante que el sujeto conozca (o no) el riesgo (85).

En relación con aquellos contextos en los que el sujeto está sometido a concretas obligaciones y deberes de actuación, la decisión contraria a éstos con base en lo que se conoce es lo que fundamenta la responsabilidad por dolo, es decir, la infracción de un deber directo o inmediato de evitación del hecho típico, mientras que la imprudencia conlleva la infracción de un deber indirecto, la no activación de una prestación personal de atención o cuidado (86). Sin embargo, no deben incluirse aquí las reglas e indicaciones del estándar de cuidado en determinados ámbitos, como las normas de tráfico o de fabricación de productos, por lo que será, igualmente, la infracción del deber objetivo de cuidado la que determine la concurrencia del dolo en estos casos (87).

Siguiendo esta idea, los resultados lesivos ocasionados como consecuencia del sesgo de automatización (es decir, aquellos en los que ha sido determinante a la hora de adoptar u omitir indebidamente una acción por el supervisor) son, por lo general, conductas imprudentes, en la medida que no suponen una actuación que infrinja un deber directo de evitación del hecho típico, sino una infracción de un deber de atención o cuidado. Esto implica que, en la práctica, serán principalmente aquellos supuestos en los que la conducta del sujeto afectado por el sesgo de automatización se pueda calificar como ignorancia deliberada (88) en los que se podría debatir la imputación a título de dolo eventual (89).

---

(82) Feijoo Sánchez, B. J., *Dolo eventual*, Ediciones Olejnik, Santiago de Chile, 2018, pp. 10-11.

(83) Mir Puig, S., *op. cit.*, pp. 265-266.

(84) Feijoo Sánchez, B. J., *Dolo eventual*, *op. cit.,* pp. 41-43.

(85) *Ibidem*, p. 29*;* Sánchez-Vera Gómez-Trelles, J., *op. cit.*, p. 105. Por lo demás, al respecto: *Vid.* Ragués I Vallès, R., *op. cit.,* pp. 118-120.

(86) Feijoo Sánchez, B. J., *Dolo eventual*, *op. cit.,* p. 12.

(87) *Ibidem*, p. 21.

(88) Coco, A., *op. cit.*, pp. 1088-1089; Ragués I Vallès, R., *op. cit.,* pp. 188-193.

(89) Se muestra la doctrina mayoritaria en contra de la posibilidad de equiparar ignorancia deliberada y dolo eventual, aunque sí es aceptada, bajo ciertas condicio-

## V. SESGOS DE AUTOMATIZACIÓN Y LOS COMPORTAMIENTOS ALTERNATIVOS CONFORME A DERECHO

Otra óptica desde la que se pueden analizar los supuestos de hecho relacionados con el sesgo de automatización es el del conocido como problema del «comportamiento alternativo conforme a derecho», con el que se hace referencia a casos de conductas imprudentes que han sido la causa de un resultado típico, pero en los que, aunque se hubiese ejecutado la acción correcta, el resultado lesivo se habría producido igualmente (90).

Conforme a la doctrina mayoritaria, estos supuestos se resuelven acudiendo a la teoría de la evitabilidad, según la cual el resultado no se puede imputar objetivamente a su causante si éste se hubiera producido incluso en el caso de que el autor hubiese realizado la conducta debida (91). Se afirma, pues, que en estos casos faltaría la «conexión de la infracción del deber» o la «realización de la lesión de cuidado en la producción del resultado» (92). Si el daño se hubiese producido también en el hipotético caso de que se hubiera ejercido un control efectivo sobre el sistema, no sería posible imputar tal resultado al supervisor humano, que solamente será responsable cuando tenga capacidad de prever y evitar razonablemente los daños causados, algo que en este contexto no siempre es posible, en la medida que las decisiones y acciones del sistema de IA pueden estar más allá del control inmediato del supervisor (93). Si no es posible determinar si el resultado se hubiese evitado con el desarrollo debido de las actuaciones del supervisor humano, solamente se podrán imputar responsabilidades a

---

nes, por la jurisprudencia: *Vid.* Mir Puig, S., Corcoy Bidasolo, M. C., «Artículo 10», en Mir Puig, S., Corcoy Bidasolo, M. C., (Dirs.), *Comentarios al Código Penal, Reformas LLOO 1/2023, 3/2023 y 4/2023,* 2.ª Edición, Tirant lo Blanch, Valencia, 2024, p. 74.

