# LA ÉTICA DE LOS ALGORITMOS
## *THE ETHICS OF ALGORITHMS*

*Fernando H. Llano Alonso*
*Catedrático de Filosofía del Derecho*
*Universidad de Sevilla*

RESUMEN

El presente trabajo analiza varios casos prácticos de ética aplicada a la IA en los que, ya sea por error y/o introducción de sesgos en el diseño de los algoritmos, o bien por la comisión de alguna imprudencia en el proceso de supervisión humana del sistema algorítmico, resultaron afectados derechos y libertades de las personas que fueron víctimas de ese fallo técnico y/o humano. Por otra parte, se estudian los tres modelos éticos que pueden servir de referencia a los diseñadores y programadores de algoritmos para que, mientras no sea del todo autónoma y general, de acuerdo con el cambio de paradigma que presumiblemente conllevará la singularidad tecnológica, la IA siga manteniéndose fiel al legado ético del humanismo ilustrado, antropocéntrico y antropogénico.

PALABRAS CLAVE

Utilitarismo. Ética deontológica, ética de la virtud, vehículo autónomo, supervisión humana.

ABSTRACT

This work analyses several practical cases of ethics applied to AI in which, either by error and/or introduction of biases in the design of algorithms, or by the commission of some imprudence in the process of human supervision of the algorithmic system, the rights and freedoms of the people who were victims of this technical and/or human failure were affected. On the other hand, we study the three ethical models that can serve as a reference for algorithm designers and programmers so that, as long as it is not fully autonomous and general, in accordance with the paradigm shift that technological singularity will presumably bring, AI remains faithful to the ethical legacy of enlightened, anthropocentric and anthropogenic humanism.

KEYWORDS

Utilitarianism, deontological ethics, virtue ethics, autonomous vehicle, human supervision.

DOI: https://doi.org/10.36151/TD.2024.106

# LA ÉTICA DE LOS ALGORITMOS

Fernando H. Llano Alonso
Catedrático de Filosofía del Derecho
Universidad de Sevilla

**Sumario:** 1. Introducción. 2. La supervisión humana de los algoritmos: ¿Efecto placebo o mecanismo eficaz de regulación? 3. La función correctora de los principios de la Ética de la IA en el diseño y desarrollo de los algoritmos. 4. El caso Herzberg. Crónica y análisis del primer error algorítmico fatal en la historia de la IA. 5. Un dilema moral y tres modelos éticos para los sistemas algorítmicos. 6. Conclusión. Notas. Bibliografía.

*«Una fracción de segundo es tiempo más que suficiente para que una computadora estudie concienzudamente todas las posibilidades. La decisión dependerá de las prioridades que haya fijado el software»*

(McEwan, 2019)

## 1. INTRODUCCIÓN

El 9 de noviembre de 2023 el diario *Financial Times* publicó un artículo sobre el caso Sarah Meredith una joven residente en Cambridge a la que le diagnosticaron fibrosis quística al nacer. Se trata de una enfermedad genética que a largo plazo causa daños graves en órganos vitales como los pulmones, el aparato digestivo o el hígado. Efectivamente, en el verano de 2021, recién cumplidos los 28 años, el hígado de Sarah empezó a fallar tras casi tres décadas de constantes tratamientos experimentales, hospitalizaciones y terapias para controlar su enfermedad crónica.

Sarah fue ingresada de urgencia en el hospital de Plymouth, donde los hepatólogos que la atendieron calcularon que a su hígado le quedaban aproximadamente dos años y le recomendaron un trasplante. Tras aceptar someterse a un trasplante, Sarah se inscribió en una lista nacional de pacientes que esperaban la donación de un hígado. Por su parte, el Servicio Nacional de Salud del Reino Unido (NHS, por sus siglas en inglés) le informó de que el promedio de espera era entonces de 68 días, si bien esta expectativa se veía ensombrecida con un dato inquietante: el hecho de que solo seis personas que padecían las mismas afecciones crónicas que Sarah habían sido incluidas en la lista para un trasplante de hígado en la historia del NHS.

Con el fin de reducir el número de pacientes que fallecían mientras esperaban un trasplante, en 2018 el NHS introdujo un algoritmo de asignación de trasplante hepático denominado National Liver Offering Scheme (NLOS, por sus siglas en inglés) que cruzaba datos de pacientes en lista de espera de trasplante con el total de hígados donados en toda Gran Bretaña para calcular la puntuación obtenida por cada paciente, estableciendo así un orden de prelación en la lista de espera, de modo que a quien obtuviera la puntuación más alta se le ofrecería el hígado.

Sin embargo, este algoritmo había sido cuestionado desde su puesta en marcha por la opacidad de su funcionamiento, la ausencia de un procedimiento de apelación y la inexistencia de un sistema de supervisión humana que permitiera corregir o anular la puntuación ante un hipotético error de cálculo del algoritmo. En efecto, cuando la familia de Sarah preguntó a varios hepatólogos del NHS Blood and Transplant si sabían cómo calculaba el algoritmo el Transplant Benefit Score (TBS, por sus siglas en inglés), es decir, cómo funcionaba el *software* a la hora de asignar la puntuación que le correspondía a cada paciente en espera de trasplante, ninguno pudo responder con seguridad, aunque se inclinaban a pensar que Sarah tendría más posibilidades de recibir el trasplante que la mayoría de pacientes debido a su edad.

Un buen día, la familia Meredith encontró en Internet un sitio web creado por Ewen Harrinson, profesor de cirugía y ciencia de datos en la Universidad de Edimburgo. El Dr. Harrinson, también cirujano práctico en trasplantes, había creado una versión sencilla y accesible del algoritmo NLOS que permitía entender cómo se calculaba el TBS. En esta especie calculadora *online* se introducían algunos datos del paciente como la edad, el sexo y algunas mediciones específicas de su química sanguínea, y, de este modo, se obtenía la puntuación que le correspondería al enfermo como probable beneficiario de un trasplante.

Tras introducir los datos de Sarah, la puntuación obtenida era inferior al número de puntos necesario para conseguir que se le trasplantara un hígado sano. El resultado de Sarah no mejoraba ni siquiera cuando se modificaban datos variables como el de su edad. La triste conclusión a la que llegaron Sarah y su familia después de probar el sistema NLOS para la asignación de órganos es que las posibilidades de que fuera seleccionada para un trasplante eran finalmente muy remotas.

En la comunidad de hepatólogos del Reino Unido fue creciendo la sensación de que el algoritmo presentaba un sesgo de edad que perjudicaba sobre todo a los pacientes más jóvenes, entendiendo por tales los menores de 45 años (ese era el caso de Sarah Meredith). Entre los expertos que habían analizado el algoritmo NLOS destacaban David Spiegelhalter (Universidad de Cambridge) y Palak Trivedi (Universidad de Birmingham), cuyos análisis revelaron que el algoritmo partía de una premisa errónea: generalizar e igualar la tasa de mortalidad de todos los pacientes en lista de espera, sin reparar en el estado de desarrollo de la enfermedad de cada paciente con independencia de su edad. De esta manera, el algoritmo priorizaba a los pacientes de mayor edad frente a los más jóvenes porque se presumía que, al tener una edad más avanzada, vivirían menos que los más jóvenes. Paradó-

jicamente, el algoritmo NLOS llegó a asignar una puntuación superior a muchos pacientes de la tercera edad cuyos órganos estaban en un estado aceptable en comparación con el de algunos pacientes más jóvenes cuya vida dependía de un trasplante urgente.

