Universidad Rey Juan Carlos | Facultad de Ciencias Jurídicas y Políticas

TRABAJO FIN DE GRADO

GRADO EN DERECHO

CURSO ACADÉMICO 2023-2024

CONVOCATORIA MARZO

ALGORITMOS DE NECESIDAD EN VEHÍCULOS AUTÓNOMOS: CONFIGURACIÓN ÉTICA Y SUPUESTOS DERIVADOS DE RESPONSABILIDAD PENAL

AUTOR: Balaguer Medrano, David

En Madrid, a (día) de (mes) de 2024

ÍNDICE

RESUMEN 3
INTRODUCCIÓN A LA PROBLEMÁTICA 4
I. IA y coches autopilotados 4
II. El algoritmo de necesidad 5
III. Atribución de responsabilidad 6
TEORÍA DEL ALGORITMO DE NECESIDAD 7
I. Metodología: el dilema del tranvía o gestión del riesgo 7
II. Un algoritmo para todos o un algoritmo personalizable 9
1. Configuración ética personalizable 10
2. Configuración ética obligatoria 13
III. Un algoritmo utilitarista estricto 14
IV. El modelo deontológico 17
1. Un plan preliminar a largo plazo 17
2. Cuestiones previas 18
2.1. Los principios de autonomía y solidaridad 18
2.2. La ponderación de intereses 20
3. Reacción a agresiones ilegítimas 21
4. Criterios del estado de necesidad agresivo 22
4.1. Espacios seguros 24
5. Criterios del estado de necesidad defensivo 25
6. Las capas de seguridad del VA: cuándo interviene el algoritmo de necesidad 28
LA IMPUTACIÓN DE RESPONSABILIDAD PENAL DERIVADA DEL ALGORITMO DE NECESIDAD 31
I. La atribución de responsabilidad en el panorama de los VA 31
II. La IA como un sujeto de Derecho responsable penalmente 32
III. El riesgo permitido 33
1. Los VA y sus algoritmos de necesidad como un riesgo permitido 33
2. La cláusula de cierre: seguros obligatorios para los VA 34
IV. Responsabilidad penal de los usuarios, programadores o fabricantes de VA 35
1. El usuario del VA como responsable penal 36
2. Los fabricantes y programadores de VA como responsables penales 37
CONCLUSIONES 39
BIBLIOGRAFÍA 42

RESUMEN

Los vehículos autónomos proyectan una mayor seguridad en carretera en el futuro próximo, pero será inevitable que se enfrenten ante las mismas situaciones de necesidad que los conductores humanos. Deberán optar en estas ocasiones entre los bienes jurídicos en peligro conforme a unas pautas adecuadas tanto a nuestro ordenamiento como a la realidad: un algoritmo de necesidad. Este algoritmo minimizaría el riesgo en comparación con los beneficios a la seguridad que aportarían estos coches autopilotados, cubriéndose así su producción, venta y uso bajo la figura del riesgo permitido. A su vez, permitirá ordenar la responsabilidad penal de los distintos intervinientes en el proceso de creación y uso de vehículos autónomos. En su diseño deberán participar expertos en Derecho, IA, comercialización y ética y usuarios, pero este trabajo pretende orientar en cuanto a las diversas posiciones de la doctrina con respecto a la programación del algoritmo de necesidad y la responsabilidad penal derivada de este.

Palabras clave: algoritmo de necesidad, vehículos autónomos, responsabilidad penal, configuración ética, riesgo permitido.

ABSTRACT

Automated vehicles present an opportunity for more secure roads in our near future, but it is inevitable that these cars will face the same situations of necessity that human drivers experience. On these occasions, it must decide between various legal interests in danger based on guidelines that are adequate to our legal system and to reality: an algorithm of necessity. This algorithm would minimize the risks compared to the benefits to security that self-driving cars are supposed to bring to the roads, so their production, selling and usage can be covered by the legal figure of permissible risk. Moreso, it would determine the criminal liability of the various actors who create and use these automated vehicles. In order to design the algorithm, legal, AI, merchandising and ethics experts and users must participate. This essay pretends to offer orientation about the diverse positions the doctrine offers on the matter of programming this algorithm of necessity and the criminal liability derived from it.

Palabras clave: algorithm of necessity, automated vehicles, criminal liability, ethical configuration, permissible risk.

INTRODUCCIÓN A LA PROBLEMÁTICA

I. IA y coches autopilotados

La Inteligencia Artificial (IA) se presenta como una gran oportunidad y una gran amenaza al mismo tiempo en la docencia, la industria, la domótica y otras muchas áreas de la sociedad. Al mismo tiempo que ofrece nuevas oportunidades de optimización de sistemas, rapidez de respuesta y amplia capacidad de procesamiento de información, nos encontramos con brechas en la seguridad de nuestros datos, campos de actuación nuevos que requieren una protección jurídica adaptada y nuevos riesgos que no se preveían para las TIC hasta el momento. Nosotros nos centraremos en el sector de la seguridad vial y la producción automovilística (cada vez más desarrollada), donde se ensalza la IA como la solución a los accidentes de tráfico y la mortalidad en carreteras. La Organización Mundial de la Salud recoge en sus datos desde 2010 una media de 1,19 millones de fallecidos en accidentes de tráfico al año[1], de los cuales un 90% se deben a un error humano (despistes, microsueños, infracción de las normas viarias…)[2]. Frente a estas estadísticas, los vehículos autónomos (VA desde este momento) prometen la solución a una conducción segura, relegando la actuación humana *in situ* a un segundo plano y, junto con ella, sus errores; a pesar de que algunos accidentes mortales causados por coches autopilotados hayan plantado una semilla de duda[3]. Sus ventajas, sin embargo, parecen superar con creces estos obstáculos susceptibles de mejora con el avance tecnológico: reducir la contaminación (con una conducción más estable), mejorar la productividad y el descanso (al no tener que dedicar la atención a conducir), mejorar la movilidad y reducir drásticamente el número de accidentes, hacer el transporte más accesible (al no tener que saber conducir)…[4]

Las clasificaciones actuales sobre Inteligencias Artificiales (IA en adelante) implantadas en los vehículos varían ligeramente en el número de niveles de automatización (entre cinco y seis niveles, variando sobre todo a la hora de combinar o distinguir los dos últimos), pero coinciden en que aún no se ha logrado una automatización completa[5]. El estándar internacional J3016 para consumidores de VA tipifica seis niveles, de los cuales existe hasta el penúltimo (nivel 4) solo como prototipo[6]: un VA en el que el conductor apenas tiene que intervenir, aunque puede hacerlo con cierto margen de maniobra. Dada la proximidad geográfica y jurídica, tomaremos como referencia el Código Ético alemán, con seis niveles del 0 al 5 de automatización[7]: desde los VA sin automatización ninguna, pasando por la asistencia en la conducción, una automatización parcial que necesita un conductor, una automatización

[1] ORGANIZACIÓN MUNDIAL DE LA SALUD, *Global status report on road safety 2023*, 2023, Génova, p. 4.
[2] FEILER, J., "The Artificially Intelligent Trolley Problem: Understanding Our Criminal Law Gaps in a Robot Driven World", en *Hastings Science and Technology Law Journal*, Vol. 14, Nº. 1, 2023, p. 16.
[3] NYHOLM, S., "The ethics of crashes with self-driving cars: A roadmap, I", en *Philosophy Compass*, Vol. 13, Nº 7, 2018, p. 1.
[4] LAW COMMISSION & SCOTTISH LAW COMMISSION, "Automated Vehicles: Analysis of Responses to the Preliminary Consultation Paper", en *The National Archives: Open Government Licence*, Reino Unido, 2019, p. 30.
[5] GOODALL, N.J., "Ethical decision making during automated vehicle crashes", en *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2424, Nº. 1, 2014, p. 2.
[6] CONSEJO DE EUROPA, Legal aspects of autonomous vehicles, Comisión de Asuntos Legales y Derechos Humanos, Asamblea Parlamentaria, 2020, AS/Jur 20, p. 5.
[7] SUÁREZ, M.F., "Inteligencia Artificial y Derecho Penal: El Dilema del Tranvía. Cuarta Revolución Industrial. Ética del Algoritmo. IA en vehículos. Causas de Justificación", en *Revista Pensamiento Penal*, Nº. 445, 2022, pp. 12-14.

condicionada que requiere la actuación del conductor en emergencias y una automatización elevada que apenas necesita de intervenciones puntuales hasta la automatización completa. Nos centraremos en la discusión ética y jurídica que plantean los VA plenamente automatizados.

II. El algoritmo de necesidad

Ahora bien, el tema a tratar no es la conducción autónoma al completo, sino las situaciones de necesidad que estos VA encontrarán en la carretera, normalmente presentadas en la doctrina a través del 'dilema del tranvía', en el cual se asumen decisiones éticas ante el riesgo para la vida de los pasajeros y los transeúntes y en que la intervención humana es imposible o inadecuada[8]; debiendo preverse la solución desde el llamado algoritmo de necesidad para determinar qué bien jurídico debe prevalecer y cómo debe actuar el algoritmo del VA en consecuencia[9]. En palabras de Coca Vila, se trata de algoritmos que resuelvan *situaciones de necesidad en que la infracción de una norma o la causación o no evitación de un daño resultan inevitables*[10].

Esta es la razón por la que nos centraremos en el último nivel de automatización completa (el VA se encarga de la totalidad de las maniobras de pilotaje). Si bien aún no se encuentran en circulación, este trabajo pretende alcanzar resoluciones anticipadas a los dilemas que plantea la conducción autónoma en que la IA sea la encargada de llevar a cabo una decisión de conducción en una situación de necesidad que resulte en la lesión de un bien jurídico y analizar los algoritmos que se introducirán en estos VA y tomarán estas decisiones.

Se suscitan entonces una serie de preguntas derivadas. La primera es si el algoritmo debería ser definido y estandarizado de forma nacional o incluso universal conforme a unos principios comunes o si, por el contrario, debería dejarse en manos de cada consumidor la elección del criterio moral a implementar por el algoritmo de necesidad de su VA[11]. Aun en caso de que el algoritmo sea personalizable (ya que no se aceptaría cualquier tipo de criterio, sino solo una gama de aquellos acogidos socialmente por un sector suficientemente relevante y razonados), y más en la pretensión de algoritmos nacionales o universales, se presenta la cuestión de cuál será su fundamento ético[12]. Junto a esto, debe discutirse la formulación de sus principios y reglas adaptados a la problemática de las situaciones de necesidad en los VA, cuál será el método adecuado para alcanzar una deliberación propicia al consenso y si es una opción ética factible desde el lenguaje de las IA[13].

Avanzados estos puntos en nuestro trabajo, nuestra siguiente tarea será delimitar la posible analogía o extracción de principios desde las causas de justificación de nuestro ordenamiento (la legítima defensa, el estado de necesidad agresivo y el estado de necesidad defensivo recientemente incorporado por parte de la doctrina en España y, sobre todo, en

[8] NYHOLM, *op. cit.*, p. 2.
[9] GRANDI, N.M., "Inteligencia artificial al volante. Una mirada sobre la atribución de Responsabilidad Penal por los resultados lesivos generados por los vehículos autónomos", en *Revista Argentina de Derecho Penal y Procesal Penal*, Nº. 27, 2020, p. 17.
[10] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 238.
[11] CONSEJO DE EUROPA, *op. cit.*, p. 1.
[12] SUÁREZ, *op. cit.*, p. 8.
[13] D'AMATO, A., DANCEL, S., PILUTTI, J., TELLIS, L., FRASCAROLI, E. & GERDES, J.C., "Exceptional Driving Principles for Autonomous Vehicles", en *Journal of Law and Mobility*, Nº 2, 2022, p. 9.

Alemania y otras corrientes de Europa[14]) y cómo deben alinearse y ponderarse los bienes jurídicos en juego, los requisitos de justificación y los principios que fundamentan su consonancia con el ordenamiento jurídico español en situaciones diferentes conforme a estas tres causas de justificación. Así pues, algunos autores diferenciarán casos en que la vida o integridad de los pasajeros del VA estén en juego o no en la ponderación de intereses. También distinguiremos supuestos en que un sujeto sea causante responsable del peligro que creó el estado de necesidad y reciba, a su vez, la agresión[15] de aquellos accidentes en que la situación de peligro y necesidad provengan en una relación no responsable pero vinculada a través de un nexo causal (y, algunos autores exigen, de imputación objetiva) a un sujeto interviniente que sea el destinatario de la subsiguiente agresión[16] u otros en los que el sujeto que se encuentre en la situación de necesidad dirija la agresión hacia un tercero ajeno[17] al cuadro en que se enmarcan estas responsabilidades. Cada una de estas opciones hace prevalecer unos principios y establece unos límites en nuestro ordenamiento jurídico.

III. Atribución de responsabilidad

Por último, el problema manifiesta más preguntas respecto a la responsabilidad que dependen de las respuestas ofrecidas a las anteriores cuestiones. En definitiva: ¿quién sería responsable por los daños a la vida, la integridad o el patrimonio producidos en estas situaciones de necesidad: el usuario, el programador, el fabricante, otra persona o nadie en particular[18]? Estos interrogantes se distribuirán por las distintas ramas de doctrina que nos iremos encontrando al respecto, ya que no es lo mismo argumentar que los algoritmos de necesidad bajo un mismo estándar nacional cubren los riesgos como riesgo permitido sin haber por tanto tipicidad, asumir que el usuario deberá ser responsable de las consecuencias de un accidente causado por la opción ética que haya elegido en su VA o que la responsabilidad deba definirse bajo ciertas normas deontológicas que pretendan adaptar el modelo de responsabilidad penal habitual al nuevo ambiente en que los VA sean el instrumento ejecutor de la acción.

Todas estas preguntas y más componen nuestro trabajo, pero lo que es evidente es la necesidad de planteárselas con antelación para crear una base doctrinal suficientemente fuerte que sea útil al legislador que en el futuro (más o menos próximo) deba regular estas cuestiones cuando los VA plenamente automatizados alcancen las carreteras.

[14] COCA VILA, I., "Entre la responsabilidad y la solidaridad. El estado de necesidad defensivo", en *Revista para el Análisis del Derecho InDret*, Nº. 1, 2011, p. 2.
[15] MARTÍNEZ ESCAMILLA, M., MARTÍN LORENZO, Mª. & MARISCAL DE GANTE, M.V., *Derecho penal: Introducción. Teoría jurídica del delito. Materiales para su docencia y aprendizaje*, Universidad Complutense de Madrid, 2012, p. 270.
[16] COCA VILA, I., "Entre la responsabilidad y la solidaridad. El estado de necesidad defensivo", en *Revista para el Análisis del Derecho InDret*, Nº. 1, 2011, p. 4.
[17] WILENMANN VON BERNATH, J., "El sistema de derechos de necesidad y defensa en el Derecho penal", en *Revista para el Análisis del Derecho InDret*, Nº. 3, 2014, p. 6.
[18] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 238.

TEORÍA DEL ALGORITMO DE NECESIDAD

I. Metodología: el dilema del tranvía o gestión del riesgo

Para responder a las preguntas relacionadas con el algoritmo, se han formulado esencialmente dos métodos: el dilema del tranvía y la gestión del riesgo. A pesar de que la doctrina se divida entre apoyar uno u otro sistema, ambos tienen una utilidad en fases distintas del planteamiento. El dilema del tranvía cumple su función en el primer plano hipotético para hallar argumentos filosófico-morales y jurídicos. La gestión del riesgo nació para acercar estos argumentos a la realidad en situaciones complejas que no se basan en certezas, sino en probabilidades[19].

El tradicional dilema del tranvía se traduce en nuestro caso concreto a un VA al que se le presenta una situación en que debe decidir entre varias opciones que podrían causar la muerte de los pasajeros del VA, de los pasajeros de otros vehículos y de peatones[20]. La formulación más típica es que el VA se encuentra con un grupo de personas en la carretera, siendo la única posibilidad de evitarlos virar hacia la acera y chocar contra un peatón, causándole la muerte[21]. Las diferentes variantes añaden o modifican elementos para afinar más la respuesta ética (y jurídica) a estos dilemas, como puede ser la provocación por parte de alguno de los actores (los peatones entraron al paso de cebra saltándose un semáforo o el vehículo conducía a una velocidad cuestionable)[22], el número de muertes causadas (sea que ambas opciones causen igual número de muertes o sean asimétricas) o los bienes jurídicos en juego (cambiando el resultado de muerte por uno de lesión o por uno de daños a la propiedad)[23].