(90) Gimbernat Ordeig, E., «Teoría de la evitabilidad versus teoría del aumento del riesgo», *Anuario de Derecho Penal y Ciencias Penales,* vol.48, 2015, p. 25.

(91) La postura dominante considera que la probabilidad de que el resultado se hubiese evitado debe ser «rayana en la seguridad», habiendo otros autores que consideran que basta con la mera posibilidad. *Vid.* Bolea Bardón, C., «La responsabilidad penal del médico por omisión desde una teoría normativa de la evitabilidad», *InDret,* núm. 4, 2018, p. 14.

(92) Gimbernat Ordeig, E., *op. cit.*, p. 26.

(93) Cámara Arroyo, S., «¿Sueñan los androides delincuentes con penas electrónicas? Sistemas de armamento autónomo y responsabilidad penal: nuevos desafíos para el Derecho penal», en García Sánchez, B., Jiménez García, F. (coords.), *La atribución de una responsabilidad jurídico penal e internacional de la Inteligencia Artificial*, Iustel, Madrid, 2023, p. 271.

través de un delito de peligro abstracto, una opción que, hoy en día, no existe en relación con el uso de sistemas de IA (94).

Gimbernat critica el resultado que se alcanza siguiendo la teoría de la evitabilidad, en la medida que supone vaciar de contenido el concepto de riesgo permitido (95). Defiende, junto con un sector minoritario de la doctrina, que para resolver estos casos se debe acudir al criterio del «incremento del riesgo», rechazando que la probabilidad o posibilidad de que el resultado también se hubiese producido aun realizando la conducta adecuada sea el criterio relevante para determinar la imputación del resultado. De esta forma, si es probable, pero no seguro, que la acción correcta hubiese causado el mismo resultado, significa que el comportamiento imprudente causante del resultado era más peligroso, lo que implicaría que tal conducta supone un incremento del riesgo en el que se puede fundamentar la imputación de responsabilidad (96).

De acuerdo con la teoría del aumento del riesgo, solo quedarían exentos de responsabilidad los casos en los que la conducta alternativa conforme a derecho habría conducido «con seguridad» a la producción del mismo resultado, de forma que la conducta imprudente no supusiese un mayor peligro que el riesgo permitido (97). En cambio, si no existe tal seguridad, sino únicamente la probabilidad de que no se hubiese producido el mismo resultado, solamente podrá fundamentarse una absolución si el sujeto demostrase que ha tomado las medidas necesarias para reducir el riesgo (98).

Por lo tanto, si aplicamos este criterio en el contexto de la supervisión de sistemas de IA, el resultado lesivo se podría atribuir al supervisor si su desempeño supusiese un incremento del riesgo de que se produzca el resultado. El desafío se presenta a la hora de determinar cuál debe ser la referencia a la hora de comparar y valorar el incremento del riesgo asociado a los distintos niveles de control humano (99). Es decir, si habiendo desarrollado las labores de supervisión de una manera diferente el riesgo habría sido menor. Por ejemplo, si el sistema advierte que existe una situación de incertidumbre y el supervisor decide continuar sin realizar otro tipo de comprobaciones o consultar otras fuentes de información distintas, o si realiza una aplicación rutinaria del sistema (i.e., afectado por el sesgo de automa-

(94) Gil Gil, A., «Sistemas de armas autónomos letales. Comentario a la posición española, con especial atención a algunos problemas jurídico-penales», *Revista General de Derecho Penal,* núm. 38, 2022, p. 43.

(95) Gimbernat Ordeig, E., «Teoría de la evitabilidad…», *op. cit.*, pp. 36-37.

(96) *Ibidem*, p. 29.

(97) *Ibidem*, p. 30.

(98) *Idem*.

(99) Cámara Arroyo, S., *op. cit.*, p. 273.

tización), estas situaciones se deberán comparar con aquellas en las que sí se efectuasen tales comprobaciones o en la que se practicase un uso no rutinario de los resultados del sistema.

Lo problemático en estos casos será demostrar que la conducta alternativa efectivamente supondría una disminución del riesgo (100), especialmente, en aquellos contextos en los que no existan referencias o guías para el desempeño de las tareas de supervisión, en los que resulta más difícil valorar si la actuación del sujeto supone un incremento del riesgo, al menos en comparación con aquellos ámbitos en los que exista una regulación mediante la que se establezcan deberes o estándares de comportamiento más concretos. La cuestión es que en el ámbito de la IA ni siquiera la reciente aprobación del RIA soluciona este problema, en la medida que, como se ha indicado a lo largo del trabajo, esta norma no establece deberes para los supervisores ni los usuarios finales de los sistemas de IA.