Por otra parte, el sistema no tenía en cuenta otros datos tan relevantes como los años de vida saludable perdidos por los pacientes más jóvenes, muchos de los cuales habían nacido ya enfermos o habían desarrollado la enfermedad durante la infancia, y tampoco contemplaba los resultados a largo plazo de los enfermos más jóvenes (cuyos hígados habían estado deteriorándose más tiempo que el de los pacientes de mayor edad, muchos de ellos enfermos hepáticos sobrevenidos debido a adicciones como el alcoholismo); en definitiva, no se tuvo en cuenta el progresivo empeoramiento de los pacientes más jóvenes debido al tiempo de espera ni la reducción de su esperanza de vida en general.

El algoritmo NLOS estaba destinado a asignar hígados disponibles a aquellos enfermos que, según las estadísticas, tenían más posibilidades de beneficiarse de los órganos donados; sin embargo, no existen aún medidores exactos para calcular con precisión cuál es el beneficio esperado, entre otras razones porque no hay datos comparables entre los pacientes que fueron trasplantados y los que no recibieron una donación.

Por si fuera poco, los problemas surgidos a raíz de la asignación hepática automatizada iban más allá de los meros defectos estadísticos. Al igual que sucede con otros procesos automatizados de toma de decisiones, el algoritmo NLOS también presentaba serios problemas de diseño humano. En primer lugar, porque la configuración del sistema restringía la capacidad de acción de los expertos humanos y les impedía impugnar sus decisiones automatizadas. Y en segundo término porque, debido a la falta de transparencia de su funcionamiento, no había forma de apelar a casos excepcionales como el de Sarah Meredith.

Afortunadamente, la historia acabó bien para Sarah Meredith, pues el 13 de septiembre de 2023 recibió por fin la llamada que estaba esperando desde hacía tanto tiempo: los cirujanos del Hospital Addenbrooke de Cambridge le informaron de que disponían de un hígado de un donante cuatro décadas mayor que ella que pudieron asignarle directamente, ya que el proceso de selección no se sometió al control del algoritmo NLOS, sino que se llevó a cabo sin puntuación automatizada, es decir, a través del acuerdo alcanzado por el equipo de hepatólogos del hospital tras estudiar minuciosamente el historial clínico de cada uno de los candidatos en lista de espera.

Como se desprende del análisis del caso Meredith, los algoritmos, especialmente los modelos derivados directamente de datos a través de un aprendizaje automático plantean diversos retos. No solo permiten a un número importante de agentes tomar decisiones sin intervención humana, sino que, por su naturaleza tan compleja y opaca, ni siquiera sus diseñadores pueden prever cómo se comportarán en muchas situaciones.

Por otra parte, cuanto menor es la implicación directa de las personas con el algoritmo o modelo final, menos conciencia adquieren aquellas de los posibles efectos secundarios no pretendidos —éticos, morales o de otra índole— de dichos modelos, que constituyen el centro de interés del presente trabajo.

## 2. LA SUPERVISIÓN HUMANA DE LOS ALGORITMOS: ¿EFECTO PLACEBO O MECANISMO EFICAZ DE REGULACIÓN?

El de Sarah Meredith es un caso paradigmático que viene a engrosar la lista de incidencias y fallos surgidos en torno a la toma de decisiones algorítmicas. A este respecto, parece inevitable que, al conocer los efectos colaterales e inesperados generados por un sistema de IA que analiza y actúa a ciegas sobre ámbitos tan sensibles como la salud, la seguridad y/o los derechos humanos, algunos tiendan a rechazar de plano todo lo relacionado con los algoritmos y el aprendizaje automático.

Ahora bien, aunque no deben tomarse a la ligera los riesgos potenciales que comporta la toma de decisiones dependiente de los algoritmos, tampoco sería acertado que, para evitarlos, se prescindiera de sus beneficios para confiarlo todo a la supervisión humana. En este sentido, cualquier sistema de IA que se apoye en última instancia, principal o exclusivamente, en la supervisión humana no conseguirá abarcar completamente el enorme volumen y la velocidad de la toma de decisiones algorítmicas. Por ende, si no quieren verse superados por la magnitud del problema y volverse insuficientes, los enfoques inspirados solo en la supervisión humana terminarán cediendo el testigo a las decisiones algorítmicas.

Tampoco parece que la sobrerregulación de la IA sea un remedio suficiente para impedir los múltiples daños que puede ocasionar la toma de decisiones algorítmicas. Si bien es cierto que las leyes y los reglamentos no solo resultan fundamentales para reducir los riesgos y los daños generados por mal uso o el funcionamiento defectuoso de los sistemas de IA, sino también, y sobre todo, para asegurar que las acciones y omisiones que se codifican en los algoritmos estén enraizadas en la ética humanista y la búsqueda del bien común de la sociedad, la solución a los problemas generados a raíz de la toma de decisiones algorítmicas debiera ser, en gran medida, también algorítmica. Al igual que sucede con la supervisión humana, un enfoque excesivamente legalista de la IA adolece del mismo defecto: no nos permite comprender plenamente la escala inconmensurable de los problemas que eventualmente puedan ser provocados por la ingente cantidad de datos y la celeridad con la que estos se procesan y gestionan en la toma de decisiones algorítmicas (Kearns y Roth, 2020: 263).

A propósito del exceso de confianza en el efecto tranquilizador de la supervisión humana y en el poder de la regulación para contener y controlar el uso sin restricciones de la IA, algunos autores han señalado que, en realidad, esta respuesta no entraña más que una ingenua presunción alejada de la realidad del mundo tecnológico de nuestro tiempo. Ciertamente, la regulación del mecanismo de supervisión humana podría ejercer sobre nosotros un efecto calmante, es decir, funcionaría como una especie de efecto placebo regulador para la sociedad; aunque hay que admitir que, en rigor, dicho efecto sería tal vez inútil en la práctica (Koulu, 2020: 720-735).

La cuestión, por tanto, no debería centrarse tanto en hallar un modo de limitar o incluso prescindir de los algoritmos cuanto en la precisión del diseño y la definición de estas herramientas. Como advierten Kearns y Roth el problema no es el algoritmo en sí, sino la falta de correspondencia entre las entradas al algoritmo y el mundo real, dado que el sesgo está in-

crustado en los datos. No cabe esperar que el algoritmo sea capaz de descubrirlo y corregirlo por sí mismo. *Mutatis muntandis*, se trataría de la constatación de una de las máximas más citadas de la ciencia informática: «garbage in, garbage out» («si entra basura, sale basura»); dicho en otras palabras, si en el proceso de aprendizaje automático introducimos datos con sesgos, el algoritmo hará inevitablemente su predicción mediatizada por dichos sesgos.

A fin de ilustrar el error consistente en alimentar con datos desviados las profecías autocumplidas de los algoritmos y la amplificación del sesgo introducido en los mismos, ambos autores proponen el siguiente ejemplo:

Imaginemos que la jefatura de policía de la ciudad X utiliza modelos estadísticos para predecir en qué barrios hay un mayor índice de delincuencia y enviar allí un mayor número de policías. Pues bien, supongamos a continuación que, a pesar de que los barrios A y B de esta ciudad X presentan unas cifras de delincuencia subyacentes muy parecidas, la jefatura de policía decide mandar más policías al barrio A que al B. Lógicamente, esta decisión provocará que se registrarán más delitos en A que en B. En este caso hipotético, si utilizásemos los datos comparativos recabados tras la actuación policial en los dos barrios, estos datos servirán para retroalimentar sesgadamente la siguiente operación del modelo: confirmarán que lo correcto fue enviar más patrullas policiales al barrio A que al barrio B, y reforzarán cada vez más esa tendencia, razón por la cual la amplificación del sesgo en una suerte de bucle *ad infinitum*, estaría garantizada (Kearns y Roth, 2020: 131-132).