Varios autores han destacado ya los inconvenientes de utilizar el dilema del tranvía. Primero, debido a la limitada capacidad técnica de los VA para identificar datos como la intención de los actores[24] y la imposibilidad de conocer a ciencia cierta el resultado de cada acción, como se presume en el célebre planteamiento[25]. Como añadido a este punto se cuestionan qué tipo de datos deberían incorporarse al VA como relevantes para tomar una decisión sobre el riesgo que supone cada vía de actuación sin suponer a su vez una discriminación sobre aquellos sujetos (por ejemplo: según las estadísticas, las personas mayores tienen más probabilidad de morir en un accidente de tráfico, ¿debe considerarse como un factor para aumentar el riesgo de lesión de estas personas por encima de los ciudadanos jóvenes o de edad media?)[26]. Incluso plantearía una cuestión ética en cuanto a la protección de los datos personales que se le prestan al VA para decidir[27]. Por último, crea un falso dilema, ya que en la

[19] GOODALL, N.J., "Machine Ethics and Automated Vehicles", en Meyer, G. & Beiker, S. (ed.): *Road Vehicle Automation*, Springer, 2014, p. 96.
[20] GRANDI, *op. cit.*, p. 17.
[21] HOLSTEIN, T. & DODIG-CRNKOVIC, G., "Avoiding the Intrinsic Unfairness of the Trolley Problem", presentada en *2018 IEEE/ACM International Workshop on Software Fairnes (FairWare)*, Sweden, 2018, p. 1.
[22] LAWLOR, R., "The Ethics of Automated Vehicles: Why Self-driving Cars Should not Swerve in Dilemma Cases", en *Res Publica*, Vol. 28, Nº. 1, 2022, p. 198.
[23] Ídem.
[24] Íbidem, 4.
[25] NYHOLM, *op. cit.*, p. 4.
[26] GOODALL, N.J., "Ethical decision making during automated vehicle crashes", en *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2424, Nº. 1, 2014, p. 8.
[27] HOLSTEIN & DODIG-CRNKOVIC, *op. cit.*, p. 4.

realidad las opciones no se limitan a dos o tres cauces de actuación, sino a muchos más[28], si bien lo normal será que el algoritmo determine que solo unos pocos son los más probables para tener en cuenta y más adecuados según la ética con que sean configurados.

No obstante, pese a su simplificación y alejamiento de la realidad, autores como Goodall defienden que el dilema sirve como punto de partida para extraer unos estándares éticos basados en argumentos razonables[29], ya que la perspectiva de una tasa nula de accidentes de tráfico es prácticamente imposible a causa de los posibles errores de funcionamiento de los VA y agentes externos imprevisibles dentro y fuera del vehículo (ciclistas, fauna, peatones…)[30]. Así, se hace imprescindible construir una base legal y ética firme que equilibre la comprensión y confianza ciudadana y el progreso tecnológico, empresarial y de seguridad en este ámbito[31]. Esto se consigue a través de herramientas como el dilema del tranvía, que imagina casos similares, pero con ciertas variaciones que generan intuiciones morales distintas para transformarlas después en argumentaciones elaboradas sobre por qué deberían tratarse de forma diferente en base a principios que reflejan esas intuiciones (o se retractan de las mismas)[32].

Por otra parte, el propio Goodall propone la gestión del riesgo como un modelo más realista, pues asume que las situaciones de necesidad en carreteras suponen un cierto grado de incertidumbre y deben discutirse, no en términos de certeza, sino bajo el concepto de riesgo[33], que engloba tanto la probabilidad de que algo ocurra como la gravedad de las consecuencias que traiga el suceso en caso de suceder. Con la multiplicación de estas dos variables se computa el valor esperado del riesgo[34], que permite entonces optar por la opción cuyo valor esperado de riesgo sea menor. Se consigue así un sistema transparente y resolutivo para situaciones de necesidad.

Pondremos como ejemplo muy simplificado el caso anteriormente expuesto del dilema del tranvía: un VA se encuentra de frente con un grupo de tres individuos que cruzan la carretera, debiendo continuar su trayectoria o virar para esquivarlos a una acera donde pasea un peatón. El VA calcula la probabilidad de chocar con cada individuo del grupo de peatones que cruza la carretera en caso de no virar y continuar su trayectoria (incluso frenando), siendo esta de 0,8; y la multiplicaría por la probabilidad de muerte o lesión de ese individuo: 0,6[35]. Finalmente, haría esta operación con todos los individuos afectados por ese cauce de actuación (tres individuos en nuestra ilustración) para hallar su resultado global de riesgo:

0,8 (probabilidad de choque) x 0,6 (probabilidad de muerte) x 3 individuos = 1,44.

[28] GOODALL, N.J., "Away from Trolley Problems and Toward Risk Management", en *Applied Artificial Intelligence*, Vol. 30, Nº. 8, 2016, p. 812.
[29] Íbidem, 819.
[30] GOODALL, N.J., "Machine Ethics and Automated Vehicles", en Meyer, G. & Beiker, S. (ed.): *Road Vehicle Automation*, Springer, 2014, pp. 94-95.
[31] BONNEFON, J.F., SHARIFF, A. & RAHWAN, I., "The social dilemma of autonomous vehicles", en *Science*, Vol. 352, Nº. 6293, 2016, p. 2.
[32] LAWLOR, *op. cit.*, p. 212.
[33] Ídem, 7.
[34] GOODALL, N.J., "Away from Trolley Problems and Toward Risk Management", en *Applied Artificial Intelligence*, Vol. 30, Nº. 8, 2016, p. 814.
[35] Íbidem. Goodall utiliza como magnitudes números enteros como el número de vidas, asumiendo que la muerte de estos actores es segura, pero formularemos nuestro ejemplo en términos probabilísticos para mayor claridad y una aplicación más ajustada de este modelo.

Calcularía en el momento la probabilidad de chocar con el peatón en caso de virar hacia la acera (0,7) y la multiplicaría por la probabilidad de muerte o lesiones (0,8), hallando el valor esperado del riesgo de esa alternativa:

0,7 (probabilidad de choque) x 0,8 (probabilidad de muerte) = 0,56.

Y así con todas las opciones que el VA pudiera llevar a cabo, decidiendo al fin cuál tiene un menor valor esperado del riesgo. En nuestro caso, virar hacia el peatón tiene menor valor esperado de riesgo, por lo que, aplicando solo el modelo de gestión del riesgo, debería ser la alternativa a elegir.

Estas son las dos propuestas que lideran la problemática, pero, de combinarse ambas, la metodología comprendería un ámbito mucho más amplio y ajustado a la realidad. El cálculo de la probabilidad del riesgo del sistema de gestión del riesgo no cuenta con factores ético-jurídicos como pueden ser la provocación de la situación de peligro, la protección de los espacios seguros en carretera[36] u otros, que se encuentran en el ámbito de los valores y la ética más que en el área de la probabilidad. En nuestro ejemplo, virar hacia el peatón de la acera es la opción elegida, pero no considera si la acera es un espacio protegido que los coches no deben traspasar o si el grupo en la carretera provocó la situación de peligro al no estar cruzando un paso de peatones, cuestiones esenciales para una decisión de justicia. Estas materias y valores son precisamente los que el dilema del tranvía explora de forma exhaustiva en sus enunciaciones más complejas y actualizadas y permitirían perfeccionar un modelo lo más universal posible en su formulación ética abstracta[37]. Armonizándolo con el sistema de gestión del riesgo, aproximaría el modelo a la realidad en su estudio probabilístico del caso concreto.

II. Un algoritmo para todos o un algoritmo personalizable

En 2016, Bonnefon y su equipo realizaron un estudio sociológico para comprender cuáles serían las reacciones de los usuarios de VA ante los criterios de conducción de estos automóviles y qué primeras intuiciones morales se podían extraer de la sociedad sobre el tema[38]. Plantea tres actuaciones distintas ante situaciones de necesidad: que el coche deba virar hacia un peatón y matarlo para salvar a diez personas en la carretera (con un fundamento utilitarista del mal menor), que el pasajero deba sacrificar la vida del pasajero para salvar a esas diez personas (con igual fundamento utilitarista) o que deba virar hacia un peatón y matarlo para salvar la vida de un solo peatón (fundamentado en un principio de solidaridad que luego analizaremos y cuestionaremos)[39].

El primer caso recibió la aprobación de la mayoría de encuestados, mientras que el tercero apenas obtuvo opiniones positivas al respecto. El segundo caso es el que más interés trae a este apartado, ya que, si bien los usuarios lo veían como una decisión moralmente adecuada y que la generalidad de la sociedad debería adoptar, afirmaban a su vez que muy pocos comprarían un VA que siguiera este criterio[40]. Es decir, un modelo que no hiciera prevalecer la vida de sus pasajeros salvo justificaciones morales muy fuertes no tendría éxito

[36] LAWLOR, *op. cit.*, p. 197.
[37] GOODALL, N.J., "Away from Trolley Problems and Toward Risk Management", en *Applied Artificial Intelligence*, Vol. 30, Nº. 8, 2016, p. 819.
[38] BONNEFON et al., *op. cit.*, p. 2.
[39] Íbidem, 5.
[40] Ídem.

en el mercado y esto repercutiría en que no se podría alcanzar el deseado nivel de seguridad en carreteras que los VA prometen[41]. Ante estos resultados, el mismo año de 2016, el representante de Mercedes Benz Christoph von Hugo anunció que el algoritmo que implementarían en sus VA priorizaría siempre la supervivencia de sus pasajeros. No obstante, para probable sorpresa de la marca, esto provocó una reacción social intensamente negativa, dejándonos con una paradoja evidente[42]. Nyholm explica que esta contradicción se debe al enfoque meramente social del trabajo de Bonnefon y su equipo, ya que a la sociedad aún le falta experiencia en cuanto al funcionamiento de los VA y esto despierta intuiciones sin suficiente justificación que podrían cambiar una vez se incorporen estas tecnologías a la vida cotidiana[43].

Siendo así, se plantean dos vertientes contrapuestas. La primera defiende una configuración ética personalizable: crear algoritmos que permitan a los usuarios elegir (e incluso modificar a su antojo) el estándar ético que utilizará el VA ante situaciones de necesidad[44]. La posición opuesta sustenta una configuración ética obligatoria[45], un algoritmo de necesidad para todos por igual (o, por lo menos, igual para todos los que se encuentren bajo un mismo ordenamiento jurídico o varios sistemas normativos que tengan principios similares).

1. Configuración ética personalizable

Algunos autores angloparlantes denominan la primera vertiente *ethical knob* por ilustrarse como una rueda de selección del criterio ético a aplicar. Esta posibilidad viene justificada por las razones antes expuestas de que una configuración ética obligatoria podría anteponer la vida de otros frente a la vida de los pasajeros del VA en ciertos casos y el gran público sería, por lo tanto, más reticente a comprar VA con sistemas éticos predeterminados, causando un retraso en su implantación y en la subsiguiente mejora de la seguridad del tráfico[46]. Algunos recuerdan además que cualquier tipo de sistema reglado debería ser exhaustivo y eso supone tres problemas: hallar un consenso social sobre la ética universal a aplicar, crear un sistema ético completo para toda situación de necesidad y traducirlo al lenguaje informático del VA[47]. También afirman que se antoja difícil argumentar que la teoría ética elegida sea la que hace más segura la conducción en todos los casos[48] y que la determinación de responsabilidad sería mucho más difusa o imposible, creando un vacío de responsabilidad y arriesgando una percepción de injusticia (siendo la justicia retributiva uno de los fines del Derecho Penal[49]).

Autores como Millar o Contissa definen este algoritmo de VA como un proxy[50], un instrumento informático que sirve para traducir la voluntad ética del conductor en el lenguaje del VA, solventando así el problema de atribución de responsabilidad en caso de accidente

[41] COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, p. 1133.
[42] NYHOLM, *op. cit.*, p. 5.
[43] Ídem.
[44] Íbidem, 3.
[45] CONTISSA, G., LAGIOIA, F. & SARTOR, G., "The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law", en *Artificial intelligence and law*, 2017, Vol. 25, Nº 3, p. 366.
[46] BONNEFON et al., *op. cit.*, p. 7.
[47] GOODALL, N.J., "Ethical decision making during automated vehicle crashes", en *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2424, Nº. 1, 2014, p. 7.
[48] WENDEL, W.B., "Economic Rationality and Ethical Values in Design-Defect Analysis: The Trolley Problem and Autonomous Vehicles", en *California Western Law Review*, Vol. 129, Nº. 56, 2018, p. 159.
[49] FEILER, *op. cit.*, p. 19.
[50] CONTISSA et al., *op. cit.*, p. 366.

causado por decisiones que involucren al algoritmo de necesidad[51] y la reticencia a comprar VA por no estar conformes con su algoritmo ético (contarían con un algoritmo a medida y bajo su propio criterio); no teniendo siquiera que justificar su opción ética. Existen varias formulaciones de este algoritmo personalizable, pero nos centraremos en la propuesta de Contissa y su equipo de 2017[52], no solo como paradigma de esta postura y para hallar sus ventajas e inconvenientes, sino también para analizar elementos que pueden incorporarse a otros modelos posteriores.

Contissa et al. proponen una escala graduable de tres modos. El modo altruista tendrá preferencia hacia la vida de terceros por encima de la vida de los pasajeros, frente al modo egoísta, que alberga preferencia por la vida de los pasajeros sobre la vida de terceros. En un punto intermedio, el modo imparcial dará igual importancia a pasajeros y terceros, de modo que decidirá la opción que minimice el número de lesionados y, en caso de riesgo equiparable, decidirá aleatoriamente[53]. Pero el sistema es gradual y puede seleccionar también los intervalos entre niveles (por ejemplo, escoger un 30% de modo altruista).

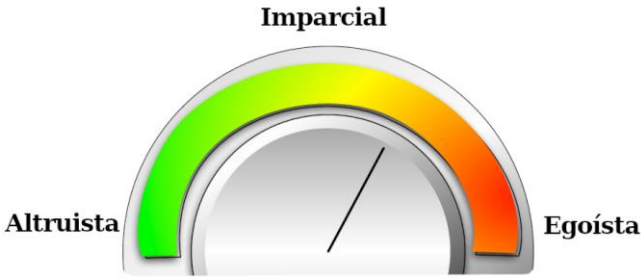

Imagen 1. El algoritmo ético personalizable (*ethical knob*)[54]

Por último, incorpora el modelo probabilístico de gestión del riesgo, que radica en que la graduación ética elegida por el usuario se transforma en una variable de peso (en importancia) de la vida de los pasajeros y, en oposición, de los terceros (si el valor que la empre será el valor restante de la primera variable) y se incorporará a un sistema probabilístico con otra variable: la probabilidad de muerte o lesión al pasajero y al tercero. Ambas variables se multiplican entre sí y por el número de actores en cada decisión para hallar cuál es la opción que crea menos perjuicio y, por ende, más adecuada[55].

Pondremos aquí una ilustración de un VA que circula por una carretera y, al girar una curva cerrada, se encuentra con un ciclista y sin margen para reducir la velocidad y recular: puede continuar su trayectoria, chocando con el ciclista con una probabilidad de muerte del 80% o girar hacia un lado estrellando el coche con una probabilidad de muerte del conductor del 40%. Si el usuario ajustara la configuración ética a un 70% de modo egoísta ya tendríamos dos variables: el peso de la vida del pasajero (0,7) y el peso de la vida de terceros (0,3). La opción de continuar supondría un valor de riesgo de 0,24 (0,8 x 0,3) y la alternativa de estrellar el coche contaría con un valor de riesgo de 0,28 (0,4 x 0,7), por lo que el VA decidiría continuar y chocar con el ciclista porque el valor del riesgo ajustado al criterio del algoritmo elegido por el usuario sería inferior.

[51] NYHOLM, *op. cit.*, p. 3.
[52] CONTISSA et al., *op. cit.*, pp. 365–378.
[53] Íbidem, 369.
[54] Íbidem, 370.
[55] Íbidem, 372.

Ahora bien, llegados a este punto, Contissa et al. reconocen a Gogoll que no vale cualquier estándar ético en el algoritmo personalizable[56]. El modo egoísta puede suponer un aumento ilegítimo de la probabilidad de accidentes más graves, por lo que propone solventarlo reduciendo el límite máximo de modo egoísta que podría ofrecer el sistema[57] para contener este problema. En vez de permitir al conductor elegir entre un 100% de modo altruista y hasta un 100% de modo egoísta, exigir (por poner un ejemplo) que los sistemas solo admitan al usuario optar desde un 100% altruista hasta un 50% (u otro margen que determine) de modo egoísta. Otra idea, alternativa o cumulativa a este límite del egoísmo, consiste en incrementar la cuantía y/o cobertura del seguro en caso de accidente[58].

Desde el estándar jurídico-moral más fundamental, este modelo no puede considerarse aceptable por simplificar en exceso la problemática y olvidar el principio básico que es la dignidad humana inherente a la vida y su carácter imponderable y no cuantificable[59]. De hecho, en caso de que la situación de necesidad no implique ningún riesgo al pasajero, Contissa et al. propone que el VA opte por una ética utilitarista[60], lo cual es una imposición que se aparta de la premisa de un algoritmo ético ajustable si lo que dice pretender este sistema es dejar las cuestiones morales al ciudadano[61].