En resumen, para valorar la posibilidad de imputar responsabilidad al supervisor en los casos analizados se deberá valorar si, en el caso de haberse realizado la conducta alternativa conforme a Derecho, se habría evitado (o es posible que se hubiese evitado) el resultado, en caso de que asumamos la teoría de la evitabilidad, o bien si la conducta imprudente ha supuesto un incremento del riesgo superior al riesgo permitido (en caso de asumir la teoría del incremento del riesgo), con las dificultades probatorias aparejadas a ambos criterios, dificultades que, no obstante, no deben condicionar las definiciones de los conceptos fundamentales de la imputación penal (101).

## VI. POSICIÓN DE GARANTE Y DELITOS DE INFRACCIÓN DE DEBER

La falta de control que supone la imprevisibilidad en el funcionamiento de los sistemas de IA más avanzados (esto es, con mayores capacidades de autonomía, aprendizaje y capacidad de adaptación) dificulta considerar que el sujeto tiene realmente el dominio del hecho, lo que supone un desafío para la imputación objetiva de un resultado lesivo basándose en este criterio (102). Sin embargo, en ocasiones el dominio del hecho por parte del sujeto no es relevante a la hora de fundamentar su responsabilidad, sino la infracción de un deber por

(100) Gil Gil, A., «Sistemas de armas autónomos…», *op. cit.*, p. 43.
(101) Schünemann, B., «El propio sistema del Derecho Penal», *InDret*, 2008, p. 7.
(102) Cámara Arroyo, S., *op. cit.*, p. 270.

parte del agente (103), como ocurre en los delitos de omisión impropia de un deber de garante (104), pudiendo ubicar aquí los supuestos en los que el supervisor de un sistema de IA se ve afectado por el sesgo de automatización. Con todo, no basta en esta clase de delitos la mera infracción del deber por parte del agente para legitimar la sanción penal, sino que deberá concurrir el injusto material que se fundamenta en la lesión o puesta en peligro de un bien jurídico protegido (105).

Si bien no resulta posible imputar responsabilidad a un sujeto por riesgos imprevisibles, también es cierto que cuando concurren especiales deberes jurídicos, este sujeto puede quedar obligado a vigilar e informarse de las eventuales consecuencias perjudiciales que se puedan producir en el desarrollo de la actividad, como ocurre, por ejemplo, en la responsabilidad del fabricante por productos defectuosos (106). Este paradigma se ha visto intensificado como consecuencia de la creciente adscripción de deberes de prevención a través de la normativa extrapenal a los focos de riesgo (por ejemplo, empresas multinacionales), definiéndose el reproche penal a partir de la infracción de los deberes especiales que vinculan al sujeto con la protección de bienes jurídicos (107).

En el contexto de la supervisión de sistemas de IA, la posición de garante vendría determinada por la obligación del responsable del despliegue de designar a un sujeto que supervise su funcionamiento en caso de que se trate de un sistema de alto riesgo *ex* artículo 26.2 RIA. Cabe también la posibilidad de que se establezca la obligatoriedad de este rol por otra regulación sectorial específica para sistemas de IA (por ejemplo, la figura del conductor en los vehículos de conducción automatizada (108)) o para actividades en las que un sujeto tenga que cumplir unos deberes de vigilancia y cuidado específicos (por ejemplo, en el ámbito sanitario).

Si bien el RIA no establece directamente los deberes del supervisor humano (sin perjuicio de que, como se ha indicado ya, sí que puedan hacerlo otras normas), el contenido de las obligaciones que conforman la posición de garante del supervisor humano estará determinado contrac-

(103) Pariona Arana, R., «La teoría de los delitos de infracción de deber y el principio de legalidad», *Revista Electrónica de Ciencia Penal y Criminología*, núm. 26, 2024, p. 6.

(104) Jakobs, G., *op. cit.*, p. 267.

(105) Pariona Arana, R., *op. cit.*, p. 7.

(106) Contreras Caimovich, L., «La responsabilidad penal del fabricante por la infracción de sus deberes de vigilancia, advertencia y retirada», *Política criminal*, Vol. 10, Núm. 19, 2015, p. 279.