De todo cuando antecede se desprende una primera conclusión: para cerciorarse de que los modelos de toma de decisiones cumplen con las normas sociales que se desea respetar, no solo es necesario diseñar con claridad estos objetivos directamente en los algoritmos con los que se va a trabajar, sino también situar sus metas en un primer plano. Hay que tener en cuenta, además, que la toma de decisiones complejas automatizada que aporta el aprendizaje de las máquinas tiene un perfil propio y diferente del de su diseñador. Tal vez este conozca bien el algoritmo que utilizó para encontrar el modelo de toma de decisiones, pero no el modelo en sí (Kearns y Roth, 2020: 23 y 26).

Los diseñadores, los programadores y los desarrolladores de las complejas redes de algoritmos que componen la IA son las personas que mejor conocen sus limitaciones, inconvenientes y peligros, y, posiblemente, también las que tienen al alcance la mejor manera de mitigarlos. Para ello, es preciso que los científicos informáticos y los tecnólogos especializados en el aprendizaje automático se comprometan e impliquen como actores centrales en los debates éticos que puedan surgir en torno a la toma de decisiones algorítmicas. A este respecto, tampoco estaría de más contar con equipos multidisciplinares en el diseño e implementación de las programaciones algorítmicas y, muy especialmente, integrar en estos equipos interdisciplinares a especialistas en la materia específica de dichos programas (Ponce Solé, 2023: 223).

Al hilo de lo anteriormente expuesto, cabe añadir que la perspectiva ético-jurídica de la IA se presenta como un enfoque indispensable porque contribuye a allanar el desnivel que en ocasiones se produce entre los principios y valores de la sociedad respecto a los criterios prácticos del mercado aplicados por las organizaciones, compañías y entidades que recopilan, almacenan y gestionan nuestros datos personales.

## 3. LA FUNCIÓN CORRECTORA DE LOS PRINCIPIOS DE LA ÉTICA DE LA IA EN EL DISEÑO Y DESARROLLO DE LOS ALGORITMOS

La IA está tan inserta en nuestra vida cotidiana y en el contexto social que resulta necesario tomar en consideración desde un punto de vista ético los efectos y las consecuencias que esta tecnología produce sobre el género humano en la era de la automatización. Esta es la principal razón por la que resulta oportuna una investigación crítica en clave ética que no solo se ocupe de la eficacia y el rendimiento de las nuevas tecnologías, sino también del bienestar de la humanidad ante el horizonte de la singularidad tecnológica (Schaich Borg, Sinnott, Armstrong y Conitzer, 2024: 190-191).

El objetivo de este enfoque ético es doble: por un lado, promover la concordancia o el alineamiento entre las intenciones de las distintas partes implicadas y los valores éticos pertinentes para el uso previsto; por otro lado, identificar, corregir o denunciar las aplicaciones que sirven a fines éticamente inaceptables que ignoran, o violan, valores fundamentales en relación con su ámbito de actuación. Este segundo objetivo (la vigilancia orientada a evitar el uso indebido de los sistemas de IA) tiene una gran relevancia porque, si se verifica un uso desalineado de los sistemas, entonces la denuncia del problema y la puesta en marcha de acciones dirigidas a su eliminación constituye un acto obligatorio y absolutamente necesario.

En realidad, la ética es un ingrediente básico para que la investigación tecnológica pueda avanzar tranquilamente, explorar sus propias posibilidades y ser beneficiosa para la humanidad, dado que propicia el mantenimiento de un alto nivel de confianza entre los actores implicados en el desarrollo de la IA. En última instancia, como sostiene Stefano Quintarelli, razonar sobre el posible impacto moral de la IA y asegurarse de que tenga efectos beneficiosos sobre la existencia de la humanidad no tiene porqué suponer un freno al pleno desarrollo de la tecnología sino; por el contrario, es la mejor receta para su éxito a largo plazo y para cosechar todos los beneficios que promete (Quintarelli, 2020: 84).

Sin embargo, para que haya una alineación entre las tecnologías y las expectativas éticas, estas deben estar claramente definidas. El establecimiento de un marco de valores lo más coherente y claro posible es, en efecto, crucial para que la investigación y el desarrollo tecnológicos sean éticamente conscientes y se practiquen a la luz de esta conciencia. A propósito de la definición de este marco axiológico relativo a la IA, cabe hacer mención de la gran labor ético-jurídica realizada a través de las declaraciones internacionales sobre los principios éticos de la IA. En este sentido, Floridi (2022: 94-95) ha destacado los tres instrumentos que, a su juicio, son los más influyentes hasta ahora en la definición del marco ético-jurídico de la IA:

– Los Principios de Asilomar. Se trata de un conjunto de 23 principios formulados bajo los auspicios del Future Life Institute en colaboración con los participantes en la Conferencia Asilomar, celebrada en Pacific Grove (California) en enero de 2017. Estos principios hacen referencia, entre otras cuestiones, a la seguridad y la protección, la transparencia y la responsabilidad, el desarrollo y uso de las tecnologías de IA de forma justa y equitativa, así como a la garantía de que los sistemas de IA se

diseñen y operen sin daño ni discriminación de ningún individuo o grupo (es decir, que todas las personas tengan acceso al uso y el beneficio de la IA).

– La Declaración de Montreal para un desarrollo responsable de la inteligencia artificial (2018), que enuncia diez principios que sientan las bases para fomentar la confianza de la sociedad en los sistemas de IA. El sexto principio es el de equidad, que precisamente en su primer apartado exhorta al diseño y el entrenamiento de los sistemas de IA de manera tal que «[…] no creen, refuercen ni reproduzcan patrones de discriminación basados en diferencias sexuales, étnicas, culturales o religiosas, entre otras».

– La Declaración sobre Inteligencia Artificial, robótica y sistemas «autónomos», publicada en marzo de 2018 por el Grupo Europeo sobre Ética de la Ciencia y las Nuevas Tecnologías de la Comisión Europea (en adelante, EGE), en la que se proponen un conjunto de principios éticos fundamentales y prerrequisitos democráticos basados en los valores establecidos en los Tratados y en la Carta de derechos fundamentales de la UE: a) dignidad humana; b) autonomía; c) responsabilidad; d) justicia, equidad y solidaridad; e) democracia; f) Estado de Derecho y rendición de cuentas; g) seguridad, protección, e integridad física y mental; h) protección de datos y privacidad; y i) sostenibilidad.

A diferencia de los Principios de Asilomar y de la Declaración de Montreal, el apartado d) de la Declaración de la EGE alude explícitamente a los sesgos discriminatorios y manifiesta que la IA debería contribuir a la justicia global y facilitar la igualdad de acceso a los beneficios y ventajas de la IA, la robótica y los sistemas «autónomos». Concretamente, el final del primer párrafo de este cuarto parágrafo proclama que:

> «[…] los sesgos discriminatorios en los conjuntos de datos utilizados para entrenar y ejecutar los sistemas de IA deben evitarse. De no ser posible, estos sesgos deben ser detectados, notificados y neutralizados en la etapa más temprana del proceso».

Consciente de la necesidad de reconocer y regular el impacto de la IA en el sistema de derechos fundamentales, la Unión Europea se ha situado a la vanguardia de la creación de un marco jurídico específico sobre IA. En este sentido, el 14 de marzo de 2017 el Parlamento Europeo aprobó una resolución sobre las implicaciones de los macrodatos en los derechos fundamentales: privacidad, protección de datos, no discriminación, seguridad y aplicación de la ley. La resolución, que marcó un hito en el comienzo de la normativa de la Unión Europea sobre IA, insiste precisamente en el hecho de que:

> «[…] los ciudadanos, los sectores público y privado, el mundo académico y la comunidad científica solo podrán aprovechar plenamente las perspectivas y oportunidades que brindan los macrodatos si la confianza pública en esas tecnologías se garantiza mediante la estricta observancia de los derechos fundamentales y el cumplimiento de la legislación vigente de la Unión en materia de protección de datos, así como la seguridad jurídica en relación con todas las partes interesadas».