Por añadidura, ofrece soluciones individuales y egoístas (por usar sus propios términos)[62] que cada usuario implementa en su VA y no tienen por qué coincidir con la voluntad social de una deliberación pública basada en argumentos que dé lugar a unas bases comunes[63], sino que pretende maximizar el interés individual como bandera bajo la que cobijar las decisiones diversas de los ciudadanos[64]. Simplemente se basaría en las preferencias polarizadas que enfrentan la vida de los pasajeros del VA frente a la vida de los terceros sin una justificación concreta y comprensible a por qué unas deberían valer más que las otras y que podría causar un clamor público, ya que el mero interés y voluntad de un ciudadano sin más no basta para justificar su priorización sobre la integridad de los derechos de los demás ciudadanos de la comunidad[65].

Además, vuelve a contar con los inconvenientes del modelo de gestión del riesgo: no contempla la provocación o negligencia de los actores[66], la figura de los espacios seguros[67] o demás cuestiones que consideramos esenciales en una solución basada en el principio de justicia y que son límites a la autonomía de la voluntad.

---

[56] GOGOLL, J. & MÜLLER, J.F., "Autonomous Cars: In Favor of a Mandatory Ethics Setting", en Science and engineering ethics, Vol. 23, Nº. 3, 2017, p. 688.
[57] CONTISSA et al., *op. cit.*, p. 375.
[58] Íbidem 378.
[59] SUÁREZ, *op. cit.*, p. 8.
[60] Íbidem, 370.
[61] NYHOLM, *op. cit.*, p. 3.
[62] CONTISSA et al., *op. cit.*, p. 366.
[63] GOODALL, N.J., "Away from Trolley Problems and Toward Risk Management", en *Applied Artificial Intelligence*, Vol. 30, Nº. 8, 2016, p. 819.
[64] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 245.
[65] COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, p. 1133.
[66] LAWLOR, *op. cit.*, p. 197.
[67] Ídem, 200.

Por otra parte, se presentan como ventajas de este método la clara delimitación de la responsabilidad y que solventa la desincentivación a la compra de VA que podrían priorizar la vida de otros sobre la de los pasajeros. Pero debemos comprender que, si la configuración ética personalizable apunta directamente al conductor que elige su parámetro moral como responsable del daño y excluye así el riesgo permitido[68], casi con total seguridad desincentivaría la compra de estos VA por el temor de la mayoría de consumidores a las represalias legales (y monetarias) en caso de accidente.

Finalmente, esta opción queda carente de fundamento al adjudicar la responsabilidad por un accidente derivado de una situación de necesidad directamente al usuario del VA por su elección ética. Esta afirmación elude un análisis esencial de la acción típica, antijurídica y culpable que todo delito exige. Como anticipo e ilustración, no profundiza en el debate sobre si programar un algoritmo para actuar en un futuro incierto de determinada manera produciendo daños personales o materiales es equivalente a hacerlo una persona con sus propios medios[69] y en el momento o si existe imprudencia o dolo en el usuario al programar el algoritmo con respecto al resultado concreto que después se produce[70].

2. Configuración ética obligatoria

De este modo, nos encontramos con la bibliografía opuesta: una configuración ética que aplica por igual a todos que puede no ostentar un pleno consenso, pero sí se sustenta por unos principios generalmente aceptados[71]. Esta propuesta entronca directamente con el interés general, un concepto jurídico indeterminado que constituye el estandarte de las Administraciones Públicas y que, por tanto, corresponde al gobierno investigar, deliberar, implementar y garantizar[72].

Se concibe como una teoría contractualista derivada de las ideas de Hobbes y concretada en el modelo de justicia de Rawls, que transforma esta idea del pacto social en una justicia como imparcialidad[73]. Este autor planteaba que una sociedad justa debía elegir sus principios y normas desde una posición original en abstracto anterior a la creación de la civilización[74] y que podía compararse el grado de justicia de las normas actuales según su mejor o peor acomodación a estas normas dictadas en la posición original. Para asegurar mejor la equidad e imparcialidad, existe un velo de la ignorancia sobre los ciudadanos que dictan estas normas, no conociendo cuál será su posición en la sociedad que están construyendo (clase social, nivel educativo, relaciones, sexo, edad…)[75]. A partir de esta base, Rawls construye su modelo de justicia para la sociedad.

Volviendo a nuestra área, varios autores trasladan la teoría de Rawls a unos VA en que la respuesta a los algoritmos de necesidad son unos estándares éticos universales razonados y

[68] GRANDI, *op. cit.*, p. 26.
[69] COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, p. 1141.
[70] FEILER, *op. cit.*, p. 22.
[71] GOODALL, N.J., "Away from Trolley Problems and Toward Risk Management", en *Applied Artificial Intelligence*, Vol. 30, Nº. 8, 2016, p. 819.
[72] CONSEJO DE EUROPA, *op. cit.*, p. 1.
[73] RAWLS, J., *Teoría de la justicia*, Fondo de Cultura Económica de España, 1999, p. 25.
[74] Ídem, 121.
[75] Ídem, 135-136.

argumentados desde una posición original y bajo el velo de la ignorancia[76]. Este estándar seguiría los principios generales que mejor protegieran los intereses de todas las personas involucradas en una situación de necesidad[77] de forma imparcial, equitativa, sin motivos discriminatorios o interesados y con respeto a la dignidad humana de cada vida en juego. Grandi[78] incluso propone que, al fundamentarse en un pacto social implantada desde los poderes públicos, las decisiones de esta ética del algoritmo imparcial deberían acogerse bajo la figura del riesgo permitido, no existiendo, por lo tanto, responsabilidad penal (sería atípico por no cumplir el requisito de imputación objetiva).

De hecho, el algoritmo personalizable no maximizaría la seguridad en el tráfico, ya que se encuadraría en el llamado dilema del prisionero, formulado por los matemáticos Merrill M. Flood y Melvin Dresher en 1950. Este dilema supone que hay dos presos que tienen la oportunidad de reducir su condena si confiesan y delatan al otro ante la policía. Si ninguno delata a su compañero, cada preso sería condenado a un año de prisión. Si ambos delatan a su compañero, no obstante, la pena sería de tres años para cada uno. Y si solo uno de ellos delata a su compañero, el que lo delató no sería condenado y el delatado pasaría cuatro años en prisión[79]. La cuestión es que ninguno de ambos sabe lo que decidirá el otro ni pueden pactar sus respuestas entre sí.

Es decir, si en la sociedad existiera un algoritmo ético personalizable, cada individuo podría considerar la ética altruista como la opción moralmente superior frente a la egoísta y, si todos optaran por ella, minimizarían el daño social en su conjunto[80]. No obstante, la opción egoísta es la más acorde a los intereses de cada individuo, porque prioriza su vida. Al no conocer a los demás ni saber el algoritmo que elegirán, el individuo no estaría dispuesto a correr el riesgo de ser el perjudicado (él eligiera la opción altruista y los demás, la egoísta), de modo que optaría por la ética egoísta[81]. Un algoritmo ético universal salvaría este problema imponiéndose de forma razonada y minimizando así los riesgos para todos los agentes implicados[82].

Vistas estas razones a favor de una configuración ética obligatoria frente a los inconvenientes de la ética personalizable de los VA, parece evidente que la primera es el camino a seguir. De esta manera, pasaremos a analizar iniciativas de diversos autores que siguen esta línea de pensamiento, pero divergen en su planteamiento y de los cuales podremos exprimir principios y métodos útiles en cuanto a la filosofía y la práctica para una solución lo más acertada y realista posible.

III. Un algoritmo utilitarista estricto

Una primera lectura sobre el tema llevó a resolverlo a través de la filosofía utilitarista. Consiste en la idea del mal menor o minimización social del daño social a través de la figura del interés preponderante y una teoría consecuencialista[83]. Es decir, opta por la alternativa que

[76] NYHOLM, *op. cit.*, p. 7.
[77] Ídem, 3.
[78] GRANDI, *op. cit.*, p. 26.
[79] GOGOLL & MÜLLER, *op. cit.*, p. 692.
[80] Íbidem, 694.
[81] Íbidem, 691.
[82] Íbidem, 695.
[83] COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, p. 1133.

minimice en mayor medida el daño o, visto desde la perspectiva opuesta, la que maximice su utilidad (de ahí el nombre)[84] u obtenga el mayor bien común, como lo definía Bentham (uno de sus firmes defensores)[85].

Lo ilustraremos de este modo, utilizando un ejemplo de Lawlor[86] a favor de una mejor comprensión de esta filosofía. Un coche y un ciclista sin casco se aproximan, cada uno en su carril, a un semáforo de una intersección al máximo de la velocidad permitida en la vía. La luz del semáforo se torna amarilla, pero, a la velocidad a la que circulan, frenar no aseguraría hacerlo a tiempo y ambos continúan la marcha confiando en que pasarán el semáforo y la intersección antes de que el semáforo se ponga en rojo. Lawlor nos informa de que esta presunción es correcta y razonable: ambos pasarán antes de que la luz pase a rojo. Sin embargo, justo cuando llegan al cruce, un motociclista con su pasajero a la espalda y ambos con los cascos de seguridad puestos decide saltarse su semáforo en rojo y cruzarse perpendicularmente a la trayectoria del coche y la bicicleta. El coche puede intentar frenar y golpear con la menor fuerza posible la parte trasera de la motocicleta. La segunda opción evitaría golpear a la motocicleta virando hacia la izquierda, pero golpearía al ciclista (que, sin casco, corre mayor riesgo de lesiones graves o incluso muerte). Por último, puede tratar de girar hacia la derecha, dejando a salvo al ciclista, pero golpeando de pleno al motociclista y su pasajero (lo cual supone un mayor riesgo de lesiones graves o muerte).

Así el ejemplo, una primera aproximación desde la perspectiva utilitarista optaría por girar a la izquierda, puesto que chocar contra un solo ciclista frente a chocar contra dos usuarios de la vía contiene un riesgo menor cuantitativamente, solo habiendo un afectado y maximizando su utilidad. Imaginemos ahora que a este sistema se incorporase el modelo de gestión del riesgo. Este valoraría de forma meramente material que el conductor y el pasajero de la moto llevan casco reglamentario y el ciclista no lo lleva: calcularía que el riesgo es de muerte para el ciclista y de meras lesiones para el motociclista y su pasajero si sigue de frente frenando. Por lo tanto, optaría por la opción que minimice los riesgos: intentar frenar, pero continuando su trayectoria y chocando con la moto[87].

Los argumentos que se presentan a favor de esta teoría se cimentan en la sencillez, ya que primero se nos ofrece el principio del mal menor como una mera traslación del método de resolución de conflictos que seguimos los seres humanos en el día a día trasladado a conflictos jurídicos de VA (y que sirven de inspiración a las causas de justificación penal): una ponderación de intereses[88]. Además, estos cálculos de probabilidades implementando el modelo de gestión del riesgo en su máxima expresión se podrían concretar de forma automática y casi infalible (cuanto más avance la tecnología)[89], al contrario que la valoración que debe hacer un

[84] GOODALL, N.J., "Machine Ethics and Automated Vehicles", en Meyer, G. & Beiker, S. (ed.): *Road Vehicle Automation*, Springer, 2014, pp. 93-102.
[85] SUÁREZ, *op. cit.*, p. 8.
[86] LAWLOR, *op. cit.*, p. 200. El ejemplo del *Pillion Passenger* que Lawlor ofrece se verá aquí ligeramente modificado a fin de funcionar de ilustración tanto para los puntos a favor del utilitarismo como sus inconvenientes y no tanto para analizar la imputación de responsabilidad y los espacios seguros.
[87] GOODALL, N.J., "Ethical decision making during automated vehicle crashes", en *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2424, Nº. 1, 2014, p. 8.
[88] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 246.
[89] Íbidem, 247.

conductor a diario cuando se encuentra frente a una situación de necesidad sin apenas margen de actuación y actúa casi por instinto[90].

Sin embargo, debemos rechazar también esta filosofía por las razones que se exponen a continuación. Volviendo a los problemas del modelo de gestión de riesgos, su combinación con el sistema utilitarista no resuelve el hecho de que sus resultados se basan en datos y la elección de qué datos deben incorporarse al cálculo de riesgos supone varias dificultades: en su obtención (el aspecto técnico), la controversia que provocaría en cuanto a la protección de datos (incluso considerándose una vulneración de privacidad)[91] y la comparación de valores normativos (sin un valor cuantitativo concreto) como el merecimiento, la provocación (puede castigar injustificadamente a personas ajenas por el peligro causado por un tercero de manera culposa o dolosa[92]) o la cualidad imponderable de la vida[93].

Del mismo modo, afirmar que las decisiones morales del ser humano en el día a día (y las causas de justificación) se basan solo en la ponderación del mal menor tampoco tiene un fondo sólido, especialmente desde el punto de vista normativo. Nuestro Derecho, como muchos otros y retornando a ese argumento contractualista[94], encuentra su raíz en la autonomía individual y el respecto a los derechos particulares de cada persona y las normas legales (y éticas) que permiten la convivencia de estos intereses en armonía[95]. Si bien a veces estas normas, de forma extraordinaria, deben resolver situaciones de necesidad mediante la ponderación de los intereses en juego en virtud de un principio de solidaridad más o menos interviniente[96], este no desplaza a la dignidad inherente a cada ser humano y al estudio detallado de todos los factores que derivan del ámbito de organización de cada actor involucrado. Ni mucho menos esto supone solventar el conflicto a través de la maximización de agregada de intereses que nombra Coca Vila[97], sino mediante la justa apreciación de las responsabilidades y posiciones jurídicas de cada uno y decidiendo por la opción más justa y adecuada.

Otro razonamiento que se opone al utilitarismo es que podría llegar a generar desincentivos para seguir las normas viales que se refieren a la autoprotección si se aplica de forma feroz o muy estricta[98]. En nuestro ejemplo, el algoritmo utilitarista consideraría la utilización del casco por el motorista y su pasajero y su ausencia en el caso del ciclista solo en virtud de hallar una mayor probabilidad de lesión. En términos simples, el VA elegiría chocar contra el conductor de una moto con casco frente al conductor de una moto sin casco porque el primero probablemente sufrirá daños más leves, condenando así a quien cumple las normas

[90] GOODALL, N.J., "Machine Ethics and Automated Vehicles", en Meyer, G. & Beiker, S. (ed.): *Road Vehicle Automation*, Springer, 2014, p. 97.
[91] GOODALL, N.J., "Ethical decision making during automated vehicle crashes", en *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2424, Nº. 1, 2014, p. 8.
[92] LAWLOR, *op. cit.*, p. 214.
[93] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 248.
[94] D'AMATO et al., *op. cit.*, p. 14.
[95] COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, pp. 1133-1134.
[96] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 250.
[97] Íbidem, 248.
[98] GOODALL, N.J., "Ethical decision making during automated vehicle crashes", en *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2424, Nº. 1, 2014, p. 8.

viarias mientras que el que las incumple se vería beneficiado[99]. Esto supone una pérdida del valor útil de estas medidas de seguridad, el usuario se preguntaría si realmente es más seguro ponerse casco o no[100].

No obstante, el utilitarismo es relegado por ciertos autores al último recurso, cuando ya se ha realizado un análisis deontológico profundo de la situación, los actores y sus intereses y derechos y el sujeto se encuentra ante una disyuntiva entre dos opciones normativamente idénticas[101]. Por ejemplo, el caso de un VA que, si frena, causará la muerte del motorista que lo sigue sin respetar la distancia de seguridad; pero, si no lo hace, atropellará a un grupo de peatones que se le han cruzado de forma sorpresiva y negligente. Tras analizar en profundidad el supuesto, Coca Vila[102] (aunque él enfrenta una vida frente a otra vida, no un grupo) asume que los intereses en juego son equivalentes y, de esta manera, debería decidirse de forma puramente aleatoria. Pero añadiendo el detalle de que sea un grupo se plantea la posibilidad de que se resuelva de forma meramente utilitarista: el mal menor es la muerte del motorista; pero no usando al motorista como medio y descartando el valor intrínseco de los intereses en juego, sino después de un exhaustivo análisis de los mismos y concluyendo que ponderan de forma equivalente[103].

Acumulando todas estas razones, no queda sino concluir que un utilitarismo estricto no es solución a nuestra problemática, quedándonos solo con la posición deontológica: un análisis completo de las posiciones de todos los actores implicados[104] en la situación de necesidad que asiente unos principios según los cuales crear unas reglas de decisión que, aunque difíciles de formular[105] y consensuar[106], son en realidad la propuesta más justa.

Algunas de estas pautas las hemos nombrado ya: imponderabilidad de la vida, humana, interés preponderante, responsabilidad o negligencia[107]… Ahora las examinaremos dentro de los distintos proyectos que los autores afines a esta solución han proporcionado.