(107) Pariona Arana, R., *op. cit.*, p. 26.

(108) Ver *supra* nota al pie n.º 52.

tualmente, en el contexto de la relación laboral o estatutaria en la que se desarrolle el uso del sistema de IA. Para aquellos sistemas de IA que, por no ser clasificados como de alto riesgo, no les resulten de aplicación las obligaciones previstas en el RIA y, por tanto, no exista una obligación legal de nombrar un supervisor humano del sistema de IA, igualmente se podrá conformar una posición de garante derivada de la existencia de una relación contractual, cuya existencia no es necesario interpretar en términos formales (es decir, un contrato en términos de derecho privado, tanto si es de naturaleza mercantil, laboral o incluso civil), sino de la efectiva asunción de una función de control o evitación de riesgos (109). Igualmente, también puede fundamentarse una posición de garante por injerencia, como consecuencia de la creación de un riesgo derivada de la puesta en funcionamiento del sistema de IA (*ex* artículo 11.b CP).

El sujeto que asume las labores de supervisión humana del sistema de IA vinculadas a esta posición de garante deberá hacerlo de manera real y efectiva, es decir, ejerciendo su dominio sobre los riesgos. De lo contrario, será el sujeto que crea el riesgo (el proveedor o el responsable del despliegue del sistema de IA), a quien se le imputará, en su caso, la responsabilidad cuando se materialice la lesión, al no haberse realizado la delegación de tales riesgos de forma efectiva (110).

En los supuestos de posición de garante por injerencia, la responsabilidad del sujeto se fundamenta en una actuar precedente que presenta un riesgo (el uso de un sistema de IA) pero que, en la medida que se cumplan los requisitos establecidos por la normativa (el RIA), serán considerados como un riesgo permitido, siempre y cuando se adopten las medidas de precaución necesarias para «contener el riesgo» (111). En el contexto de la supervisión de sistemas de IA estas medidas consistirían en realizar las comprobaciones y actuaciones necesarias para que el sujeto no se vea afectado por el sesgo de automatización, lo que se puede traducir en el desarrollo de un «control humano significativo» o, al menos, una aplicación no rutinaria de la información de salida generada por el sistema de IA. Sin embargo, viene siendo aceptado que la valoración de la idoneidad de las medidas de precaución no genera problemas en relación con las estructuras de imputación de la responsabilidad, sino que se trata de una cuestión de prueba, por lo que los argumentos para resolver estos casos deberán ir vinculados a la presunción de inocencia y al juego entre pruebas de cargo y de descargo (112).

---

(109) Feijoo Sánchez, B. J., *Homicidio y lesiones imprudentes…*, *op. cit.*, p. 37.

(110) *Ibidem*, p. 38.

(111) Dopico Gómez-Aller, J., *Omisión e injerencia en Derecho Penal,* Tirant lo Blanch, Valencia, 2006, p. 541.

(112) *Ibidem*, p. 551.

## VII. CONCLUSIONES

1. A pesar de que el nuevo marco regulatorio aplicable a la IA en Europa aporta seguridad jurídica y distribuye obligaciones y responsabilidades entre los sujetos que participan en la cadena de valor del desarrollo e implementación de sistemas de IA, su impacto se ve limitado en relación con aquellos sujetos que no quedan bajo su ámbito de aplicación material, como los supervisores y usuarios finales de los sistemas de IA, así como por su enfoque restringido a unas actividades, ámbitos y usos concretos.

2. Si bien la supervisión humana de los sistemas de IA se presenta como una de las principales medidas de control de riesgos asociados al uso de sistemas de IA, no debe utilizarse como excusa para convertir a los sujetos encargados de las labores de supervisión en chivos expiatorios a los que imputar responsabilidades indebidamente. A pesar de que se trate de los sujetos que se encuentran más próximos al daño, no siempre tendrán posibilidad de ejercer un control efectivo sobre el funcionamiento del sistema, así como de prever los resultados lesivos derivados de su uso.

3. El sesgo de automatización es un riesgo inherente al uso de sistemas y herramientas que permiten automatizar tareas y procesos, siendo la contrapartida a los beneficios esperados en términos de eficiencia y reducción de riesgos. Como fenómeno psicológico, resulta complicado valorar el sesgo de automatización como una conducta imprudente, así como la imputación de responsabilidades al confluir los deberes de distintos sujetos en torno al control de las causas y los efectos de este sesgo (proveedores, responsables del despliegue, supervisores, etc.).