A su vez, en el parágrafo 20 de esta misma resolución, el Parlamento insta a la Comisión, a los Estados miembros y a las autoridades encargadas de la protección de datos a que:

> «[…] definan y adopten las medidas que se impongan para minimizar la discriminación y el sesgo algorítmicos y a que desarrollen un marco ético común sólido para el tratamiento transparente de los datos personales y la toma de decisiones automatizada que sirva de guía para la utilización de los datos y la aplicación en curso del Derecho de la Unión[1]».

A propósito de las declaraciones de principios de la IA, Floridi ha advertido que existe una suerte de hipertrofia de los principios éticos que son proclamados cada vez que una organización internacional aprueba una declaración, aun a riesgo de incurrir en innecesarias repeticiones de —o superposiciones con— los que han sido enunciados en otras declaraciones precedentes, recurrencia que conduce a la confusión y la ambigüedad.

Frente a esta profusión de principios éticos, Floridi propone la idea de que la IA no debe ser entendida como un nuevo tipo de inteligencia, sino como una forma de actuar sin precedentes. Por ello, sostiene el autor, de todas las áreas de la ética aplicada, la bioética es la que más se parece a la ética digital en el tratamiento que esta hace de nuevas formas de agentes, pacientes y entornos (Floridi, 2013). Sin embargo, aunque los principios bioéticos se adaptan perfectamente a los nuevos retos éticos planteados por la IA, no es fácil explicarlos. A este respecto, Floridi añade a los cuatro principios clásicos de la bioética (*beneficencia, no maleficencia, autonomía y justicia*) un quinto principio: la *explicabilidad*, entendida como un principio que incluye tanto el sentido *epistemológico* de *inteligibilidad* como la respuesta a esta pregunta: «¿cómo funciona?», del sentido ético de responsabilidad (*accountability*) d respuesta a la pregunta: «¿quién es responsable del modo en que funciona?» (Floridi, 2022: 96).

Estos cinco principios son válidos tanto para expertos —por ejemplo, los diseñadores o los ingenieros de productos— como para los no expertos, entre ellos los pacientes o losclientes (Watson y Floridi, 2020: 1-32).

A continuación, comentaremos sucintamente y por separado cada uno de estos cinco principios.

*i)* Beneficencia

De los cuatro principios tradicionales de la bioética, el de beneficencia es el más fácil de interpretar por su claridad. La mayoría de las declaraciones de principios de la IA coinciden a la hora de identificar este término con la idea de «bienestar» de los seres humanos y de todas las criaturas sintientes (Declaración de Montreal), y conciben este principio como un criterio-guía que prescribe que los desarrollos de la IA se orienten al bien común, el beneficio de la humanidad y del planeta (Conferencia de Asilomar).

*ii)* No maleficencia

Este principio complementa al de beneficencia, en la medida en que advierte de las diversas consecuencias negativas que se derivan del uso abusivo e inadecuado de las tecnologías de IA, entre ellas las violaciones de la intimidad personal. En este sentido, la Declaración de Montreal sostiene que quienes diseñan y desarrollan los sistemas de IA deben asumir su responsabilidad y actuar contra los riesgos derivados de la innovación tecnológica. Sin embargo, el noveno principio de esta declaración (el de responsabili-

dad) introduce el siguiente matiz al respecto: cuando un sistema de IA provoca daños o perjuicios y se demuestra que, a pesar de haber sido utilizado según lo previsto, es responsable de un daño, no resulta razonable culpar a las personas involucradas en su desarrollo y uso.

*iii)* Autonomía

Este principio se relaciona con el anterior en la medida en que la autonomía funcional de la máquina inteligente implica una suspensión del control humano en la realización de determinada tarea automatizada. No parece razonable atribuir al usuario la responsabilidad de los eventuales efectos negativos generados por el funcionamiento erróneo del sistema. En otras palabras, en caso de erosión del control del funcionamiento del sistema, la autonomía funcional puede considerarse una buena razón para limitar, también, la responsabilidad de los individuos implicados en su diseño y desarrollo como programadores, constructores, etc. (Powers y Ganascia, 2021: 31-32).

Ahora bien, si ningún actor humano puede ser considerado plenamente responsable de los efectos producidos por el funcionamiento de la IA, y si no basta tampoco con imputarle la culpa al propio sistema de IA, ¿cómo puede resolverse este dilema? A este respecto, hay dos alternativas posibles:

En primer lugar, recuperar algún tipo de control significativo sobre el funcionamiento autónomo de nuestras herramientas de IA. En segundo lugar, ampliar el concepto de responsabilidad más allá del paradigma del control (Quintarelli, 2020: 90).

*iv)* Justicia

Pese a su importancia, el principio de justicia es el que más acepciones presenta en las diferentes declaraciones de principios de la IA. La Declaración de Montreal apela al principio de justicia para que «[…] no se creen, refuercen ni reproduzcan patrones de discriminación basados en diferencias sociales, sexuales, étnicas, culturales o religiosas, entre otras» (sexto principio), mientras que uno de los Principios de Asilomar —concretamente, el decimoquinto— identifica la justicia con la prosperidad económica compartida; por lo demás, en otras declaraciones el término justicia tiene otros significados (por ejemplo, se hace equivalente a la equidad). En suma, como afirma Floridi, las distintas formas de caracterizar la justicia remiten, en última instancia, a una falta de claridad más amplia sobre la IA como reserva de acción inteligente creada por los seres humanos (Floridi, 2022: 100).

*v)* Explicabilidad

Para salir de esta confusión asociada a los significados divergentes atribuidos al principio de justicia, Floridi propone un quinto principio específicamente referido a la IA que complementa a los principios clásicos de la bioética: el principio de explicabilidad. La agregación del principio de explicabilidad, que incluye tanto el sentido epistemológico de «inteligibilidad» como el sentido ético de «responsabilidad», es «[…] la pieza crucial que falta para completar el puzle ético de la IA».

En efecto, este quinto principio complementa los otros cuatro porque, para que la IA sea beneficiosa y no dañina, es necesario que estemos en disposición de comprender el bien o el daño que está haciendo a la sociedad y de qué manera; por otra parte, para que la IA promueva y no limite la autonomía humana, nuestra decisión sobre quién debe decidir tiene que estar informada por un conocimiento previo: cómo actuaría la IA en nuestro lugar y, en tal caso, cómo podrían mejorarse sus prestaciones; por último, para que la IA sea justa, necesitamos saber a quién responsabilizar ética o jurídicamente en caso de que produzca un resultado grave o negativo, determinación que, a su vez, requeriría una comprensión adecuada de las razones por las que se produjo tal resultado (Floridi, 2022: 100).

## 4. EL CASO HERZBERG. CRÓNICA Y ANÁLISIS DEL PRIMER ERROR ALGORÍTMICO FATAL EN LA HISTORIA DE LA IA

La noche del 18 de marzo de 2018 un vehículo de prueba de la empresa Uber conducido por robot circula a 69 km/h por una avenida de cuatro carriles en Tempe (Arizona). En el automóvil autónomo va sentada en el asiento de copiloto Rafaela Vásquez, cuyo único cometido era mantenerse atenta a la conducción del coche robotizado para vigilar que no cometiera ninguna infracción del código de circulación.

Alrededor de las 22 horas, en medio de la North Mill Avenue, aparece sorpresivamente, como salida de la oscuridad, una mujer empujando una bicicleta en la que lleva todas sus pertenencias. Se trata de Elaine Herzberg, una indigente de mediana edad que cruza la autovía ajena a la circulación. Cuando el sistema de detección del Uber autónomo percibe a la Sra. Herzberg y los algoritmos dan la orden de frenado de emergencia, ya es demasiado tarde para evitar el atropello de la viandante, que muere una hora más tarde en el hospital.