IV. El modelo deontológico

1. Un plan preliminar a largo plazo

Antes de adentrarnos en la naturaleza y fundamento de los principios y reglas que deben conformar este modelo, mencionaremos una idea que introduce Goodall a raíz de los VA y la

[99] GOODALL, N.J., "Machine Ethics and Automated Vehicles", en Meyer, G. & Beiker, S. (ed.): *Road Vehicle Automation*, Springer, 2014, p. 99.
[100] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 271.
[101] SUÁREZ, *op. cit.*, p. 15.
[102] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 270.
[103] SUÁREZ, *op. cit.*, p. 11.
[104] COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, p. 1134.
[105] WILENMANN, *op. cit.*, p. 3.
[106] GOODALL, N.J., "Machine Ethics and Automated Vehicles", en Meyer, G. & Beiker, S. (ed.): *Road Vehicle Automation*, Springer, 2014, p. 99.
[107] COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, p. 1135.

formulación de las normas que deberán incorporar: una estrategia incremental de despliegue ético de los VA basada en tres fases que se amplían al mismo tiempo que avanza la tecnología para ir adaptándose a las mejoras que incluya[108].

La primera fase será una implementación de una ética racional elaborada por los expertos en ética, ingenieros, juristas y productores de VA que equilibre los intereses de los productores y trabajadores implicados en el diseño, fabricación y comercialización del producto y los intereses en juego en cada caso concreto de los usuarios de la vía. Sin embargo, el propio Goodall asume que cualquier conjunto normativo estará incompleto[109], de ahí su segunda fase.

Una segunda fase, pensada a medio plazo, involucra un entrenamiento de la propia IA para desarrollar sus propias fórmulas de resolución de situaciones de necesidad en base a gran cantidad de datos con respecto a simulaciones o grabaciones reales de humanos ante problemáticas morales complejas en la conducción y sus reacciones[110]. Esta etapa hibridaría el enfoque de IA bajo los principios del modelo de ética racional predeterminado mientras aún no se pueda asegurar que estas tecnologías extraen unos principios ideales y acordes[111].

Finalmente, una etapa en que la metodología de las IA sea suficiente y pueda comprenderse desde nuestro lenguaje. Aún no se han alcanzado los avances suficientes como para hacer transparente el proceso mediante el cual la IA extrae principios y toma decisiones (traducirlo al "lenguaje natural"). En cuanto se consiga, a largo plazo, Goodall afirma que la IA podría desarrollar un sistema ético completo y meramente supervisado por los humanos[112].

Hasta que la ciencia alcance a conquistar la IA, la segunda y tercera fases serán solo un proyecto que merece la pena considerar. Ahora sí nos adentraremos en esa primera fase normativa que debe ordenar sus principios y reglas.

2. Cuestiones previas

2.1. Los principios de autonomía y solidaridad

Para enfrentar nuestro sistema deontológico, Vila se aproxima a la respuesta a través de la justificación penal de nuestro ordenamiento[113], ya que, al hablar de situaciones de necesidad, se suele acudir a las bases de las causas de justificación, si bien no es exactamente la misma situación (la decisión viene predeterminada desde una posición no participante en los VA, al contrario que las situaciones de necesidad habituales, en que es uno de los actores el que debe tomar la decisión en una fracción de segundo). Estas figuras parten de un principio esencial, el de autonomía, y se valoran conjuntamente con el principio de solidaridad, estableciéndose límites mutuamente.

[108] GOODALL, N.J., "Ethical decision making during automated vehicle crashes", en *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2424, Nº. 1, 2014, p. 10.
[109] Íbidem, 8.
[110] Íbidem, 9.
[111] Íbidem, 11.
[112] Ídem.
[113] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 235.

El principio de autonomía se origina en una sociedad liberal en que cada ciudadano tiene un ámbito de libre organización que no puede coartarse ni ser injerido por otro[114], pero esto le coloca en una posición de responsabilidad por las consecuencias de sus actos en la fórmula del *casum sentit dominus* (asunción personal de la desgracia): si vulnera negativamente el ámbito de libre organización de otro individuo, le coloca en una posición susceptible de mayor carga perjudicial[115] (un mayor deber de tolerancia). Así se justifican la legítima defensa y el estado de necesidad defensivo contra quienes sean responsables de la causación del peligro y aquellos con una relación especial que no llega a la responsabilidad normativa, pero supone una vinculación a tener en cuenta por la doctrina[116] (a través de la causalidad y la imputación objetiva, como veremos después). Ahora bien, en una primera instancia, rige la regla del *laissez faire* siempre que no se interfiera en la esfera de libertad de los demás[117].

Sin embargo, nuestro ordenamiento cuenta también con el principio de solidaridad, en virtud del cual el individuo puede ver sacrificada su libertad (de forma no merecida o retributiva) a favor de la libertad de otro individuo o individuos de la comunidad en aras de un deber de cuidado hacia la colectividad[118]. El estado de necesidad agresivo se concibe en base a este principio: una situación de necesidad que permite al necesitado desplazar el peligro que lo amenaza hacia un ciudadano ajeno y no responsable[119] (bajo unos requisitos). Este es el caso que plantea Wendel de un vehículo que, tras una curva sin visibilidad, debe decidir si chocar contra otro vehículo que obstaculiza la vía, matando a sus cinco pasajeros, o girar hacia un peatón que cruza por la calzada, el cual no tuvo ningún papel en el peligro[120].

Como hemos mencionado, estos dos principios se limitan mutuamente. La solidaridad no funciona como un chivo expiatorio para justificar conductas que instrumentalicen al individuo como un número cuantificable para calcular daños[121] (como vimos con el utilitarismo). Partimos de un *casum sentit dominus*: si la amenaza proviene exclusivamente del individuo amenazado (*dominus*) y no de una fuente responsable del peligro exógena a él, no puede justificarse que lo desplace a terceros ajenos[122], debiendo examinarse con atención quién es el *dominus* en cada incidente atendiendo a quién posee en su ámbito de libre organización la capacidad de evitar o haber evitado previsiblemente la situación de peligro[123]. Pero tampoco el

---

[114] COCA VILA, I., "Entre la responsabilidad y la solidaridad. El estado de necesidad defensivo", en *Revista para el Análisis del Derecho InDret*, Nº. 1, 2011, p. 17.
[115] SUÁREZ, *op. cit.*, p. 6.
[116] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 252.
[117] SUÁREZ, *op. cit.*, p. 14.
[118] PAWLIK, M., "Solidaridad como categoría de legitimación jurídico penal: el ejemplo del estado de necesidad agresivo justificante", en *Revista de Estudios de la Justicia*, Nº. 26, 2017, p. 229.
[119] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 253.
[120] WENDEL, *op. cit.*, p. 155.
[121] SUÁREZ, *op. cit.*, p. 5.
[122] COCA VILA, I., "Entre la responsabilidad y la solidaridad. El estado de necesidad defensivo", en *Revista para el Análisis del Derecho InDret*, Nº. 1, 2011, p. 13.
[123] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 262.
Se presenta el supuesto en que un VA solo salvaría a sus pasajeros dirigiendo el coche hacia un carril de frenado de emergencia, pero sacrificando a dos niños que en él juegan. En los VA, la decisión del algoritmo deberá ser preestablecida, no ya como el conductor que pondera al momento desde una perspectiva subjetiva e individual, sino desde una posición externa y más objetiva. Vila advierte que no debemos asumir que la posición de dominus se atribuye automáticamente al conductor, ya que el VA no tiene conductor y se presume que circulará siguiendo

principio de autonomía hace intocable al ciudadano, ya que responde de las injerencias en la libertad de otros y el hecho de formar parte de la sociedad implica una serie de deberes de tolerancia (o incumbencia) para con los demás[124] siempre que esté justificado.

De hecho, estas causas de justificación de situaciones de defensa y necesidad cuentan con unos requisitos para comprobar si prima la solidaridad o la autonomía, respondiendo a tres cuestiones: si hubo o no una amenaza inminente de un mal sobre los bienes jurídicos del sujeto necesitado, si esa amenaza fue o no provocada por la propia víctima y una ponderación de los intereses concomitantes (ya sea a través de la necesidad racional o de una proporcionalidad más o menos estricta)[125].

2.2. La ponderación de intereses

Este requisito merece mención especial porque se convierte en una exigencia más rígida cuanto menos entra en juego la responsabilidad del tercero perjudicado. Tiene en cuenta los bienes jurídicos afectados (los intereses de cada actor) y considera cuál es la probabilidad de que se vean perjudicados y cuál la magnitud de ese posible perjuicio[126]. De este modo, será preferible chocar contra un escaparate (sin daños al pasajero) antes que dar muerte a un peatón[127] o hacer peligrar solo la integridad física de un sujeto frente a hacer peligrar la vida de otro (aunque es difícil de calcular, cobrando relevancia el modelo de gestión del riesgo).

No obstante, siempre debemos recordar que la vida es un bien imponderable frente a otra vida, no pudiendo cuantificarse ni tan siquiera acumulativamente para determinar qué vida o interés vale más[128]. Pongamos un ejemplo que nos ayudará a entender esta imponderabilidad y el concepto siguiente: la ilusoria primacía de deberes de omisión frente a los de actuación. Un VA circula por la derecha en una carretera de dos carriles y pasa una curva de escasa visibilidad. En el carril por el que circula, se cruzan súbitamente dos peatones (sin haber un paso habilitado) y, al mismo tiempo, en el carril izquierdo, un ciclista. El VA podría continuar y atropellar al peatón o virar y atropellar al ciclista, causando la muerte en cualquiera de los casos.

Los intereses en juego enfrentan vida contra vida, al no poder acumular las vidas de los peatones frente a la del ciclista, por lo que no podemos ponderar una sobre la otra. De acuerdo con el art. 124 del Reglamento General de Circulación (RCG), el peatón no debe cruzar la calzada. También el ciclista debería circular por el arcén si lo hubiera o lo más escorado a la derecha del carril si este no existiera, de acuerdo con el art. 17 de la Ley sobre Tráfico, Circulación de Vehículos a Motor y Seguridad Vial (LSV) y el 36 RGC. De esta manera, tampoco el merecimiento ha lugar, pues ambos son responsables de ocasionar un peligro.

---

las normas viales y sin producir ningún riesgo no previsto. Sino que, analizando quién de entre todos los agentes tuvo en su ámbito de actuación la libertad previsible de evitar esa fuente de peligro. En el ejemplo, el carril de frenado de emergencia es un cauce normativo para que los vehículos acudan en caso de avería, una actuación previsible, predeterminada y conforme a Derecho. De este modo, son los niños que accedieron a ese espacio previsto para la circulación urgente de los coches (también vinculado al concepto de los espacios seguros que propone Lawlor y contemplaremos a continuación) los que dentro de su ámbito de libertad organizativa crearon un riesgo exclusivamente de su responsabilidad y que, siendo así *domini*, deben soportar la desgracia conforme al principio de asunción personal de la desgracia.

[124] Íbidem, 4.

[125] D'AMATO et al., *op. cit.*, p. 11.

[126] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 256.

[127] SUÁREZ, *op. cit.*, p. 13.

[128] Íbidem, 8.

Finalmente, algunos alegan que el VA debería continuar porque virar implica una conducta activa con mayor desvalor que no frenar como una conducta omisiva. De acuerdo con Vila, esta teoría carece de unos valores fundados que le den sentido y además se topa con que cualquier acción puede expresarse como una omisión y cualquier omisión como acción[129]: no frenar equivale a continuar y viceversa, virar hacia la izquierda es lo mismo que no continuar de frente… No está creando un riesgo nuevo, sino que es afrontado por dos riesgos distintos a los que debe hacer frente y dirigir[130].

Por consiguiente, contamos con dos posibilidades: confiar en la probabilidad (con un cierto toque utilitarista) o decidir de forma aleatoria. Uno de los elementos de la ponderación es la probabilidad de producir el daño previsto; un modelo de gestión del riesgo perfeccionado y con datos suficientes, ante dos intereses equivalentes, podría optar por aquel con menor probabilidad de causar un perjuicio más grave[131]. Sin embargo, como aún no contamos con esta tecnología, tal vez sería adecuado dejarla para la segunda fase de la estrategia a largo plazo de Goodall. Por ahora, parece que la decisión aleatoria es la más acorde al principio de igualdad[132], ya que evita discriminaciones y aporta cierta sensación de justicia[133].

3. Reacción a agresiones ilegítimas

Los casos más claros en los que un VA puede atentar contra los bienes jurídicos de otros justificadamente coinciden con ese primer principio de autonomía: si se coarta el ámbito de libertad organizativa de un individuo, está legitimado reprimir esa coacción[134]. Si bien no es el prototipo de peligro en la carretera, se puede resolver utilizando los criterios de la legítima defensa: cuando un sujeto en la vía dirija contra el pasajero del VA una agresión ilegítima (inminente, actual y antijurídica) que no haya sido provocada por el pasajero, el VA podrá lesionar los bienes jurídicos del ofensor siempre que utilice el método menos dañino desde una perspectiva ex ante de entre todos los medios a su alcance[135].

Adoptemos el ejemplo de un VA parado por una congestión del tráfico (que esta tecnología pretende evitar, pero ya vimos que no es infalible cuando elementos externos con libertad de actuación entran en juego; por ejemplo, unas obras pueden ralentizar el tráfico) en una carretera con dos carriles cuando, de pronto, se sitúa una motocicleta en paralelo y efectúa varios disparos contra el VA. No pudiendo huir hacia ninguna dirección, lo único que podría intentar es utilizar el poco margen de maniobra para embestir de lado a la motocicleta, pudiendo causarle lesiones graves e incluso la muerte. Esta situación (de película) presenta una agresión ilegítima e inminente no provocada por el pasajero del VA y ante la cual se presenta la

[129] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 266.
[130] WENDEL, *op. cit.*, p. 157.
[131] GOODALL, N.J., "Away from Trolley Problems and Toward Risk Management", en *Applied Artificial Intelligence*, Vol. 30, Nº. 8, 2016, p. 814.
[132] SUÁREZ, *op. cit.*, p. 15.
[133] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 270.
[134] WILENMANN, *op. cit.*, p. 5.
[135] MARTÍNEZ et al., *op. cit.*, p. 275.

embestida como una opción de necesidad racional, ya que no tiene ningún otro medio menos gravoso para hacer frente a la agresión[136].

4. Criterios del estado de necesidad agresivo

El estado de necesidad agresivo se contempla normativamente como una situación de peligro inminente a los bienes jurídicos de una persona y no provocada por esta ante la cual ese mismo perjudicado o un tercero no puede oponerse sino lesionando los bienes de un tercero ajeno al conflicto para evitar el mal, siempre que el mal que cause no supere en gravedad a aquel que lo amenaza y que el sujeto no deba sacrificarse por su oficio o cargo[137]. La cuestión del deber de sacrificio por cargo u oficio puede solventarse por un mecanismo de identificación del pasajero del VA cuando se trate de estos individuos, como harían los miembros de las Fuerzas y Cuerpos de Seguridad, para que el algoritmo lo considere. La diferencia radica en la regla de proporcionalidad entre el mal que amenaza al necesitado y el mal que este desplaza hacia el perjudicado[138] y en el hecho de que el perjudicado no provocó el peligro y es, por ende, ajeno a la situación de conflicto[139].

A modo de ejemplo, imaginemos un VA que, ante un peatón que cruza repentinamente la carretera y sin tiempo para frenar sin chocar con él, tiene la opción de virar hacia la acera, chocando y matando a un perro que por ella circulaba. Cumple con los elementos de una situación de necesidad no provocada por el pasajero del VA pero que amenaza la vida de un peatón negligente (como mínimo) y desplaza el mal hacia un bien jurídico que sería el patrimonio del dueño del perro. Atentar contra el bien jurídico patrimonial (la vida del perro) es menos grave que atentar contra una vida humana, por lo que obedecería a la proporcionalidad y se trataría de un estado de necesidad agresivo justificante.

Ahora bien, si rige una estricta proporcionalidad, el VA deberá desechar cualquier opción que contemple atentar contra la vida de un tercero no involucrado, ya que hemos dicho que es imponderable[140] y no puede solucionarse mediante el estado de necesidad justificante. Si bien algunos proponen el estado de necesidad exculpante cuando se enfrentan males de igual valor, en que la conducta es antijurídica pero no cabe exigirle otro comportamiento (inexigibilidad como elemento de la culpabilidad)[141]. Esta causa de exculpación se basa en que el sujeto plenamente capaz se encuentra bajo una presión extraordinaria como para requerirle que tome la decisión que sería más correcta en frío[142]. Pero pierde su sentido al trasladarlo a los VA, ya que, además de que el VA no tiene subjetividad penal, el programador o fabricante que creó su algoritmo se encuentra con una lejanía suficiente con el hecho al diseñar el programa como para no encontrarse frente a frente en esa situación de necesidad crítica en el momento[143],

[136] MUÑOZ RUIZ, J., *El delito de conducción temeraria: análisis dogmático y jurisprudencial*, Dykinson, 2014, pp. 309-310.
[137] MARTÍNEZ et al., *op. cit.*, p. 292.
[138] CONTISSA et al., *op. cit.*, p. 369.
[139] WILENMANN, *op. cit.*, p. 6.
[140] SUÁREZ, *op. cit.*, p. 13.
[141] MUÑOZ RUIZ, *op. cit.*, p. 314.
[142] MARTÍNEZ et al., *op. cit.*, p. 330.
[143] SUÁREZ, *op. cit.*, p. 6.

de modo que no está bajo la presión que hace inexigible que el individuo tome la opción menos ajustada a Derecho[144].