4. Sobre el papel, se podría considerar como una conducta cuidadosa la adopción de las medidas adecuadas por parte del supervisor para evitar verse afectado por el sesgo de automatización. Sin embargo, resulta difícil determinar cuáles son los umbrales de riesgo permitido que permitan valorar una conducta como descuidada, en la medida que los supervisores de sistemas de IA no tienen establecidas obligaciones o criterios respecto al desarrollo de tal actividad en el RIA, presentándose otras fuentes de la determinación del riesgo permitido, como la *lex artis* o el principio de confianza, como limitadas en este contexto, ante la aún incipiente implantación de estos sistemas en la realidad social.

5. Los supuestos en los que se produzcan resultados lesivos como consecuencia del sesgo de automatización en el ejercicio de labores de supervisión serán, en principio, considerados como supuestos de imprudencia inconsciente, en la medida que, por definición, el

sujeto afectado no se representa el riesgo de que se produzca el resultado lesivo. Sin embargo, si la falta de representación por parte del sujeto se produce como consecuencia de la infracción de deberes de vigilancia y evitación de resultados, se puede plantear la posibilidad de considerarlos como supuestos de imprudencia consciente, siendo clave a estos efectos la forma en la que se interpreten el contenido y los límites de las obligaciones del supervisor del sistema de IA, en particular a la hora de plantear la posibilidad de imputar responsabilidad a título de dolo eventual.

6. De acuerdo con la doctrina relativa a los comportamientos alternativos conforme a derecho, no podrá imputarse responsabilidad al supervisor cuando el resultado lesivo se hubiese producido aunque hubiese ejercitado un control humano significativo del sistema de IA (o bien si la ausencia de tal control no ha supuesto un incremento del riesgo, si nos acogemos a este otro criterio, doctrinalmente minoritario). En cualquier caso, la prueba tanto de la posibilidad de evitar el resultado, como del incremento del riesgo, respecto de la realización de la conducta debida, se presenta como previsiblemente difícil al tratase los sistemas de IA, por lo general, de elementos técnicamente complejos.

7. Los sistemas de IA que presentan mayores niveles de autonomía tienden a presentar mayores grados de imprevisibilidad en su funcionamiento, lo que genera dificultades en relación con la imputación objetiva de los resultados lesivos derivados de su uso como consecuencia de la falta del dominio del hecho por parte del supervisor. No obstante, para los supuestos en los que existan deberes especiales de actuación, como los derivados de la posición de garante del supervisor de un sistema de IA, la doctrina de los delitos de infracción de deber aporta soluciones que permiten solventar algunos de los problemas de imputación derivados de la falta de dominio del hecho en estos contextos.

## VIII. BIBLIOGRAFÍA

Beck, J., Burri, T., «From «Human Control» in International Law to «Human Oversight» in the New EU Act on Artificial Intelligence», *Forthcoming* en Amoroso, D., Santoni de Sio, F. (eds.*), Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*, Elgar, 2023. Disponible en: https://ssrn.com/abstract=4236554 (recuperado el 15 de septiembre de 2024).

Bolea Bardón, C., «La responsabilidad penal del médico por omisión desde una teoría normativa de la evitabilidad», *InDret,* núm. 4, 2018.

Cámara Arroyo, S., «¿Sueñan los androides delincuentes con penas electrónicas? Sistemas de armamento autónomo y responsabilidad penal: nuevos desafíos para el Derecho penal», en García Sánchez, B., Jiménez García, F. (coords.), *La atribución de una responsabilidad jurídico penal e internacional de la Inteligencia Artificial*, Iustel, Madrid, 2023, pp. 225-286.

Coco, A., «Exploring the Impact of Automation Bias and Complacency on Individual Criminal Responsibility for War Crimes», *Journal of International Criminal Justice*, vol. 21, núm. 5, 2023, pp. 1077-1096.

Contreras Caimovich, L., «Reglas extrajurídicas y creaciones de riesgos toleradas o desaprobadas en los delitos culposos de homicidio y lesiones*, Política criminal*, vol. 13, núm. 25, 2018*,* pp. 387-443.

— «La responsabilidad penal del fabricante por la infracción de sus deberes de vigilancia, advertencia y retirada», *Política criminal*, Vol. 10, núm. 19, 2015, pp. 266-296.