¿Cómo pudo producirse ese error algorítmico que causó la primera víctima de la IA? ¿Por qué no actuó antes la Sra. Vásquez para evitar el impacto del vehículo que debía supervisar contra la Sra. Herzberg? ¿De qué modo estaba programado el *software* responsable de tomar la decisión última para que fallara, hasta el punto de que confundió a la desdichada peatona con un falso positivo?

El atropello de Elaine Herzberg se debió a un cúmulo de circunstancias adversas y de errores humanos e informáticos. Veamos seguidamente la sucesión de contratiempos y fallos que provocaron el fatal accidente y que nos permitirán entender cómo fue posible.

En primer lugar, la Sra. Herzberg cruzó la avenida de forma temeraria, en medio de la oscuridad, porque quería llegar pronto al campamento de personas sin hogar en el que vivía. No en vano, hay que señalar que ese tramo de carretera es precisamente uno de los trayectos en los que Uber prueba sus coches robotizados (hay que tener en cuenta que este tipo de pruebas son posibles desde que, en 2015, un decreto de Doug Ducey, gobernador de Arizona, autorizó a circular por las carreteras de este estado a vehículos controlados por un robot, incluso sin necesidad de que hubiera un ser humano dentro; a cambio de la aprobación de esta legislación tan permisiva con los coches autopilotados, Uber instaló su flota de automóviles en el estado del Gran Cañón).

En segundo lugar, Uber aprovechó la flexibilidad de la ley de seguridad vial de Arizona para reducir el número de *robot babysitters*, es decir, controladores humanos de los coches autónomos (la empresa tenía 400 vehículos de este tipo circulando por las carreteras de Tempe). A diferencia de su principal competidor, Waymo —filial de los coches autónomos de Google—, cuyos vehículos contaban con dos personas por automóvil para garantizar la seguridad en la conducción, Uber —compañía más preocupada por recortar gastos por los costes derivados de los sueldos de sus *robot babysitters* para obtener beneficios que por garantizar la seguridad de los demás conductores y la integridad física de los peatones— redujo a solo uno el número de conductores humanos a bordo de sus Volvo xc90 robotizados.

En tercer lugar, cabe subrayar que la conducta de la conductora auxiliar encargada de vigilar el buen funcionamiento del sistema de conducción inteligente fue imprudente. La normativa del estado de Arizona obliga a los vigilantes de los coches autónomos a llevar las manos sobre el volante en todo momento y a mantener permanentemente la atención en todo cuanto suceda tanto en el interior del automóvil como en la carretera para reaccionar a tiempo ante cualquier imprevisto que pueda presentarse y sobre el que la máquina no sepa qué decisión tomar. Sin embargo, la conductora auxiliar, Rafaela Vásquez no fue diligente en el cumplimiento de su obligación de supervisar el buen funcionamiento del sistema de conducción inteligente del Volvo xc90 en el que se desplazaba, y tampoco estuvo pendiente de la carretera porque solo estaba pendiente de la pantalla del móvil que llevaba en su regazo, en la que veía un programa de televisión.

La señora Vásquez parecía confiar plenamente en los sistemas de IA que guían la conducción del coche autónomo. En efecto, la flota de Volvo xc90 robotizados de Uber disponía de un complejo sistema de herramientas de IA capaz de aprender por sí mismo. Incluía dispositivos de radar, cámaras delanteras y laterales, sensores de navegación y escáneres láser (el denominado 'sistema Lidar', un aparato instalado en el techo del vehículo, similar a una sirena de la policía, que va disparando a su alrededor millones de pulsos láser por segundo para detectar objetos, medir distancias y formar a partir de ellos imágenes tridimensionales).

Como se ve, ni siquiera las predicciones de los algoritmos de un sistema de conducción inteligente tan complejo como el que llevan instalados los coches Uber son infalibles. En realidad, el problema son los falsos positivos que, como en el caso del atropello mortal de Elaine Herzberg, distorsionan los datos recabados del exterior del automóvil con los que los algoritmos elaboran predicciones de lo que podría suceder en los segundos que siguen a la detección de un objeto y determinan en qué dirección se moverá, si se trata de un peatón o un animal que invade la calzada, el tiempo aproximado de reacción antes de que se produzca el impacto con el supuesto objeto…

En el caso que nos ocupa, el sistema de detección percibió la presencia de la Sra. Herzberg con su bicicleta en mitad de la calzada seis segundos antes del atropello, tiempo suficiente para desviar la trayectoria del coche o para frenar. Sin embargo, los algoritmos del automóvil de Uber que se desplazaba en modo de conducción automático por la North Mill Avenue de Tempe mientras la conductora visionaba un programa de televisión en su

móvil dudaron y no se pusieron de acuerdo. En primera instancia, el *software* del vehículo autónomo creyó haber detectado un objeto desconocido, un segundo más tarde estimó que se trataba de un coche y, solo cuando faltaban 1.3 segundos para el impacto, confirmó que es una bicicleta y dio la orden de ejecutar el frenado de emergencia.

Teniendo en cuenta el desarrollo tecnológico ya existente en 2018, suficiente al menos para garantizar a los usuarios de un vehículo Uber autónomo la seguridad en la conducción de un *software* equipado con algoritmos capaces de hacer predicciones y dar instrucciones precisas durante la conducción, la pregunta que cabe hacer es cómo no se activó antes el frenado de emergencia pese a la detección de un objeto en la carretera con un margen de 6 segundos, una eternidad para un sistema de IA.

Según las revelaciones de dos antiguos empleados de Uber, tiempo después de estos acontecimientos, la empresa habría elevado el nivel de tolerancia a los falsos positivos en aquellos tramos de carretera elegidos para realizar las pruebas de conducción autónoma de sus vehículos con el único fin de ser más competitiva que sus rivales directos en el mercado de los VTC (Google, General Motors e Intel). La reducción de la seguridad vial a cambio de la mayor fluidez y puntualidad en el recorrido de prueba realizado por el Uber, es decir, sin detenerse cada dos por tres a causa de presuntos obstáculos que pudieran aparecer en la carretera, era un sacrificio aceptable si con ello se conseguía convencer a los inversores y a los potenciales usuarios de que estaban ante el vehículo del futuro.

El riesgo de esta estrategia empresarial tan agresiva y temeraria es que los fabricantes de automóviles autónomos den por descontado que los accidentes de tráfico se producen indefectiblemente, que los algoritmos que gobiernan sus sistemas de conducción cometen errores y que dichos errores cuestan vidas humanas. A propósito del alto riesgo asumido por las empresas de vehículos autónomos, Anthony Levandowski, el ingeniero responsable del diseño del coche de conducción automática de Google que robó datos de esta empresa para fundar su propia empresa, adquirida con posterioridad por Uber, afirmó en la Cumbre AV de coches autónomos de 2019 que la industria del automóvil necesita un avance fundamental en IA para que se produzcan progresos significativos en la tecnología de este tipo de vehículos: si de verdad queremos impulsar una tecnología, advirtió en su ponencia Levandowski, entonces la seguridad no puede ser la prioridad dado que, si así fuera, no se conseguiría nada.

En relación con esta lógica del mercado tecnológico, despojada de principios éticos inspirados por el paradigma de la moral humanista, vemos que hay casos extremos y paradójicos en los que el comportamiento de un automóvil autónomo responde solo a los principios *éticos* internos de la empresa, que quedan plasmados en el *software* del vehículo (Colomba, 2019: 94).