Por ejemplo, un conductor sobrepasa una curva y se percata de un árbol caído en la vía: si continúa de frente, chocaría con la posibilidad de causarle la muerte; si virara hacia la acera, chocaría y mataría a un transeúnte, pero salvaría su vida. Un conductor humano que eligiera la segunda opción tendría razones para alegar un estado de necesidad exculpante: el poco margen de tiempo, la tensión súbita del momento y el afán de supervivencia lo avalan, no se le podría exigir a nadie actuar de otra forma (aunque puede hacerlo)[145]. Pero si el VA estuviera al control del vehículo, este instinto de supervivencia, tensión o poco margen de tiempo no se podrían atribuir ni al VA ni al programador que lo diseñó mucho antes en una posición segura, por lo que los principios que hemos visto optarían por intentar frenar, pero continuar de frente, arriesgando la vida de los pasajeros.

Para concluir, en caso de encontrarse el VA en una situación de necesidad no provocada por el pasajero, solo podrá desplazar el riesgo hacia un tercero ajeno si el mal que pretende dirigirle es menor a aquel que amenaza al pasajero[146]. Si, por el contrario, esta misma situación solo puede evadirse atacando el bien jurídico de la vida de un tercero ajeno al conflicto, hemos visto que no hay justificación ni exculpación posible, de modo que los pasajeros del VA solo podrán dirigir el mal contra quien lo provocó o soportar la desgracia[147].

Es cierto que vuelve a situar sobre la mesa el impacto negativo que la posibilidad de que el VA priorice otras vidas sobre las de sus pasajeros tiene en la comercialización de los VA. Sin embargo, cabe remarcar que este sistema limita esta posibilidad a casos muy concretos en que se arriesgue la vida tanto de los pasajeros como la de los terceros ajenos al conflicto de forma equivalente y no exista ninguna otra opción de rescate; conjuntamente, se prevé que la implantación de los VA en la circulación reduzcan aún más la probabilidad de que estas situaciones de necesidad ocurran. Y, por último, se respalda en principios y fundamentos bien argumentados que nuestro ordenamiento decidió implementar y lleva poniendo en práctica en el día a día varios lustros, eliminando así la sensación de injusticia a través del razonamiento.

Aquí cabe el modelo de gestión del riesgo, capaz de calcular la probabilidad de afección de los bienes jurídicos en juego y de su gravedad[148]. En este respecto, deberá consensuarse si una mínima probabilidad de afectar a la vida del tercero implica *per se* desechar esa alternativa o si hay un baremo concreto bajo el cual se puede desplazar el mal hacia el tercero porque la probabilidad de causarle la muerte es suficientemente baja como para tenerla en cuenta.

---

[144] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 240.

[145] MARTÍNEZ et al., *op. cit.*, p. 295. Perspectiva que ni siquiera es compartida por toda la doctrina, ya que las teorías monistas no la avalan, pero que merece nuestra atención por realizar la comparativa entre los conductores humanos y los VA partiendo desde la teoría diferenciadora que sí concibe este estado de necesidad exculpante desde el dilema de la tabla de Carneades propuesto por el filósofo homónimo.

[146] MUÑOZ RUIZ, *op. cit.*, p. 313.

[147] COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, p. 1135.

[148] GOODALL, N.J., "Away from Trolley Problems and Toward Risk Management", en *Applied Artificial Intelligence*, Vol. 30, Nº. 8, 2016, p. 814.

4.1. Espacios seguros

Solo como un apunte adicional, Lawlor reformula en 2022 esta figura del tercero ajeno al conflicto como los llamados espacios seguros (*safe zones*), que no siempre coinciden con espacios físicos protegidos por la norma, sino más bien con posiciones normativas con un derecho de preservación preferente porque el sujeto que se encuentra en ellas lo hace de forma lícita y no indebida[149]. Defiende que los VA en casos dilemáticos no deberían virar hacia los sujetos que se encuentren en esos espacios seguros, salvo que el riesgo de lesión a sus bienes jurídicos sea ínfimo (para Lawlor, muy inferior al 1%)[150]. Si bien no podemos aceptar esta restricción para todos los casos, encaja perfectamente con el modelo de la justificación penal para el estado de necesidad agresivo que hemos analizado: cuando el mal a provocar es menor al mal que lo amenaza, puede justificarse en nuestro ordenamiento; pero cuando los males se ponderan igual, entran en juego los espacios seguros, un punto de partida para valorar qué opciones deberían descartarse en estas disyuntivas[151]. Así, si los bienes jurídicos en juego son equivalentes, no podrá desplazarse el mal si hay una probabilidad mayor al 1% de perjudicar al bien jurídico de un tercero ajeno al conflicto.

Un ejemplo de espacios seguros derivados de la norma son las aceras, pasos de peatones y zonas peatonales (siempre que los peatones circulen ahí de forma debida), pues el art. 121.5 RGC prohíbe a los vehículos circular por aceras y demás zonas peatonales. También el art. 25.1 a) LSV avisa de que el conductor de un vehículo no tiene preferencia respecto a los peatones en estos espacios.

Las mismas normas (arts. 30.1 y 77 f) de LSV y 89.1 RGC) determinan que el carril de sentido contrario solo podrá invadirse en estricto caso de necesidad por un obstáculo en la vía, pero siempre asegurándose de que pueda efectuar el cambio sin provocar ningún peligro. Así, esta regulación otorga una protección especial menor al carril de sentido contrario.

Pero Lawlor va más allá y se aproxima ahora al VA y otros vehículos también como espacios seguros, sea que haya conductor que circule debidamente o que no lo haya y los pasajeros no controlen la situación y circulen de forma debida en el VA sin manifestar una conducta irresponsable[152]. De esta manera, invita a considerar a todos los actores y considerar que, si no se encuentran en un sitio en que no deben o haciendo algo que no deben, esos actores están en un espacio seguro y no pueden verse afectados salvo que el mal dirigido hacia ellos sea menor que el mal que amenaza al necesitado.

Este método supone una mejora en claridad y seguridad jurídica para los usuarios de la vía, que conocen que la probabilidad de ver en peligro su vida disminuye en caso de circular o participar debidamente de los espacios viales. Razón además de prevención general, ya que motiva a seguir estas normas de debido comportamiento[153]. Se encuadraría concretamente al atender a los terceros ajenos al peligro que sufrirían un perjuicio si juzgáramos en virtud de los criterios del estado de necesidad agresivo, alegando que ya no pueden verse afectados si los males son equivalentes.

[149] LAWLOR, *op. cit.*, p. 200.
[150] Íbidem, 206.
[151] Íbidem, 197.
[152] LAWLOR, *op. cit.*, p. 203.
[153] Íbidem, 214.

5. Criterios del estado de necesidad defensivo

Entre la legítima defensa y el estado de necesidad agresivo, la doctrina alemana concibió y cierta parte de la doctrina española comenzó a adoptar una figura intermedia, el estado de necesidad defensivo[154]: el necesitado se defiende contra aquel que ve vinculado su esfera de libertad organizativa a la fuente del peligro (no una causación exógena), pero no derivado de una agresión antijurídica y responsable[155]. Se trata de peligros que provienen de hechos imprudentes o que no constituyen acciones (por animales o cosas si hay una especial posición jurídica del propietario o poseedor) ante los cuales el necesitado reacciona atentando contra un bien jurídico de la fuente de peligro[156]. Así, el sujeto aún sigue siendo protegido por la eximente, en mayor medida que el estado de necesidad agresivo, pero con unos requisitos más estrictos que los supuestos de legítima defensa.

Atiende a los dos principios ya propuestos de autonomía y solidaridad como títulos de responsabilidad penal. Por una parte, esa esfera de libertades de cada individuo comporta una responsabilidad por los efectos que esta tenga al interferir en la esfera de intereses de otro ciudadano[157], es decir, impone unos deberes negativos. Pero el compromiso no acaba ahí, sino que también formar parte de una comunidad impone unos deberes positivos de solidaridad o cuidado que pueden incluir una obligación de sacrificio (o un deber de tolerancia graduable) [158] y cuyo incumplimiento también supone responsabilidad. En efecto, la mayoría de los casos que hemos propuesto hasta ahora podrían reformularse como situaciones de necesidad defensiva si, en vez de desplazar el mal hacia un tercero ajeno o uno responsable de la fuente de peligro, el VA lo trasladara hacia un sujeto materialmente relacionado con ese peligro que interfiere en los intereses ajenos de forma previsible, pero no dolosamente.

Dicha definición se nutre de la doctrina de los deberes intensificados de solidaridad o la incumbencia preferente, que suponen una alteración en la posición jurídica de ciertos sujetos que, de forma justificada, exige un deber de tolerancia más intenso[159]. Se construye a partir de otros conceptos como la provocación, la posición de garantía derivada del actuar precedente o la omisión del deber de socorro agravada del art. 195.3 del Código Penal (CP)[160]. La provocación de un riesgo para otros, ya comentada a raíz de la legítima defensa, implica un deber de tolerancia elevado frente a la reacción por parte del defendido por la creación responsable del riesgo que interfiere en su esfera de libertades, ligado a la idea kantiana de

[154] COCA VILA, I., "Entre la responsabilidad y la solidaridad. El estado de necesidad defensivo", en *Revista para el Análisis del Derecho InDret*, Nº. 1, 2011, p. 2.
[155] WILENMANN, *op. cit.*, p. 16.
[156] REAL ACADEMIA ESPAÑOLA, "Estado de necesidad defensivo", en *Diccionario panhispánico del español jurídico (DPEJ)*. Recuperado en 31 de febrero de 2024, de https://dpej.rae.es/lema/estado-de-necesidad-defensivo
[157] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 252.
[158] PAWLIK, *op. cit.*, pp. 222-247.
[159] COCA VILA, I., "Entre la responsabilidad y la solidaridad. El estado de necesidad defensivo", en *Revista para el Análisis del Derecho InDret*, Nº. 1, 2011, p. 18.
[160] Ídem. También se contempla por algunos autores el nacimiento de este concepto a partir de la responsabilidad penal de las comercializadoras que cumplan todos los deberes de cuidado por un producto que se torna defectuoso por caso fortuito. Discurren diversas discusiones entre teorías (razón por la cual no se incluye aquí) sobre si la comercializadora debería ser responsable por el defecto en estos bienes vendidos al no tratarse de una acción antijurídica el haber comercializado el producto conforme a los estándares objetivos de cuidado. Pero, aun en el caso de no generar una responsabilidad penal por el producto fortuitamente defectuoso, sí puede considerarse un actuar precedente legal que la sitúe en esa posición de incumbencia preferente para retirar los productos defectuosos del mercado.

libertad[161]. La posición de garantía de la comisión por omisión derivada del actuar precedente o injerencia dispone que el sujeto será garante de los bienes jurídicos de otro individuo si su acción u omisión previa lo hace peligrar[162], de modo que es posible que alguien adquiera unos deberes especiales por su propia acción u omisión anterior.

No obstante, Vila contempla que tanto la responsabilidad como el deber de tolerancia pueden regularse. Lo ejemplifica a través del delito de omisión del deber de socorro agravada: aquel que no socorriere a una persona desamparada y en peligro manifiesto y grave, pudiendo hacerlo sin peligro, y cuya causa de agravación es que la misma persona ocasionado el accidente fortuita o imprudentemente sobre la víctima[163]. Este artículo tipifica un punto intermedio de responsabilidad entre la omisión pura y la comisión por omisión, unos deberes cualificados fundados en que el conductor no acudió en defensa de la víctima, siendo más cercano al mismo que cualquier otro tercero ajeno al choque de bienes jurídicos[164].

De igual modo, Vila propone que la legítima defensa supone un deber de tolerancia en sentido estricto que sigue a la acción libre y responsable del agresor (principio de autonomía), el estado de necesidad agresivo impone un deber de sacrificio (más rígido en sus presupuestos) sobre el tercero ajeno al conflicto derivado del principio de solidaridad y el estado de necesidad defensivo abarca tanto el deber de solidaridad intensificado (o deber de colaboración) como una autonomía del sujeto que, en su ámbito de libertad, ha afectado a otro de manera no responsable[165]. De este modo, en el estado de necesidad defensivo pueden lesionarse intereses que, al ponderarse, sean superiores a los que el necesitado salvaguarda, pero en menor medida que siguiendo los criterios de la legítima defensa. La doctrina fundamenta esta eximente de dos maneras: o bien la incluye en el estado de necesidad genérico debido a que la ponderación de intereses situaría el interés del necesitado por encima del interés de aquel que causó materialmente (aunque no normativamente) el peligro, o bien usa una analogía favorable al reo (conforme al ordenamiento) con la legítima defensa y el estado de necesidad agresivo por su similitud en estructura y principios[166].

El ejemplo más básico de estado de necesidad defensivo, ahora que hemos comprendido el espectro que abarca y sus elementos, nos lo proporciona Holstein[167]: un VA circula por una calle al máximo de velocidad permitida (50 km/h) cuando, de pronto, cruza la calzada de forma imprudente un grupo de jóvenes; el vehículo puede intentar frenar, chocando con ellos y lesionándolos, o virar hacia otro peatón que circula por la acera y ajeno al trance. Desde el estado de necesidad defensivo, estaría justificada la primera alternativa por haber causado imprudentemente el supuesto de necesidad.

En cuanto a ese vínculo entre el peligro y el sujeto *cuasirresponsable*, ciertas posiciones afirman que basta un nexo de causalidad que conecte a ambos, poder reconducir el peligro al ámbito de actuación del sujeto desde la *conditio sine qua non*, lo que se llama una ubicación

[161] WILENMANN, *op. cit.*, p. 5.
[162] MARTÍNEZ et al., *op. cit.*, p. 177.
[163] Íbidem, 184.
[164] COCA VILA, I., "Entre la responsabilidad y la solidaridad. El estado de necesidad defensivo", en *Revista para el Análisis del Derecho InDret*, Nº. 1, 2011.
[165] Íbidem, 29.
[166] REAL ACADEMIA ESPAÑOLA, "Estado de necesidad defensivo", en *Diccionario panhispánico del español jurídico (DPEJ)*. Recuperado en 31 de febrero de 2024, de https://dpej.rae.es/lema/estado-de-necesidad-defensivo
[167] HOLSTEIN & DODIG-CRNKOVIC, *op. cit.*, p. 32.

fáctica de la fuente del peligro[168]. Sin embargo, autores como Robles o Vila oponen que una persona que atente con un nexo causal, pero sin una previsibilidad concreta de que su acción derivará en un riesgo hacia otros peatones o pasajeros de vehículos, en realidad no puede tratarse como si fuera participante de ese riesgo ya que no constituiría ni tan siquiera una acción típica al no cumplir los requisitos del tipo objetivo[169]. Así, para conectar una acción humana con la creación de un riesgo o peligro para los vehículos (VA en lo que a nosotros respecta) y tener relevancia en el Derecho que sea una línea divisoria con respecto a aquellos que no tuvieron tal relación responsable[170], solo puede exigirle un deber de colaboración (y responsabilidad graduada) a aquel que tanto desde la causalidad como desde la imputación objetiva pueda vinculársele su acción y la fuente del peligro[171].

Pese a ello, existe aquí una pega cuando trasladamos estas reglas a los VA. Es comprensible que, en el estado de desarrollo actual de los VA, las reglas de la imputación objetiva no sean traducibles aún al algoritmo de modo que observando a una persona de la que parece desprenderse un peligro pueda conocer si era previsible para ella que su actuar precedente causaría dicho riesgo y, por consiguiente, no sabría distinguir quién causó el riesgo de forma imputable y quién lo hace solo causalmente, sin relación de imputación objetiva. Dado que la jurisprudencia parece apoyar de manera más firme la ubicación fáctica del peligro [172] (probablemente por la dificultad que supone en una situación de necesidad comprender estas cuestiones) y que los VA no comprenden estas peculiaridades aún, cabe que propongan que, en la primera fase racional, el principio a seguir exija solo una relación de causalidad con respecto al sujeto fuente del peligro y, a medida que avanzamos hacia la fase híbrida y la metodología de IA y su traducción al lenguaje natural[173], pueda ir incorporándose el refuerzo de ser conductas imputables objetivamente. Siempre fundamentado en el riesgo permitido por los beneficios proporcionalmente mayores que acarrean los VA y hasta un avance informático y de IA suficiente para traducirlo a lenguaje algorítmico.