Dopico Gómez-Aller, J., *Omisión e injerencia en Derecho Penal,* Tirant lo Blanch, Valencia, 2006.

Enqvist, L., «Human oversight' in the EU artificial intelligence act: what, when and by whom?», *Law, Innovation and* Technology, vol. 15, iss.2, 2023, pp. 508-535.

Feijoo Sánchez, B. J., *Homicidio y lesiones imprudentes: requisitos y límites materiales*, Ediciones Olejnik, Santiago de Chile, 2021.

— *Dolo eventual*, Ediciones Olejnik, Santiago de Chile, 2018.

Giannini, A., Kwik, J., «Negligence failures and negligence fixes. A comparative analysis of criminal regulation of AI and autonomous vehicles», *Criminal Law Forum*, núm. 34, 2023, p. 43-85.

Gil Gil, A., «Sistemas de armas autónomos letales. Comentario a la posición española, con especial atención a algunos problemas jurídico-penales», *Revista General de Derecho Penal,* núm. 38, 2022.

Gimbernat Ordeig, E., «Teoría de la evitabilidad versus teoría del aumento del riesgo», *Anuario de Derecho Penal y Ciencias Penales,* vol.48, 2015, pp. 21-62.

Giuca, M., «Disciplinare l'intelligenza artificiale. La riforma francese sulla responsabilità penale da uso di auto a guida autónoma», *Archivio Penale*, núm. 2, 2022.

Goddard, K., Roudsari, A., Wyatt, J. C., «Automation bias: a systematic review of frequency, effect mediators, and mitigators», *Journal of the American Medical Informatics Association*, núm. 19, 2012, pp. 121-127.

Grupo de Expertos de Alto Nivel Sobre Inteligencia Artificial de La Comisión Europea, *Directrices éticas para una IA fiable*, 2019.

Hacker, P., «The European AI liability directives, Critique of a half-hearted approach and lessons for the future», *Computer law and security review*, núm. 51, 2023.

Harbers, M., Peeters, M. M. M., Neerincx, M. A., «Perceived Autonomy of Robots: Effects of Appearance and Context», en Aldinhas Ferreira, M., Silva Sequeira, J., Tokhi, M., E. Kadar, E., Virk, G. (eds.) *A*

*World with Robots. Intelligent Systems, Control and Automation: Science and Engineering*, Springer, Cham, 2017, pp.19-33.

Jakobs, G., *Derecho Penal: Parte General, Fundamentos y teoría de la imputación,* 2.ª edición, Marcial Pons, Madrid, 1997.

Kautz, H. A., «The third AI summer: AAAI Robert S. Engelmore memorial lecture», *AI Magazine*, vol. 43, núm. 1, 2022, pp. 105-125.

Lazcoz Moratinos, G., *Sistemas de inteligencia artificial en la asistencia sanitaria: cómo garantizar la supervisión humana desde la normativa de protección de datos*, AEPD, Madrid, 2022.

Mir Puig, S., Corcoy Bidasolo, M. C., «Artículo 10», en Mir Puig, S., Corcoy Bidasolo, M. C., (Dirs.), *Comentarios al Código Penal, Reformas LLOO 1/2023, 3/2023 y 4/2023,* 2.ª Edición, Tirant lo Blanch, Valencia, 2024, pp. 71-76.

Mir Puig, S., *Derecho Penal: Parte general*, 8.ª edición, Reppertor, Barcelona, 2006.

Miró Llinares, F., «General Report», *Revue Internationale de Droit Pénal,* vol. 95, iss. 1, 2024, pp. 19-87.

Mongillo, V., «Responsabilidad penal y extrapenal de las personas jurídicas por delitos relacionados con la inteligencia artificial: vías de intervención legal y principales obstáculos», *Revista Electrónica de Responsabilidad Penal de Personas Jurídicas y Compliance*, núm. 4, 2024.

Palma Herrera, J. M., «Inteligencia artificial y neurociencia. Algunas reflexiones sobre las aportaciones que pueden hacer al Derecho Penal», Peris Riera, J. M., Massaro, A. (coords.), *Derecho penal, inteligencia artificial y neurociencias*, Roma Tre-Press, Roma, 2023, pp. 249-270.

Paredes Castañón, J. M., *El riesgo permitido en Derecho Penal*, Ministerio de Justicia e Interior, Madrid, 1995.