La fuerza performativa de la tecnología, que, como en el caso de la primera víctima registrada de la IA, reduce drásticamente el abanico de opciones posibles de actuación e invita a plantear algunos interrogantes sobre las transformaciones estructurales y el condicionamiento que la autorregulación de las empresas tecnológicas impone a la libertad de actuación de las personas (Lessig, 1999: 7-8).

Desde un punto de vista operativo, la lógica del presupuesto ético-moral que justifica el automóvil autónomo sería la siguiente: en primer lugar, las carreteras por las que circulen los coches *self-driving* son más seguras a medio plazo que las carreteras por donde circulan los coches guiados solo por conductores humanos; esto implicaría que, *ceteris paribus*, es inmoral no apostar por la fabricación masiva de automóviles autónomos. De lo que antecede se desprende, en segundo lugar, la siguiente conclusión: la *ratio* moral que impulsa la producción de automóviles autónomos no es punitiva («sufre quien lo merece»), sino *utilitarista*: según este enfoque ético, en el caso que hemos estudiado el del atropello mortal de la señora Herzberg, primaría el cumplimiento del objetivo empresarial, a saber, la puntualidad en cubrir el recorrido del trayecto elegido para hacer la prueba sobre el principio ético-moral de seguridad (Casadei y Zanetti, 2019: 45-46).

Por otra parte, esta autorregulación basada en la ética empresas de vehículos autónomos como los que componen la flota de Uber, Google, Intel o General Motors, también pone de relieve la cuestión del grado de responsabilidad que cabe atribuir a los diseñadores y desarrolladores de los algoritmos, esto es, cuestiona la sostenibilidad ética de las decisiones que toman y de la elección de opciones morales que hacen, en la medida en que estos asumen una responsabilidad pública al decantarse por determinado resultado o por un modelo de funcionamiento específico (De Vanna, 2019: 82).

Llegados a este punto, algunos especialistas en ética de la IA se han preguntado qué criterio ético aplicar a los vehículos autónomos. ¿Debería establecerse una ética única a través de un pacto social? ¿O bien dejar que las empresas se autorregulen? ¿O tal vez consentir la coexistencia de diferentes enfoques? ¿En el futuro podremos elegir nuestros coches autónomos en función de que su *software* que se alinee más con nuestros valores y no en función de sus prestaciones? ¿Cómo desearíamos que fueran diseñados los algoritmos del sistema que conducen ese automóvil? ¿De manera altruista o de modo que preserven ante todo la vida de su dueño? (Sigman y Bilinkis, 2023: 183; Schaich, Borg, Sinnott, Armstrong y Conitzer, 2024: 17).

En suma, podríamos concluir este epígrafe coincidiendo con el parecer de quienes piensan que el asunto de la confiabilidad de las máquinas inteligentes va más allá de una simple cuestión ética. Es probable que en el futuro los vehículos automatizados no sean conducidos por personas ni por ordenadores, sino por empresas que actuarán a través de sus operadores humanos y máquinas. Una cuestión esencial para este campo —y para la inteligencia artificial en general— es cómo deben ganarse nuestra confianza las empresas que desarrollan y despliegan estas tecnologías.

Una empresa que es digna de confianza comparte sin problemas su filosofía de seguridad, hace una promesa al público y la cumple. Una empresa que comparte su filosofía de seguridad no tiene inconveniente en informar a los usuarios con total transparencia con un mensaje parecido a este: «Esto es lo que hacemos; por esto pensamos que es razonablemente seguro; y por esto puede Vd. creernos». Su promesa sería la siguiente: «Solo comercializamos lo que razonablemente creemos que es seguro; seremos sinceros sobre nuestras limitaciones y fallos; y cuando fallemos, lo corregiremos». Y la empresa cumple esa promesa gestionando adecuadamente las expectativas del público, supervisando el ciclo de

vida de su producto o servicio y mitigando los daños de forma rápida, completa y pública (Smith, 2020: 682).

## 5. UN DILEMA MORTAL Y TRES MODELOS ÉTICOS PARA LOS SISTEMAS ALGORÍTMICOS

En 1967, la filósofa inglesa Philippa Foot, que defendía una ética contemporánea de las virtudes inspirada por la ética aristotélica (de acuerdo con la cual una acción es éticamente correcta si hacerla fuera propio de una persona virtuosa), publicó un artículo en la *Oxford Review* sobre del problema del aborto y la doctrina del doble efecto. En este estudio, la autora proponía varios ejemplos de dilemas éticos para explicar la doctrina del doble efecto: esta doctrina se basa en una distinción entre lo que un hombre *prevé* como resultado de su acción voluntaria y lo que, en sentido estricto, *pretende*. En otras palabras, este agente moral pretende, *strictu sensu*, tanto las cosas que se propone en cuanto fines como las que se propone en cuanto medios para conseguir sus fines. Estos últimos pueden ser lamentables en sí mismos, aunque deseados en aras de la consecución del fin propuesto.

Entre los diversos ejemplos utilizados por Foot para ilustrar el dilema que plantea la doctrina del doble efecto, quizás el problema del conductor del tranvía (*the trolley driver problem*) es el más extrapolable a la cuestión de la ética del diseño y desarrollo de los algoritmos. En realidad, se trata de un dilema al que se recurre cada vez que se pretende generar patrones de comportamiento humano en la toma de decisiones de los vehículos autónomos, en la tecnología de los drones bélicos configurados con IA para minimizar daños colaterales, o en los protocolos de ética médica —como en el caso de Sarah Meredith— concebidos pata tomar decisiones que afecta a la vida humana (Campione y Pietropaoli, 2024: 133-135).

Pasemos ahora a analizar el referido dilema moral que propone Foot:

> «Supongamos que un juez o magistrado se enfrenta a alborotadores que exigen que se encuentre a un culpable de un determinado delito y amenazan con vengarse sangrientamente de un sector particular de la comunidad. Como se desconoce el verdadero culpable, el juez se ve a sí mismo capaz de evitar el derramamiento de sangre incriminando únicamente a una persona inocente y haciéndola ejecutar. Al lado de este ejemplo se coloca otro en el que un piloto cuyo avión está a punto de estrellarse está decidiendo si se dirige de una zona más a una menos habitada. *Para que el paralelismo sea lo más cercano posible, puede suponerse más bien que es el conductor de un tranvía fuera de control, que solo puede conducir de una vía estrecha a otra; cinco hombres trabajan en una vía y un hombre en la otra; cualquiera que esté en la pista en la que entra está destinado a ser asesinado*. En el caso de los disturbios, la turba tiene cinco rehenes, por lo que, en ambos ejemplos, se supone que el intercambio es la vida de un hombre por la vida de cinco» (Foot, 1967: 8).

Las respuestas posibles a este dilema variarán en función de la teoría ética por la que nos decantemos:

En primer lugar, si optásemos por un enfoque ético consecuencialista o teleológico, como el utilitarista, la acción más correcta sería la que genera el mayor bien posible o que

comporta más bien que mal. En el caso del dilema del conductor del tranvía fuera de control, una persona que actuase conforme a las directrices morales de la ética utilitarista no dudaría en responder que la mejor opción, la más deseable, sería la de desviar el tren a la vía en la que hay un solo hombre.

Por el contrario, si nuestro patrón moral fuera la ética deontológica y nos fijásemos más en el valor propio de la acción que en las consecuencias de la misma, entonces la solución al dilema del conductor del tranvía no dependería de la cantidad de bien o felicidad que comportara la decisión. Desde el punto de vista del paradigma ético principialista o deontológico, representado por la ética kantiana, el maquinista debería actuar conforme a unas reglas de comportamiento universalizables, al margen de las consecuencias que pudieran producirse de esa actuación. En todo caso, la ética formal de Kant está coronada por unos principios éticos humanistas, iusracionalistas, ilustrados y antropocéntricos que fomenten tanto la autonomía de la voluntad como la constitución de una sociedad abierta de personas racionales y de ciudadanos libres e iguales.