Por aclararlo a través de un ejemplo adaptado de Lawlor[174], imaginemos un día de viento muy violento en una carretera de montaña con una vía peatonal adyacente. El viento tumba a dos peatones que andaban por la vía peatonal al asfalto en la trayectoria de un VA, que solo puede chocar contra ese peatón y matarlo, virar hacia el precipicio y perder la vida o virar hacia el arcén y matar a otro peatón que no fue tumbado por el viento. El VA sería incapaz de distinguir que los dos peatones fueron lanzados por el viento (caso fortuito), si lo hicieron adrede o si fueron empujados por alguien, además de que el otro peatón es totalmente ajeno al conflicto y ni él ni el pasajero del VA realizan ninguna conducta indebida. Así, combinando la teoría de los espacios seguros, la sencillez de la ubicación fáctica del peligro y los deberes intensificados de solidaridad que se les imponen a los dos peatones aún por un caso fortuito, la solución más correcta conforme al estado de necesidad defensivo sería que los dos peatones lanzados contra la vía deberían soportar la desgracia.

---

[168] COCA VILA, I., "Entre la responsabilidad y la solidaridad. El estado de necesidad defensivo", en *Revista para el Análisis del Derecho InDret*, Nº. 1, 2011, p. 6.
[169] Íbidem, 10.
[170] Íbidem, 28.
[171] Íbidem, 35.
[172] Íbidem, 20.
[173] GOODALL, N.J., "Ethical decision making during automated vehicle crashes", en *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2424, Nº. 1, 2014, p. 11.
[174] LAWLOR, *op. cit.*, p. 198.

6. Las capas de seguridad del VA: cuándo interviene el algoritmo de necesidad

Por terminar de cerrar nuestro sistema deontológico del algoritmo de necesidad de los VA, echaremos un vistazo a la propuesta de D'Amato y su equipo: unos principios que van a coincidir con nuestros fundamentos de derecho del algoritmo y, sobre todo, unas capas de seguridad del VA que van a responder a la pregunta de cuándo debe el VA actuar conforme a estos principios (ni antes ni después) en función de la proximidad e inminencia del riesgo[175].

Los principios de D'Amato y compañía empiezan por enunciar que ningún usuario de la vía puede resolver una situación de necesidad desplazando el mal hacia un tercero no inicialmente involucrado en el peligro[176], cuestión que ya hemos matizado: salvo que el interés del tercero ajeno al conflicto pondere en un nivel inferior al interés del necesitado, este será el modelo a seguir[177].

Aporta como segundo principio un énfasis en el cumplimiento de las normas de tráfico. Hasta ahora solo nos hemos centrado en cómo resolver las situaciones de necesidad ante las que se puede encontrar un VA, siempre argumentado conforme a las normas y principios de nuestro ordenamiento. Pese a ello, en el día a día cotidiano y especialmente en los momentos inmediatamente anteriores a la situación de necesidad (cuando se detecta un riesgo que aún no es inminente), por supuesto, el VA deberá respetar las reglas de circulación tal cual se estipulan[178].

Ahora bien, D'Amato et al. asumen que habrá situaciones en que la necesidad de prevenir un accidente o un riesgo para algún usuario de la vía requiera que el VA se salte algunas normas, de nuevo acudiendo a la ponderación de intereses en casos de emergencia que no supongan un riesgo relevante a la esfera de actuación de otros usuarios de la vía[179]. Un ejemplo de esto puede ser excederse de la velocidad máxima de la vía para llegar a tiempo a un hospital que salve la vida de su ocupante[180] al que no alcanzaría sin incumplir esta norma y en una vía con muy poco tráfico, con las premisas ya descritas.

Además, con carácter preventivo, afirman que el programa de trayectoria de los VA deberá disponer unas condiciones suficientes para una movilidad óptima[181] (reducir la velocidad, mantener una distancia de seguridad prudencial y acorde a la velocidad del VA, adelantar con una distancia suficiente…) con respecto a los demás vehículos y peatones, teniendo así margen para reaccionar. Esta es una de las razones que hace al VA una opción deseable para la seguridad en carretera[182]. De hecho, establecen un sistema de seguridad para el caso hipotético de que el VA sufra una avería que limite su capacidad de obtener o procesar la información: el VA actuará conforme a los datos inmediatamente anteriores al fallo técnico ejecutando la maniobra más adecuada según los algoritmo y probabilidades para frenar el vehículo apartado del tráfico y sin riesgo para los pasajeros y demás usuarios[183].

---

[175] D'AMATO et al., *op. cit.*, p. 18.
[176] Íbidem, 15.
[177] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 235.
[178] D'AMATO et al., *op. cit.*, p. 16.
[179] Íbidem, 15.
[180] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 239.
[181] D'AMATO et al., *op. cit.*, p. 17.
[182] WENDEL, *op. cit.*, p. 132.
[183] D'AMATO et al., *op. cit.*, p. 23.

Nos adentramos entonces en las tres capas o espacios de seguridad que estos autores conciben: la capa de colisión, la capa del deber de cuidado y la capa de conducción cómoda. La capa de comodidad en la conducción se refiere a esas distancias que mencionábamos que permiten al VA reaccionar con un amplio margen de maniobra como para no afectar a ningún bien jurídico y sufriendo solo una ligera incomodidad en la conducción[184]. Como ilustración nos valdrá el VA que ve al vehículo de delante frenar para incorporarse a una salida y decide reducir la velocidad y adelantarlo por el carril izquierdo.

La capa del deber de cuidado se calcula entre dos puntos: desde la distancia mínima a la que un VA puede llegar a maniobrar objetivamente (sin tener en cuenta las posibles ramificaciones de actuación de otros usuarios, sino simplemente basado en los datos de distancias, velocidad, adherencia y estado de la vía…) con una modificación de trayectoria incómoda o incluso violenta, pero sin llegar a crear un riesgo de afectar a la distancia de colisión[185]. Normalmente se corresponde con la distancia de seguridad reglamentaria o un valor ligeramente inferior.

En tercer lugar, el recorrido de colisión es aquella desde la propia estructura del VA hasta la distancia mínima bajo la que un VA no podría evadir una situación de necesidad sin chocar contra el otro usuario y provocar un accidente[186].

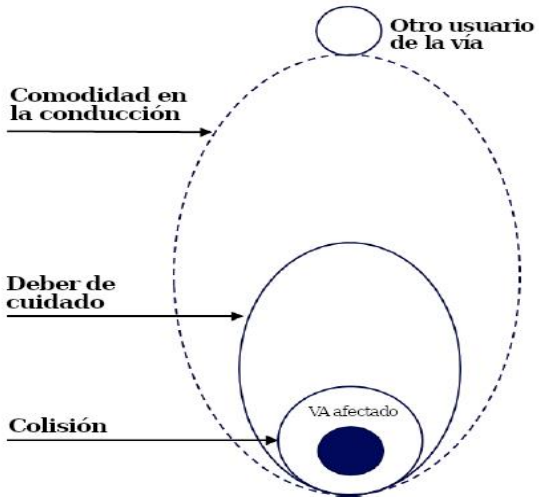

Imagen 2. Las capas de seguridad de los VA en relación con otros usuarios de la vía[187].

Lo ideal sería que todos los usuarios discurrieran sin entorpecer a otros usuarios ni que estos se adentren más allá de la capa de comodidad en la conducción ni tener que violar ninguna de las normas de tráfico, de modo que simplemente el VA se adherirá a la alternativa que restablezca una conducción segura con la mínima interferencia en la capa de conducción cómoda, si es posible[188]. Pero nuestro trabajo consiste en plantearnos qué hacer cuando vaya más allá y cuándo entra en la ecuación el algoritmo de necesidad: D'Amato y su equipo dan unas pautas. En segundo lugar, cuando se amenace la capa del deber de cuidado, el VA deberá seguir cumpliendo las reglas de tráfico y distancias de seguridad con los demás; en caso de no

[184] Íbidem, 17.
[185] Íbidem, 19.
[186] Íbidem, 20.
[187] Ídem.
[188] Íbidem, 22.

ser posible, solo podrá vulnerar la capa de cuidado del usuario del que derive el peligro[189]. Por último, nos encontramos con violaciones del espacio de colisión: cuando pueda salvarse el peligro sin arriesgar el espacio de cuidado o, mejor aún, la capa de comodidad en la conducción, el VA optará por esa opción[190]. No obstante, cuando un usuario se introduzca en la capa de colisión del VA es el momento que suscita nuestro trabajo desde el principio: la creación de un riesgo inminente de accidente a los bienes jurídicos de los pasajeros que haga inevitable una reacción que viole la norma o cause un daño a otros actores[191]. Es entonces cuando las reglas deontológicas del algoritmo de necesidad cobran sentido y deben implementarse para decidir la opción más ética y justa para todos los actores intervinientes.

[189] Ídem.
[190] Íbidem, 23.
[191] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 238.

LA IMPUTACIÓN DE RESPONSABILIDAD PENAL DERIVADA DEL ALGORITMO DE NECESIDAD

I. La atribución de responsabilidad en el panorama de los VA

Una sentencia de la Audiencia Provincial de Navarra en 2011[192] (aún los últimos niveles de VA eran meros prototipos) enjuiciaba el caso de un hombre que, bajo la influencia del alcohol, fue encontrado en su coche con el motor y las luces encendidos, pero varado en medio de uno de los carriles de la calzada. Estaba acusado de un delito de conducción de un vehículo a motor bajo la influencia del alcohol. Lo curioso es que su Fundamento Tercero afirma que el conductor era culpable porque, al verificar la IA del coche, comprendieron que no alcanzaba una automatización que le permitiera circular por su cuenta sin intervención humana. Peris interpreta que el Tribunal habría inaplicado el tipo si la IA hubiera conducido en vez del hombre[193]. Esto suscita preguntas en cuanto a la posibilidad de que la intervención de una IA suponga la ruptura del vínculo de responsabilidad con respecto a su pasajero (o conductor).

Pero entonces nos topamos con otra vertiente. Una sentencia americana dilucidaba si un conductor podía ser responsable por exceder el límite de velocidad debido a un fallo de funcionamiento del sistema de velocidad de crucero si no existía voluntad de delinquir. El Tribunal de Apelación de Kansas dictaminó que un conductor no puede eludir las obligaciones de conducción (y la responsabilidad por sus efectos) que normalmente llevaría a cabo él mismo delegando a sistemas informáticos o mecánicos el control del vehículo y la responsabilidad en su lugar[194]. Se nos plantea la pregunta contraria: ¿debe un usuario responder por la actuación dirimida por el algoritmo de un VA? En caso afirmativo, ¿será responsable aun con ocasión de una infracción proveniente de un fallo de funcionamiento del sistema informático del VA?

Actualmente, la automatización de VA solo integra los niveles 1 al 3 del Código Ético Alemán: proporcionan asistencia en la conducción que se manifiesta en control de movimiento lateral y longitudinal, velocidad de crucero, trayectoria sin saltarse la línea de carril…[195], siempre necesitando un ser humano atento y/o a cargo. Siendo así que la responsabilidad recae sin duda sobre el conductor del vehículo, si los riesgos derivan de su gestión de trayectoria y maniobras, o sobre el fabricante, si se da la ocasión de que el perjuicio a un determinado bien jurídico proviene de un defecto en el producto[196]. El nivel 4 de automatización sigue suponiendo una línea difusa; pese a todo, la mayoría parece aceptar que el pasajero del VA sigue siendo responsable de prestar atención a la conducción e intervenir en caso de peligro[197].

Es en el nivel 5 donde se evidencia la necesidad de establecer un marco jurídico que atribuya la responsabilidad conforme a los principios de nuestro ordenamiento[198]. Mas no ya desde la situación crítica de tensión que sufre un sujeto ante una situación de necesidad, sino

[192] Sentencia de la Audiencia Provincial de Navarra 1498/2011 (Sala de lo Penal, Sección 3ª), de 15 de diciembre de 2011 (recurso 17/2011).
[193] MORILLAS FERNÁNDEZ, D.L., "Implicaciones de la inteligencia artificial en el ámbito del Derecho Penal", en Peris Riera, J.M. & Massaro, A. (Dir.): *Derecho Penal, Inteligencia Artificial y Neurociencias*. Roma Tre-Press, Italia, 2023, p. 86.
[194] FEILER, *op. cit.*, p. 27.
[195] SUÁREZ, *op. cit.*, p. 12.
[196] CONSEJO DE EUROPA, *op. cit.*, p. 7.
[197] LAW COMMISSION & SCOTTISH LAW COMMISSION, *op. cit.*, p. 3.
[198] CONSEJO DE EUROPA, *op. cit.*, p. 7.

reconducido a la nueva casuística[199] que supone un atentado contra un bien jurídico causado por el algoritmo de necesidad de un VA diseñado desde un momento inconexo anterior y creado de forma incierta (no sabe si el VA se enfrentará o no a esa situación de necesidad)[200].

El ordenamiento penal español (y muchos otros Estados de Derecho liberales[201]) basa la imposición de penas y medidas de seguridad en cinco fines: retribución, prevención general positiva y negativa, prevención especial (evitar que el delincuente vuelva a hacerlo, por un tiempo) y rehabilitación[202]. Más concretamente, no serán nuevos para el lector los requisitos para hallar a alguien responsable. Primero deben cubrir una acción, conducta humana externa manifestada por acción u omisión[203], típica tanto desde una relación de causalidad e imputación objetiva[204] como integrada por un dolo de cualquier tipo o una imprudencia[205]. Esta acción típica será antijurídica si no concurre una causa de justificación (como el estado de necesidad) y culpable si tiene capacidad para comprender la acción y su ilegalidad y comportarse conforme a la norma, siendo exigible para cualquier ciudadano medio que la cumpla (y para algunos, también punible, analizando si es más útil imponer la pena o no hacerlo[206]).

Esta es la preocupación que surge cuando introducimos algoritmos de necesidad de VA: crear un modelo que cumpla en la medida de lo posible con los fines de la pena descritos y al mismo tiempo evite un vacío de responsabilidad que no sea capaz de atribuir la conducta a ningún sujeto por no subsumirse en una acción típica, antijurídica y culpable y cree una sensación de injusticia, indefensión y silencio por parte la Administración de Justicia[207].

II. La IA como un sujeto de Derecho responsable penalmente

Simplemente como un breve apunte, debemos tener en cuenta que, si bien el art. 10 CP describe el primer elemento del delito como una acción sin especificar a su autor, no se concibe separada de la noción de acción ejecutada por un ser humano[208]. Es decir, la actuación de una IA no puede equipararse a la acción humana y, por tanto, no procede atribuirle responsabilidad a la IA por sus acciones dañinas.

En primer lugar, por mucho que la IA lleve a cabo procesos de elaborar respuestas sofisticados e incluso indescifrables, parte de unos parámetros iniciales y unas metas dispuestos por el programador. Dicho de otro modo, su actuación obedece a estas pautas y no a su propia autonomía (aunque pueda tener un amplio margen de medios para alcanzar ese fin)[209].

---

[199] NYHOLM, S., "The ethics of crashes with self-driving cars: A roadmap, II", en *Philosophy Compass*, Vol. 13, Nº 7, 2018, p. 2.
[200] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 240.
[201] FEILER, *op. cit.*, p. 19.
[202] MARTÍNEZ et al., *op. cit.*, p. 42.
[203] FEILER, *op. cit.*, p. 20.
[204] MARTÍNEZ et al., *op. cit.*, pp. 99 y 104.
[205] FEILER, *op. cit.*, p. 20. Feiler hace referencia al modelo americano de mens rea, pero se corresponde con el modelo español: la acción intencional es un dolo de primer grado, el dolo eventual equivale a la acción temeraria sin una confianza razonable en que pueda evitar el riesgo y la negligencia se liga a la imprudencia (faltando solo el dolo de consecuencias necesarias o de segundo grado).
[206] MARTÍNEZ et al., *op. cit.*, pp. 81 (antijuridicidad y culpabilidad) y 347 (punibilidad).
[207] NYHOLM, S., "The ethics of crashes with self-driving cars: A roadmap, II", en *Philosophy Compass*, Vol. 13, Nº 7, 2018, p. 4.
[208] MARTÍNEZ et al., *op. cit.*, p. 87.
[209] MORILLAS, *op. cit.*, p. 74.

Desde la perspectiva jurídica, las IA no alcanzan la posición moral única del ser humano sobre la que se edifica la responsabilidad penal al no alcanzar una sensibilidad (capacidad de reaccionar manifestado en un dolor físico, emocional o de otra índole), una intención (dolosa o imprudente) a la hora de ejecutar una acción[210], pero especialmente una sapiencia que hace a la persona capaz de ser consciente sobre los actos que lleva a cabo y comprender el alcance y gravedad de las consecuencias que siguen a los mismos (impidiendo su culpabilidad)[211].

Por último, una respuesta penal dirigida a una IA no tendría más que un efecto retributivo para la víctima, pero en realidad no podría cumplir los fines de prevención general ni rehabilitación sobre el VA: la IA no se ve disuadida o intimidada por la norma porque no es capaz de percibir estas sensaciones ni comprender el valor de la norma y tampoco podrá corregirse por sí misma en cuanto a su motivación para evitar la reincidencia porque carece de ella y solo podrá ver alterado su algoritmo manualmente por un humano[212].