Pariona Arana, R., «La teoría de los delitos de infracción de deber y el principio de legalidad», *Revista Electrónica de Ciencia Penal y Criminología*, núm. 26, 2024.

Pastor Muñoz, N., *Riesgo permitido y principio de legalidad: la remisión a los estándares sociales de conducta en la construcción de la norma jurídico-penal*, Atelier, Barcelona, 2019.

Perin, A., *Prudenza, dovere di conoscenza e colpa penale. Proposta per un método di giduzio,* Editoriale Scientifica, Nápoles, 2020.

— «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, núm. 1, 2019, pp. 3-28.

Picotti, L., «Traditional Criminal law categories and AI: crisis or palingenesis? General report», *Revue Internationale de Droit Pénal*, Vol. 94, issue 1, 2023, pp. 11-53.

Quintero Olivares, G., «La robótica ante el derecho penal: el vacío de respuesta jurídica a las desviaciones incontrolada», *Revista Electrónica de Estudios Penales y de la Seguridad*, núm. 1, 2017.

Ragués I Vallès, R., *La ignorancia deliberada en Derecho Penal*, Atelier, Barcelona, 2007.

ROMEO CASABONA, C. M., «La discusión sobre la atribución de responsabilidad penal a sistemas de inteligencia artificial, en particular a sistemas autónomos», en ROMEO CASABONA, C. M., RUEDA MARTÍN, M. Á. (eds.), *Derecho Penal, ciberseguridad, ciberdelitos e inteligencia artificial, volumen II: Inteligencia artificial y responsabilidad penal*, Granada, Comares, 2023, pp. 57-82.

ROMEO CASABONA, C. M., «El tipo del delito de acción imprudente», en ROMEO CASABONA, C. M., SOLA RECHE, E., BOLDOVA PASAMAR, M. A., *Derecho Penal, Parte General*, 2.ª ed., Comares, Granada, 2016, pp. 123-137.

ROMEO CASABONA, C. M., *Conducta peligrosa e imprudencia en la sociedad de riesgo*, Comares, Granada, 2005.

SALVADORI, I., «Agentes artificiales, opacidad tecnológica y distribución de la responsabilidad penal», *Cuadernos de política criminal*, núm. 133, mayo 2021, pp. 137-174.

SÁNCHEZ LÁZARO, F. G.: «Argumentos consecuencialistas, concurrencia de imprudencias y evaluación de soluciones técnico jurídicas», Estudios Penales y Criminológicos, vol. XXX, 2010, pp. 447 y ss.

— «Strafbarkeit nicht voll zurechenbarer Rechtsgutsverletzungen nach Versuchsgrundsätzen im Fahrlässigkeitsbereich», *Goltdammer´s Archiv für Strafrecht*, vol. 152, 2005, pp. 700 y ss.

SÁNCHEZ-VERA GÓMEZ-TRELLES, J., «Nuevas tendencias normativistas en el concepto y la prueba del dolo», *Derecho Penal y Criminología*, vol. 26, núm. 79, 2005, p. 99-112.

SILVA SÁNCHEZ, J. M., *El riesgo permitido en Derecho penal económico*, Atelier, Barcelona, 2022.

— «Mandato de determinación e imprudencia», *InDret*, núm. 2, 2012.

— *La expansión del Derecho penal: Aspectos de la Política criminal en las sociedades postindustriales*, 2.ª edición, Editorial Bdef, Montevideo-Buenos Aires, 2006.

SKITKA, L. J., MOSIER, K. L., BURDICK, M., «Does automation bias decision-making?», *International Journal of Human-Computer Studies*, vol. 51, núm. 5, 1999, pp. 991–1006.

STRUENSEE, E., «Consideraciones sobre el dolo eventual», *InDret*, núm. 4, 2009.

TRIGUERO, I., MOLINA, D., POYATOS, J., DEL SER, J., HERRERA, F., «General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, societal implications and responsible governance», *Information Fusion*, vol. 103, 2024, pp. 6-9.

VALLS PRIETO, J., «Sobre la responsabilidad penal por la utilización de sistemas inteligentes», *Revista Electrónica de Ciencia Penal y Criminología*, núm. 24-27, 2022.

VERDICCHIO, M., PERIN A., «When Doctors and AI Interact: on Human Responsibility for Artificial Risks», *Philosophy and Technology*, vol.35, núm. 11, 2022.