En este sentido, aunque la decisión de desviar el tranvía hacia la vía en la que solo hay una persona sería considerada como ilícita por la ética deontológica, en última instancia esta sería sensible a la necesidad de ponderar las consecuencias globales de acciones morales ilícitas si están dirigidas a un fin mayor. Para definir cuáles serían esos principios de justicia universales que inspiran las reglas de comportamiento de un buen ciudadano dentro de una sociedad abierta y bien ordenada, podría ser útil la teoría contractualista-liberal de John Rawls, en la que, al contrario de lo que sucede con el utilitarismo, las personas aceptan por anticipado actuar (previo acuerdo con los demás participantes en el pacto social) según un principio de igual libertad sin un conocimiento previo de sus fines más particulares, de manera que convienen en adecuar sus concepciones del bien a lo que prescriben los principios de justicia, que son el resultado del acuerdo original entre los individuos —principio de libertades o de distribución de igual número de esquemas de libertades para todos; y principio de diferencia, que expresa un sentido de amistad cívica y solidaridad moral que incluye la igualdad en la estimación social, y excluye los privilegios o el servilismo dentro de la sociedad— (Rawls, 1971: 60-62).

Por último, para la ética de la virtud, la moral surge de los rasgos internos de las personas, de las virtudes, que se contraponen tanto a los principios de la ética deontológica, en la que la moral proviene de las normas, y a los de la ética consecuencialista, en la que la moral depende de los resultados del acto. En consonancia con el enfoque moral de la ética de la virtud, Philippa Foot sostiene que el fin no puede justificar los medios, como propone el utilitarismo. Por lo tanto, para la teoría de la virtud la respuesta al dilema del conductor del tranvía pasaría por una solución opuesta a la de la teoría de la ética consecuencialista: el conductor debería dejar que el tranvía siga por la vía principal y no hacer ninguna maniobra de desvío, aunque fuera a costa de arrollar a los cinco operarios que trabajan en esa vía. Según Foot, si el conductor no interviene y se queda quieto, no sería responsable moral de las cinco muertes, mientras que, si desvía el tranvía a la pista secundaria en la que se encuentra un hombre, el conductor sería parcialmente responsable de su muerte y su actuación constituiría una participación en el mal moral.

Como se ha comentado antes, el dilema del conductor del tranvía es perfectamente aplicable al diseño y el desarrollo de los sistemas algorítmicos. Una buena muestra de su utilidad lo encontramos precisamente en el diseño del *software* para controlar automóviles autónomos (Hazel Si Min y Taeihagh, 2019: 5791).

Desde el año 2016, el Massachusetts Institute of Technologie (M.I.T, por sus siglas en inglés) viene desarrollando en su web un programa pedagógico de supuestos prácticos en los que se actualiza el clásico dilema del tranvía, ya lejano del imaginario de la sociedad tecnológica, y este se sustituye por el coche autónomo. Esta plataforma *online*, denominada *Moral Machine* (en adelante, MM), se presentan diferentes escenarios dilemáticos en los que un vehículo sin conductor debe elegir entre dos males. Basándose en las respuestas recabadas de los millones de personas que visitaron la web y participaron en esta prueba, se ha podido establecer una imagen muy aproximada de la compleja y heterogénea perspectiva de los seres humanos sobre las decisiones tomadas por máquinas inteligentes.

Veamos, a continuación, la representación gráfica y la explicación de la primera combinación de partida, que reproduce el dilema del automóvil automático sin frenos inspirándose en el dilema clásico del tranvía:

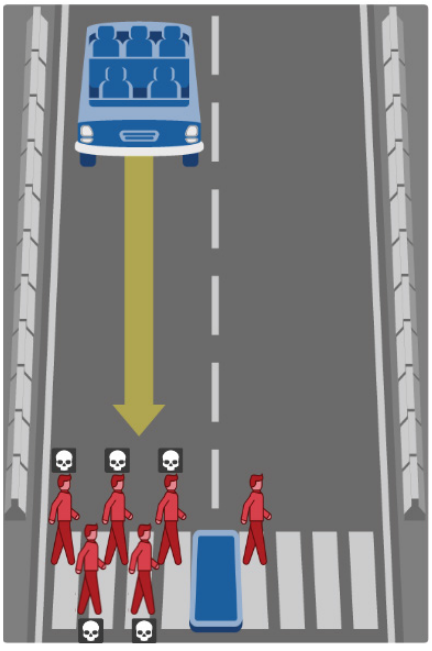

Figura 1. En este caso, el coche autónomo con fallo en los frenos sigue adelante y atraviesa el paso de cebra arrollando a los cinco peatones.

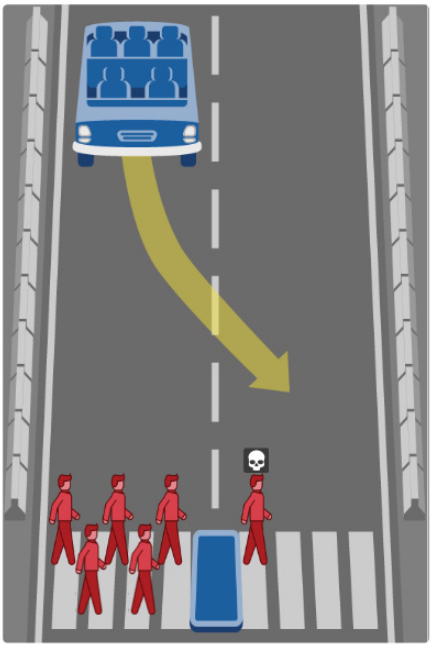

Figura 2. En este caso, el coche autónomo con fallo en los frenos gira, invade el otro carril, atraviesa el paso de cebra y arrolla al único peatón que cruza el paso de peatones

Además de su utilidad para la programación de la IA que toma decisiones, la finalidad de esta prueba de simulación se dirige principalmente a alinear los algoritmos morales

con los valores humanos mediante una «conversación global» sobre cuáles son esos valores aceptados y consensuados (Bonnefon, Shariff y Rawan, 2016).

La plataforma MM permite a sus usuarios diseñar una infinitud de escenarios para compartir, examinar y discutir *online* con otros usuarios. Las respuestas recopiladas proporcionan datos de gran valor, en la medida en que reflejan la percepción humana de la autonomía de las máquinas. Aunque el problema original del tranvía es un experimento mental útil que los filósofos utilizan para explorar las relaciones entre las intuiciones morales humanas y las predicciones de distintas teorías filosóficas, en puridad no es una guía útil para el diseño de robots corporales en el mundo físico. Para diseñar un robot ético (como un vehículo autónomo), hay que salir del estrecho marco figurativo del problema del tranvía y formular una norma social adicional.

Una situación figurada como la que nos plantea el dilema mortal del tranvía, que no ofrece opciones ni resultados éticamente buenos (en ambos caos se produce un atropello mortal), debería desencadenar un pensamiento contrafáctico, de modo que el sistema algorítmico de conducción autónoma aprenda que una situación que antes no tenía nada de especial, como entrar en una calle estrecha, exige una reducción de velocidad, para preservar la opción de realizar una parada de emergencia que salve la vida de una persona o de un grupo de personas que puedan cruzar la carretera (Kuipers, 2020: 429).

Dado que la IA actúa de manera diferente a los humanos en determinadas condiciones, tener idea de si una entidad es humana o IA hará que su comportamiento sea más predecible para los demás, aumentando tanto la eficacia como la seguridad (Walsh, 2017: 113-114).