Estas son las razones que motivan que debamos rechazar la posibilidad que el Consejo de Europa contempla de hacer realidad una personalidad o subjetividad (y, por lo tanto, posible responsabilidad penal) de las IA[213] que resuelva todo daño causado por estos algoritmos, en especial los que incorporarán los VA. De modo que no tendría sentido responsabilizar al VA o su algoritmo[214] y la cuestión vuelve a recaer en una responsabilidad meramente humana.

III. El riesgo permitido

1. Los VA y sus algoritmos de necesidad como un riesgo permitido

Como ya hemos anunciado anteriormente, al seleccionar el método deontológico como la técnica adecuada para diseñar el algoritmo de necesidad, éste responde a los fundamentos ético-jurídicos de la norma y la sociedad[215]. Siendo así que coincidimos con la postura de algunos autores que afirman que el algoritmo de los VA, al integrarse con los principios sociales y jurídicos, debería quedar comprendido en el concepto de riesgo permitido[216]. Dicho esto, considerar una conducta como un riesgo permitido debe cumplir dos requisitos, de acuerdo con Jakobs: unas consecuencias positivas de mayor calado que las negativas y una indeterminación de las potenciales víctimas del riesgo que impone[217].

El primer requisito consiste en que, al acordar social y normativamente permitir un riesgo desde una perspectiva *ex ante*[218], amplíe el ámbito de libertad de la generalidad de la

[210] Íbidem, 76.
[211] GRANDI, *op. cit.*, p. 27.
[212] MORILLAS, *op. cit.*, p. 76.
[213] CONSEJO DE EUROPA, *op. cit.*, p. 8.
[214] NYHOLM, S., "The ethics of crashes with self-driving cars: A roadmap, II", en *Philosophy Compass*, Vol. 13, Nº 7, 2018, p. 3.
[215] GOODALL, N.J., "Away from Trolley Problems and Toward Risk Management", en *Applied Artificial Intelligence*, Vol. 30, Nº. 8, 2016, p. 819.
[216] GRANDI, *op. cit.*, p. 26.
[217] JAKOBS, G., *Derecho penal. Parte general: Fundamentos y teoría de la imputación*, Marcial Pons, 1997, p. 244.
[218] Íbidem, 250. Aquí Jakobs hace referencia con la posición *ex ante* a que no se trata ya de un conductor que toma una decisión de girar el volante hacia aquel que sufrirá el perjuicio, sino que es el fabricante o programador, el usuario en el caso de una configuración ética personalizable o el experto si se trata de un algoritmo universal definido institucionalmente quien pondera objetivamente la situación de necesidad y los bienes jurídicos vulnerables y sugiere un cauce de acción aceptado por la sociedad y el ordenamiento.

ciudadanía[219]; es decir, que este traiga consigo unos beneficios claramente superiores a las limitaciones que produce[220]. En el caso de las tecnologías es evidente que todas suponen nuevos riesgos para la sociedad, pero la mayoría se consideran adecuadas porque los efectos positivos de conexión mediante los teléfonos, de movilidad en los vehículos o de seguridad por la videovigilancia (entre muchos otros) superan los efectos nocivos que estos pueden producir[221]. En el caso de los VA, se presupone que reducirán el número de siniestros viales, razón suficiente para permitirlos (cuando cumplan ciertos estándares) a pesar de sus peligros[222]. Lo mismo podría predicarse de los algoritmos de necesidad, ya que son un accesorio funcional de estos vehículos que reduce al mínimo los efectos negativos de los VA que se ponen en la balanza al ponderar si los beneficios son mayores a los riesgos[223] y además coincide en su juicio con el razonamiento que ofrece nuestro ordenamiento.

La indeterminación de las víctimas potenciales supone otro requisito que Vila ya adelantaba como un punto a favor de legitimar las soluciones del algoritmo de necesidad, ya que aporta soluciones genéricas e independientes de los sujetos concretos a confrontar[224]. Esto implica que el algoritmo bien podría favorecer o perjudicar a la misma persona según la posición en la que se encuentre ante la situación de necesidad: puede que la solución correcta sea que el VA atropelle al peatón negligente que cruza la carretera sin mirar, pero se decide conforme a los axiomas del algoritmo, no en virtud de quiénes sean los actores[225]. De esta manera, no hay discriminación hacia ningún sujeto gravándolo con una carga adicional injusta, sino que, de haber sido el pasajero del VA ante la misma situación, el algoritmo habría optado de igual modo por atropellar al peatón negligente (ahora una persona distinta)[226].

Cumpliendo ambos requisitos, vemos que los VA y sus algoritmos de necesidad introducen riesgos permitidos y, por ende, no traen una consecuencia penal para sus programadores, fabricantes o usuarios cuando insertan este riesgo en la sociedad.

2. La cláusula de cierre: seguros obligatorios para los VA

El sistema de riesgo permitido permite descartar una responsabilidad penal derivada del algoritmo para los pasajeros, programadores o fabricantes de VA, pero no impide asumir en el plano civil un seguro obligatorio para los usuarios de VA que compense los daños producidos a víctimas de accidentes[227] de igual modo que rige el seguro de responsabilidad civil obligatorio en los vehículos a motor[228]. Nyholm argumenta que, al introducir un riesgo en la sociedad como

[219] Íbidem, 243.
[220] GRANDI, *op. cit.*, p. 5.
[221] SUÁREZ, *op. cit.*, p. 12.
[222] Íbidem, 16.
[223] GRANDI, *op. cit.*, p. 20.
[224] COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, p. 240.
[225] Ídem (nota al pie de página nº 23).
[226] JAKOBS, *op. cit.*, p. 244.
[227] CONSEJO DE EUROPA, *op. cit.*, p. 8.
[228] GÓMEZ LIGÜERRE, C., CAMPAMÁ, Mª.C. & SÁNCHEZ ÁLVAREZ, V., "Ciclomotores y seguro. Daños personales causados por ciclomotores y seguro obligatorio", en *Revista para el Análisis del Derecho InDret*, Nº. 3, 2003, p. 6.

es la utilización de VA, los usuarios de estos vehículos como colectivo deben tomar una posición de aseguramiento a la sociedad por estos riesgos[229].

De hecho, Jakobs ya anunciaba que, aunque la circulación de vehículos a motor sea un riesgo permitido, no evita que se le califique de riesgo especial por incrementar el peligro sobre lo cotidiano. Así, cuando una persona hubiera adoptado toda medida necesaria para prevenirse del riesgo que imponen los vehículos a motor sobre la población (o los VA en nuestro trabajo) y aún así fuera víctima afectada concretamente por el mismo, surge un deber de asumir un seguro obligatorio para compensar el peligro[230].

Cuestión aparte son los beneficios que este sistema reportaría. El primero, que no supondría mucha diferencia en estructura con respecto al sistema actual de responsabilidad civil de los vehículos a motor, esperando determinar unas cuantías razonables de modo que convenzan tanto al asegurado como a la aseguradora del efecto positivo que suponen al compararlos con los daños o los ingresos previstos, respectivamente[231]. De este modo, se cumple la prevención general en ambos sentidos por el coste del seguro, pero sin suponer una carga excesiva. Trae otro beneficio al cubrir ese vacío de responsabilidad que preocupa a algunos autores[232] al asegurar cierto grado de cobertura y compensación a las víctimas de estos casos[233] (que, por cierto, se prevén muchas menos gracias a la implantación de estos VA).

Como colofón, al imponer el seguro obligatorio en VA se subvierte una objeción existente contra la responsabilidad objetiva por daños de vehículos a motor: a saber, que la responsabilidad subjetiva o por culpa es mejor incentivo para ser atento, previsor y cuidadoso en la conducción, puesto que la responsabilidad objetiva actúa con independencia de las medidas de precaución que se hayan dispuesto[234]. Pese a que esta perspectiva tiene sus fallas en cuanto a vehículos a motor se refiere, cuando pasamos a referimos a VA pierde todo su sentido: los detractores de la responsabilidad objetiva podrán estar tranquilos al saber que los VA, al contrario que los conductores de carne y hueso, siempre adoptarán todas las medidas adecuadas y necesarias para evitar los accidentes, ya que ese es su fin primordial y la razón por la que los consideramos tan beneficiosos para la sociedad como para implementarlos a la vida cotidiana (cuando estén suficientemente testeados y desarrollados).

IV. Responsabilidad penal de los usuarios, programadores o fabricantes de VA

Habiendo incluido el modelo deontológico del algoritmo de necesidad de los VA dentro del concepto de riesgo permitido[235], este provee un estándar bajo el cual la fabricación, diseño, comercialización y uso de este producto quedan inmunes de responsabilidad penal al no cumplir los requisitos de la imputación objetiva[236]. Ahora bien, hay ciertas situaciones en las que la

[229] NYHOLM, S., "The ethics of crashes with self-driving cars: A roadmap, II", en *Philosophy Compass*, Vol. 13, Nº 7, 2018, p. 3.
[230] JAKOBS, *op. cit.*, p. 984.
[231] GÓMEZ et al., *op. cit.*, p. 15.
[232] NYHOLM, S., "The ethics of crashes with self-driving cars: A roadmap, I", en *Philosophy Compass*, Vol. 13, Nº 7, 2018, p. 4.
[233] GÓMEZ et al., *op. cit.*, p. 10.
[234] Ídem.
[235] GRANDI, *op. cit.*, p. 26.
[236] FEILER, *op. cit.*, p. 24.

actuación de fabricantes, programadores o usuarios se exceden de estos estándares y no pueden protegerse bajo la inmunidad del riesgo permitido.

1. El usuario del VA como responsable penal

Ya contemplamos esta posición en cuanto a la configuración ética personalizable del algoritmo de necesidad[237]. Como probamos entonces, no podemos delegar al usuario la elección ética del VA, debiendo instituirse un modelo deontológico. Pero, aún si lo aceptáramos, sería problemático en cuanto a responsabilidad se refiere. Componen el riesgo permitido la diferencia de efectos beneficiosa y la indeterminación de la víctima[238]. Para empezar, los riesgos introducidos por los VA que usaran este algoritmo (especialmente en su vertiente egoísta) serían mayores a sus beneficios, como dicta el ya analizado dilema del prisionero, perjudicial para la sociedad al no minimizar daños[239]. En cuanto a la indeterminación de la víctima, si la teoría del algoritmo personalizable asemeja la elección ética del algoritmo de necesidad con la ejecución personal del delito, cabe la posibilidad de que su conducta se acoja al estado de necesidad como causa justificante. Pero ya hemos visto que es difícil equiparar estas dos situaciones porque el usuario no se encuentra en el momento crítico, de modo que la causa de justificación tendría un difícil ajuste[240]. De hecho, Vila lo compara con una tentativa de delito a distancia desde el momento en que el usuario ha tomado una decisión de puesta en peligro del bien y pierde el control sobre el medio técnico que la transforma en un resultado según su plan[241]. Teniendo esto en cuenta y que el dilema del prisionero supone que casi todos elegirían un algoritmo egoísta, la consecuencia evidente sería una responsabilidad penal generalizada de casi todos los usuarios de VA que tuvieran accidentes. Esto echaría abajo el proyecto de normalizar los VA por su propio peso, el concepto de riesgo permitido no lo abarcaría al incrementar injustificadamente el riesgo a la sociedad en general: más accidentes de tráfico que los que promete evitar y menos compradores por el temor a ser responsables penales[242].

Asentada esta argumentación, volvemos al algoritmo deontológico. Evidentemente, no podemos responsabilizar al usuario por no haber estado suficientemente atento al volante[243], pues evitarle esa preocupación es precisamente la función del VA y no tendrá margen de tiempo suficiente[244], ni por la actuación de su VA según el algoritmo de necesidad, porque el usuario no gobierna el VA[245] ni toma la decisión que se lleva a cabo[246] y queda cubierto por el riesgo permitido y la argumentación basada en las causas de justificación.

No obstante, como el algoritmo deviene tras una discusión doctrinal y profesional y se implanta de manera institucional y generalizada, se nos presenta un deber para el usuario que

[237] Véase en este trabajo, en la sección *II. Un algoritmo para todos o un algoritmo personalizable* del capítulo *Teoría del algoritmo de necesidad*, en concreto el apartado *1. Configuración ética personalizable* (9-12).
[238] JAKOBS, *op. cit.*, p. 244.
[239] Véanse las páginas 13 y 14 de este trabajo.
[240] COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, p. 1137.
[241] Íbidem, 1139.
[242] GRANDI, *op. cit.*, p. 25.
[243] MORILLAS, *op. cit.*, p. 83.
[244] NYHOLM, S., "The ethics of crashes with self-driving cars: A roadmap, I", en *Philosophy Compass*, Vol. 13, Nº 7, 2018, p. 3.
[245] LAW COMMISSION & SCOTTISH LAW COMMISSION, *op. cit.*, 111.
[246] FEILER, *op. cit.*, p. 21.

sí debería estar protegido penalmente: el deber de mantener el algoritmo intacto, salvo que deba actualizarse el software para que las respuestas que tome el algoritmo sean conforme a la evolución interpretativa que le vaya dando la doctrina (si es que hay alteraciones)[247]. En caso contrario, la respuesta que diera el algoritmo de necesidad y estuviera desfasada o difiera de la nueva norma deberá tomarse como no abarcada por el riesgo permitido y entonces sí estaría penada[248]. Del mismo modo se penaría a cualquiera que saboteara el algoritmo (ya sea el propio o el de otra persona sin que esta lo supiera o contra su voluntad) a su gusto y se produjera un accidente en que el algoritmo no siguiera las normas derivadas del algoritmo deontológico[249].

2. Los fabricantes y programadores de VA como responsables penales

Extender la voluntad del fabricante del algoritmo de necesidad de los VA hasta el resultado concreto de causación de un riesgo y resolución de una situación de necesidad con afectación a un bien jurídico es demasiado extremo[250], especialmente considerando que tanto el VA como el algoritmo de necesidad se engloban dentro del riesgo permitido y su fabricación se realiza conforme a unas pautas normativas[251]. Lo mismo cabe decir del programador, que actúa conforme al fabricante y programa el algoritmo en abstracto para tomar decisiones de necesidad, no sobre el caso concreto[252]. Por consiguiente, si se ha determinado que el modelo deontológico del algoritmo de necesidad es el adecuado y permitido según la norma, surge un deber para los fabricantes y programadores de proyectar el algoritmo ajustado a este modelo, del mismo modo que sus productos mantienen los estándares de calidad o seguridad[253]. Así, se incluiría dentro del riesgo permitido. La cuestión es cuál sería la responsabilidad penal de estos individuos en caso de sobrepasar el riesgo permitido por diseñar o fabricar un algoritmo que no cumpla el patrón legal y convertirlo así en un riesgo excesivo[254].

Un supuesto de responsabilidad penal (*lege ferenda*) podría derivar de la fabricación o programación de algoritmos de necesidad que se aparten de la fórmula legalmente establecida. En primera instancia, si el fabricante o programador alterara el algoritmo a voluntad (causando un riesgo mayor al permitido por la sociedad[255]), porque se trata de una conducta dolosa frontalmente opuesta a lo que la norma dispondrá acerca de los principios que regirán los algoritmos de necesidad de VA[256]. Aparte de la responsabilidad civil derivada de la regulación en defensa del consumidor, este producto excedería el riesgo inherente y habitual permitido hacia la salud de los consumidores[257]. De ser así y sufrir un usuario un accidente en el que el

[247] GRANDI, *op. cit.*, p. 20.
[248] LAW COMMISSION & SCOTTISH LAW COMMISSION, *op. cit.*, p. 40.
[249] COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, p. 1142.
[250] FEILER, *op. cit.*, p. 23.
[251] GRANDI, *op. cit.*, p. 20.
[252] FEILER, *op. cit.*, p. 22.
[253] GRANDI, *op. cit.*, p. 23.
[254] JAKOBS, *op. cit.*, p. 245.
[255] WENDEL, *op. cit.*, p. 146.
[256] COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, p. 1142.
[257] JUANATEY, C., "Responsabilidad penal omisiva del fabricante o productor por los daños a la salud derivados de productos introducidos correctamente en el mercado", en *Anuario de Derecho Penal y Ciencias Penales*, Nº. 57, 2004, p. 69.

VA no siguió las reglas establecidas para el algoritmo generalizado, se ha vulnerado lo que el Derecho ha establecido como la mejor forma de distribuir los riesgos que invocan los VA[258] y ha creado un resultado lesivo concreto que se liga a esa conducta típica. Podrá el legislador entonces decidir si es más conveniente un tipo que condene la desviación solo en caso de producirse un resultado lesivo concreto[259] o un tipo de peligro que abarque el mero apartamiento del diseño deontológico que pueda agravarse en caso de producir un resultado lesivo concreto.

Caso distinto sería que, tras seguir todas las instrucciones de producción y programación y superar las pruebas conforme al modelo estándar, surja *a posteriori* un defecto en el producto por el que el VA no siguiera el algoritmo predefinido, sino que existiera una alteración. Aquí la doctrina se divide en argumentaciones muy amplias que no podemos desarrollar por completo, pero partimos de los arts. 12 y 13 del Texto Refundido de la Ley General de Consumidores y Usuarios: todo empresario tiene unos deberes específicos de información sobre los riesgos del producto y retirada de productos que no cumplan las condiciones exigidas o supongan un riesgo previsible e intolerable contra la salud o seguridad de las personas. A partir de estos artículos, varios autores asumen que la distribución de una mercancía con un riesgo (aunque permitido) cumpliendo todos los requisitos entraña una posición de garante por un actuar precedente (injerencia) que impone un deber de vigilancia razonable a la empresa atendiendo a posibles nuevos riesgos o defectos del bien[260]. Si surge un riesgo o se descontrola uno existente que atente contra la salud y seguridad de las personas, la empresa tiene la obligación de informar a sus usuarios y, de ser excesivo o insalvable, retirar estos productos del mercado. Dado el caso de que la empresa no informe al usuario o haga lo posible por retirarlo, cabe afirmar una equivalencia entre esta omisión y una acción por su conocimiento privilegiado y su dominio sobre el foco del riesgo al no avisar ni evitarlo[261]. Si resulta en el caso concreto que el algoritmo resuelve de forma dispareja con el modelo deontológico y afectando a un bien jurídico, la solución habitual será imputar la lesión a la empresa por imprudencia[262].

De este modo, deber hacerse responsable al productor o programador que dolosamente produjera y distribuyera un algoritmo de necesidad distinto al autorizado. Mas también al productor (aunque exista discusión doctrinal) si cumplió todos los requisitos al diseñarlo y fabricarlo, pero después surgiera un defecto que alterase el algoritmo con riesgo a la integridad de los usuarios y la empresa no informara a los consumidores e intentara por todos los medios retirar el producto del mercado.

[258] COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, p. 1139.
[259] WENDEL, *op. cit.*, p. 144.
[260] JUANATEY, *op. cit.*, p. 65.
[261] Íbidem, 74.
[262] Íbidem, 75.

CONCLUSIONES

A lo largo de todo el trabajo hemos ido desgranando un modelo de algoritmo de necesidad de VA congruente con el ordenamiento jurídico y la doctrina que manejamos. Un sistema basado en el célebre dilema del tranvía para extraer unos principios, pero que en la realidad implementará un sistema probabilístico (adquirido del modelo de gestión del riesgo) para determinar el valor esperado del riesgo en cada alternativa que el VA pueda plantearse.

Hemos demostrado que este algoritmo ético no puede dejarse en manos de cada individuo porque, además de requerir de unas pautas mínimas más de lo que aparenta, desembocaría en un dilema del prisionero que no minimizaría los riesgos que origina, no contemplaría conceptos jurídicos esenciales como la imponderabilidad de la vida humana o la provocación de una situación de necesidad y ahuyentaría a la clientela debido a su sistema de responsabilidad por el mero uso del algoritmo personalizable. En su lugar, una configuración ética universal (o nacional) es la respuesta: unos principios que satisfacen las premisas de nuestro ordenamiento, que ya de por sí cuenta con un consenso social (desde un modelo contractualista) que dejará conformes a empresas de VA, usuarios y demás personas que pudieran encontrarse en una situación de necesidad con un VA. Asimismo, vela por la respuesta que mejor salvaguarde el interés general y minimice la materialización de los riesgos (salvando así el dilema del prisionero).

Nos adentramos en este algoritmo ético para todos y comprobamos que un sistema estrictamente utilitarista, pese a su sencillez, de nuevo no incorpora valores normativos como el merecimiento o la imponderabilidad de la vida humana e infla el principio de solidaridad, instrumentalizando al individuo y aplastando su dignidad individual (el principio de autonomía). Conjuntamente, desincentiva la autoprotección de los usuarios de la vía al designarlos como blancos más probables porque protegerse reduce el valor esperado de riesgo al chocar contra ellos. Razones suficientes para desechar el utilitarismo feroz en aras de una filosofía deontológica que realiza un análisis exhaustivo y completo de todos los actores y sus posiciones ante la situación de necesidad y fundamenta unos principios consensuados y, nuevamente, coincidentes con aquellos que rigen nuestro ordenamiento.

De todas maneras, antes de examinar los principios de actuación del algoritmo, debemos preguntarnos cuándo debe tomar acción este principio concretamente[263]. Tras conceptualizar las tres capas de seguridad de que podría disponer el VA, comprendemos que en tanto no se vean vulnerada la capa de comodidad, el VA conducirá con plena libertad. Cuando algún sujeto u objeto se introduzca en ella, la labor del VA se limitará a la conducción preventiva. Y en el momento en que algún usuario acceda a la capa del deber de cuidado, esto servirá de advertencia al VA de una posible violación inminente de la capa de colisión y podrá establecer maniobras de evasión. Finalmente, de penetrar en la última capa, la de colisión, este es el instante en el que el algoritmo de necesidad debe intervenir según la información de que disponga y participan los principios que vemos a continuación.

Este algoritmo de necesidad deontológico distinguirá tres situaciones de necesidad distintas cuyos factores difieren y, basculando entre el principio de autonomía y el de solidaridad, dan lugar a soluciones también diferentes.

[263] En el desarrollo del trabajo, este apartado se expone después de esclarecer los principios del algoritmo de necesidad porque serían importantes para comprender la definición y funcionamiento de estas capas de seguridad.

En primer lugar, en caso de sufrir el pasajero del VA una agresión ilegítima por otro ciudadano que no hubiera sido provocada por el VA o su pasajero, el algoritmo podrá reaccionar (si no puede evitarlo de otra manera sin agredir a terceros ajenos al conflicto) contra el agresor utilizando el medio a su alcance que sea menos perjudicial *ex ante*.

Un segundo grupo lo protagonizan situaciones de peligro inminente a los bienes jurídicos del pasajero del VA que no han sido provocadas por éste y ante las cuales tampoco tiene deber de sacrificarse por su cargo u oficio (cuestión que puede determinarse con mecanismos especiales de identificación del pasajero), pudiendo desplazar el riesgo hacia un tercero ajeno a la situación de peligro si no existe otra salida. El requisito que se suele asociar es la proporcionalidad: que el mal causado por el VA no supere en gravedad al mal que amenaza al necesitado. Sin embargo, hemos probado que, cuando tratamos de algoritmos de necesidad predeterminados, ya no cabe la comparación de vida contra vida a través de un estado de necesidad (incluso exculpante) debido a que la vida humana es imponderable y se desprovee del contexto crítico de tensión e inexigibilidad que le da sentido. De tal forma que, si el mal que causaría al tercero es menor a aquel que amenaza al necesitado, estaría justificado que el algoritmo optase por atentar contra el tercero; pero, si se enfrentan la vida del pasajero o pasajeros frente a la vida del tercero o terceros ajenos al conflicto, los pasajeros tendrán el deber de soportar la desgracia. Añadimos, para mayor claridad sobre el tercero ajeno al conflicto, la figura de los espacios seguros. Son espacios seguros aquellos en los que el tercero no debe soportar la desgracia, salvo que el mal que los amenace sea menor que aquel que se dirija contra el necesitado: una posición preferente debido a que el sujeto actúa de forma lícita y no crea ningún riesgo para los demás usuarios.

El último criterio atiende a la posición de incumbencia preferente que exige unos deberes de tolerancia más intensos para ciertos sujetos, pudiendo verse víctimas de una situación que, sin plena responsabilidad, pueda vinculársele en una medida superior a la estricta proporcionalidad, pero sin alcanzar las reglas de la legítima defensa. Vimos que este vínculo entre la fuente del peligro y el sujeto con deberes intensificados de solidaridad sería deseable que lo reforzaran los requisitos de la imputación objetiva. No obstante, dado que la jurisprudencia adopta mayoritariamente la mera relación de causalidad y que los VA no están aún preparados para discernir la previsibilidad de un riesgo por un sujeto (presupuesto de la imputación objetiva), consideramos adecuado que, por ahora, sea la ubicación fáctica del peligro la solución a estas situaciones excepcionales hasta la evolución suficiente de los VA.

Este último punto nos recuerda que todo grupo de normas es incompleto y que la técnica de los VA es aún prototípica. De acuerdo con estas aserciones, se proyecta una estrategia progresiva de evolución de las respuestas éticas del algoritmo que se perfecciona en consonancia con la mejora tecnológica con tres fases: una primera de ética racional (véase, el groso de este trabajo), una fase híbrida entre la ética racional y el método IA (entrenando a la IA para resolver por sí misma a partir de la observación de casos reales y de estudio) y una última fase plenamente basada en el método IA con su traducción al lenguaje natural para poder ser monitorizada por el ser humano.

Acto seguido, profundizamos en si son viables supuestos de responsabilidad penal relacionados con el algoritmo de necesidad de VA. En cualquier caso, hablamos de responsabilidad penal de seres humanos, puesto que las IA no son sujetos de Derecho y, por lo tanto, no pueden ser responsables penales en ningún caso. Apreciamos que, al establecerse un algoritmo de necesidad deontológico y para todos que produce más efectos beneficiosos que

nocivos y resuelve en abstracto sin discriminar por razón de los actores en conflicto, el riesgo permitido cubre la producción, programación y uso de VA y sus algoritmos éticos, siempre que se adhieran al modelo deontológico determinado por la norma. Como cláusula de cierre del sistema, se considera acertada la instauración de un seguro obligatorio de responsabilidad civil por el uso de VA, acorde a los riesgos que introduce en la sociedad.

Siguiendo este razonamiento, el legislador deberá establecer un tipo penal que haga responsable a aquel usuario (sea intencionalmente o por no actualizar el software en plazo), fabricante o programador que se aparte de la pauta establecida en la ley para los criterios del algoritmo. A su vez, nos posicionamos a favor de aquella doctrina que defiende la responsabilidad penal del fabricante o distribuidor que, aún cumpliendo todos los requisitos de diseño y fabricación, conoce de un defecto o alteración sobrevenido que altera el algoritmo con riesgo a la integridad de los usuarios y aun así no informa a los consumidores e intenta por todos los medios a su alcance lograr la retirada o suspensión del producto en el mercado, pudiendo ser responsable de las lesiones concretas que produzcan a cualquier consumidor a título imprudente.

BIBLIOGRAFÍA

BONNEFON, J.F., SHARIFF, A. & RAHWAN, I., "The social dilemma of autonomous vehicles", en *Science*, Vol. 352, Nº. 6293, 2016, pp. 1573-1576.

COCA VILA, I., "Entre la responsabilidad y la solidaridad. El estado de necesidad defensivo", en *Revista para el Análisis del Derecho InDret*, Nº. 1, 2011.

COCA VILA, I., "Coches autopilotados en situaciones de necesidad. Una aproximación desde la teoría de la justificación penal", en *Cuadernos de política criminal*, Nº. 122 (II), 2017, pp. 235-275.

COCA VILA, I., "La exculpación de hechos lesivos programados. Una primera reflexión a propósito de la configuración antijurídica de los algoritmos de necesidad de coches autónomos", en *Revista de Derecho Penal y Procesal Penal*, Nº. 6, 2018, pp. 1130-1143.

CONSEJO DE EUROPA, *Legal aspects of autonomous vehicles*, Comisión de Asuntos Legales y Derechos Humanos, Asamblea Parlamentaria, 2020, AS/Jur 20.

CONTISSA, G., LAGIOIA, F. & SARTOR, G., "The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law", en *Artificial intelligence and law*, 2017, Vol. 25, Nº 3, pp. 365–378.

D'AMATO, A., DANCEL, S., PILUTTI, J., TELLIS, L., FRASCAROLI, E. & GERDES, J.C., "Exceptional Driving Principles for Autonomous Vehicles", en *Journal of Law and Mobility*, Nº 2, 2022.

FEILER, J., "The Artificially Intelligent Trolley Problem: Understanding Our Criminal Law Gaps in a Robot Driven World", en *Hastings Science and Technology Law Journal*, Vol. 14, Nº. 1, 2023.

GOGOLL, J. & MÜLLER, J.F., "Autonomous Cars: In Favor of a Mandatory Ethics Setting", en *Science and engineering ethics*, Vol. 23, Nº. 3, 2017, pp. 681-700.

GÓMEZ LIGÜERRE, C., CAMPAMÁ, Mª.C. & SÁNCHEZ ÁLVAREZ, V., "Ciclomotores y seguro. Daños personales causados por ciclomotores y seguro obligatorio", en *Revista para el Análisis del Derecho InDret*, Nº. 3, 2003.

GOODALL, N.J., "Ethical decision making during automated vehicle crashes", en *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2424, Nº. 1, 2014, pp. 58-65.

GOODALL, N.J., "Machine Ethics and Automated Vehicles", en Meyer, G. & Beiker, S. (ed.): *Road Vehicle Automation*, Springer, 2014, pp. 93-102.

GOODALL, N.J., "Away from Trolley Problems and Toward Risk Management", en *Applied Artificial Intelligence*, Vol. 30, Nº. 8, 2016, pp. 810-821.

GRANDI, N.M., "Inteligencia artificial al volante. Una mirada sobre la atribución de Responsabilidad Penal por los resultados lesivos generados por los vehículos autónomos", en *Revista Argentina de Derecho Penal y Procesal Penal*, Nº. 27, 2020.

HOLSTEIN, T. & DODIG-CRNKOVIC, G., "Avoiding the Intrinsic Unfairness of the Trolley Problem", presentada en *2018 IEEE/ACM International Workshop on Software Fairnes (FairWare)*, Sweden, 2018, pp. 32-37.

JAKOBS, G., *Derecho penal. Parte general: Fundamentos y teoría de la imputación*, Marcial Pons, 1997.

JUANATEY, C., "Responsabilidad penal omisiva del fabricante o productor por los daños a la salud derivados de productos introducidos correctamente en el mercado", en *Anuario de Derecho Penal y Ciencias Penales*, Nº. 57, 2004, pp. 54-75.

LAW COMMISSION & SCOTTISH LAW COMMISSION, "Automated Vehicles: Analysis of Responses to the Preliminary Consultation Paper", en *The National Archives: Open Government Licence*, Reino Unido, 2019.

LAWLOR, R., "The Ethics of Automated Vehicles: Why Self-driving Cars Should not Swerve in Dilemma Cases", en *Res Publica*, Vol. 28, Nº. 1, 2022, pp. 193-216.

MARTÍNEZ ESCAMILLA, M., MARTÍN LORENZO, Mª. & MARISCAL DE GANTE, M.V., *Derecho penal: Introducción. Teoría jurídica del delito. Materiales para su docencia y aprendizaje*, Universidad Complutense de Madrid, 2012.

MORILLAS FERNÁNDEZ, D.L., "Implicaciones de la inteligencia artificial en el ámbito del Derecho Penal", en Peris Riera, J.M. & Massaro, A. (Dir.): *Derecho Penal, Inteligencia Artificial y Neurociencias*. Roma Tre-Press, Italia, 2023, pp. 59-91.

MUÑOZ RUIZ, J., *El delito de conducción temeraria: análisis dogmático y jurisprudencial*, Dykinson, 2014, pp. 307-338.

NYHOLM, S., "The ethics of crashes with self-driving cars: A roadmap, I", en *Philosophy Compass*, Vol. 13, Nº 7, 2018.

NYHOLM, S., "The ethics of crashes with self-driving cars: A roadmap, II", en *Philosophy Compass*, Vol. 13, Nº 7, 2018.

ORGANIZACIÓN MUNDIAL DE LA SALUD (OMS), *Global status report on road safety 2023*, 2023, Génova.

PAWLIK, M., "Solidaridad como categoría de legitimación jurídico penal: el ejemplo del estado de necesidad agresivo justificante", en *Revista de Estudios de la Justicia*, Nº. 26, 2017, pp. 222-247.

RAWLS, J., *Teoría de la justicia*, Fondo de Cultura Económica de España, 1999.

REAL ACADEMIA ESPAÑOLA, "Estado de necesidad defensivo", en *Diccionario panhispánico del español jurídico (DPEJ)*. Recuperado en 31 de febrero de 2024, de https://dpej.rae.es/lema/estado-de-necesidad-defensivo

Sentencia de la Audiencia Provincial de Navarra 1498/2011 (Sala de lo Penal, Sección 3ª), de 15 de diciembre de 2011 (recurso 17/2011).

SUÁREZ, M.F., "Inteligencia Artificial y Derecho Penal: El Dilema del Tranvía. Cuarta Revolución Industrial. Ética del Algoritmo. IA en vehículos. Causas de Justificación", en *Revista Pensamiento Penal*, Nº. 445, 2022.

WENDEL, W.B., "Economic Rationality and Ethical Values in Design-Defect Analysis: The Trolley Problem and Autonomous Vehicles", en *California Western Law Review*, Vol. 129, Nº. 56, 2018, pp. 129-163.

WILENMANN VON BERNATH, J., “El sistema de derechos de necesidad y defensa en el Derecho penal”, en *Revista para el Análisis del Derecho InDret*, Nº. 3, 2014.