Si una persona se pone delante de un vehículo que circula a 69 km/h, el conductor humano medio puede que sea incapaz de reaccionar con la rapidez suficiente para adoptar una maniobra evasiva, mientras que un sistema de IA sí podría hacerlo. Sin embargo, en otras situaciones, sobre todo las que requieren «sentido común», es probable que la IA (al menos de momento) sea muy inferior al ser humano. Los coches de IA pueden ser expertos en una autopista, pero evaluar la presencia de elementos complejos o inusuales —por ejemplo, una persona que empuja una bicicleta— puede ser más difícil para ellos. Del mismo modo que a un niño pequeño le hablamos de forma diferente, a una IA le daremos instrucciones distintas a las que damos a los humanos en aras de nuestra protección, pero también de la protección de la propia IA (Turner, 2019: 322).

## 6. CONCLUSIÓN

A lo largo de este recorrido por la casuística más reciente en torno a las implicaciones éticas de las decisiones tomadas por la IA, hemos podido comprobar que los modelos predictivos que sirven para el entrenamiento de los ordenadores en el procesamiento y análisis de datos se construyen no sólo a partir de los datos, sino también de las decisiones que tomamos sobre a qué datos prestar atención y cuáles omitir. Estas decisiones no solo tie-

nen que ver con la logística, los beneficios y la eficiencia. Son fundamentalmente morales (O'Neil, 2016: 218).

En primera instancia, puede parecer plausible modelizar sistemas éticos con técnicas de IA, ya que las prescripciones en las que se basan dichos sistemas han sido introducidas por humanos. Sin embargo, los intentos de modelizar el razonamiento ético han puesto de manifiesto las enormes dificultades a las que se enfrentan los investigadores para hacerlo. La primera dificultad proviene del modelado del razonamiento deóntico, es decir, el razonamiento sobre obligaciones y permisos. La segunda se debe a los conflictos de normas que se producen constantemente en el razonamiento ético. La tercera está relacionada con el entrelazamiento del razonamiento y la acción, que exige que estudiemos la moralidad del acto, *per se*, pero también los valores de todas sus consecuencias.

El reto técnico actual consiste en fusionar estos tres enfoques, es decir, crear uno que no sea uniforme, que pueda gestionar conflictos entre normas y que utilice modelos causales para evaluar las consecuencias de las acciones. Aunque existe un interés general en la creación de una máquina moral de este tipo (es decir, que se comporte de acuerdo con las normas de un enfoque moral), todas estas perspectivas abarcan distintos marcos normativos —el utilitarismo, la deontología y la ética de la virtud— que deben simularse. Los detalles de las simulaciones suelen resultar insuficientes, sobre todo para los filósofos. Además, surgen cuestiones sobre la utilidad práctica de tales máquinas morales, así como sobre las dificultades para ponerlas en práctica (Powers y Ganascia, 2021: 40-41).

En realidad, la ética es un ingrediente básico para que la investigación tecnológica pueda avanzar sosegadamente, para que explor sus propias posibilidades y para que sea beneficiosa para la humanidad, en la medida en que propicia el mantenimiento de un elevado nivel de confianza entre los actores implicados en el proceso de la IA. En última instancia, como sostiene Stefano Quintarelli, razonar sobre el posible impacto moral de la IA y asegurarse de que tenga efectos beneficiosos sobre la existencia de la humanidad no tiene porqué suponer un freno al pleno desarrollo de la tecnología; por el contrario, como ya se ha dicho, es la mejor fórmula para que alcance su éxito a largo plazo y, por ello, para que cosechemos todos los beneficios que promete (Quintarelli, 2020: 84).

## NOTAS

1. Resolución del Parlamento Europeo, de 14 de marzo de 2017, sobre las implicaciones de los macrodatos en los derechos fundamentales: privacidad, protección de datos, no discriminación, seguridad y aplicación de la ley (2016/2225(INI)), §§ 1 y 20.

## BIBLIOGRAFÍA

BONNEFON, Jean-François, Azim SHARIFF e Iyad RAHWAN (2016): «The Social Dilemma of Autonomous Vehicles», *Science*, 352 (6293), 1573-1576.

CAMPIONE, Roger y Stefano PIETROPAOLI (2024): *Los artefactos de la inteligencia jurídica: personas y máquinas*, Madrid: Dykinson.
CASADEI, Thomas y Gianfrancesco ZANETTI (2019): «Tra dilemmi etici e potenzialità concrete: le sfide dell´autonomous driving» en S. Scagliarini (dir.), *Smart Roads e driverless cars: tra diritto, tecnologie, etica publica*, Turín: G. Giappichelli Editore, 41-54.
COLOMBA, Vittorio (2019): «Driverless cars e intelligenza artificiale» en S. Scagliarini (dir.), *Smart Roads e driverless cars: tra diritto, tecnologie, etica publica*, Turín: G. Giappichelli Editore, 87-95.
DE VANNA, Francesco (2019): «Autonomous driving e questione della responsabilità: alcuni nodi teorici», en S. Scagliarini (dir.), *Smart Roads e driverless cars: tra diritto, tecnologie, etica publica*, Turín: G. Giappichelli Editore, 77-86.
FLORIDI, Luciano (2022): *Etica dell´intelligenza artificiale. Sviluppi, opportunità, sfide,* Milano: Raffaelo Cortina.
FOOT, Philippa (1967): «The Problem of Abortion and the Doctrine of the Double Effect», *Oxford Review*, 5, 5-15.
HAZEL SI MIN, Lim y Araz TAEIHAGH (2019): «Algorithmic Decision-Making in AVs: Understanding Ethical and Technical Concerns for Smarts Cities», *Sustainability*, 11(20), 5791.
KEARNS, Michael y Aaron ROTH (2020): *El algoritmo ético. La ciencia del diseño de algoritmos socialmente responsables*, Madrid: La Ley-Wolters-Kluwer.
KOULU, Riikka (2020): «Proceduralizing control and discretion: Human oversight in artificial intelligence policy» *Maastricht Journal of European and Comparative Law*, 27(6), 720-735.
KUIPERS, Benjamin (2020): «Perspectives on Ethics of AI: Computer Science», en M. D. Dubber, F. Pasquale y S. Das (eds.), *The Oxford Handbook of Ethics of AI*, Oxford: Oxford University Press, 421-441.
LESSIG, Lawrence (1999): *Code and Other Laws of Cyberspace*, Nueva York: Basic Books.
O´NEIL, Cathy (2016): *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*, Londres: Penguin Books.
PONCE SOLÉ, Juli (2023): «Seres humanos e inteligencia artificial: discrecionalidad artificial, reserva de humanidad y supervisión humana», en E. Gamero Casado (dir.), *Inteligencia Artificial y sector público. Retos, límites y medios* Valencia: Tirant lo Blanch, 196-225.
POWERS, Thomas y Jean-Gabriel GANASCIA (2021): «The Ethics of the Ethics of AI», en M. D. Dubber, F. Pasquale y S. Das (eds.), *The Oxford Handbook of Ethics of AI* (ed.), Oxford, Oxford: University Press, 27-51.
QUINTARELLI, Stefano (2020): *Intelligenza Artificiale. Cos´è davvero, come funciona, che effetti avrà*, Milán: Bollati Boringhieri.
RAWLS, John (1971): *A Theory of Justice*, Cambridge (Massachusetts) – London: The Belknap Press of Harvard University Press.
SCHAICH BORG, Jana, Walter SINNOTT y Vincent ARMSTRONG-CONITZER (2024): *Moral AI and How We Get There*, Milton Keyes – Dublin: Penguin Random House.
SIGMAN, Mariano y Santiago BILINKIS (2023): *Artificial. La nueva inteligencia y el contorno de lo humano*, Barcelona: Debate.
SMITH, Bryant Walker (2020):«Ethics of Artificial Intelligence in Transport» en M. D. Dubber, F. Pasquale y S. Das (eds.), *The Oxford Handbook of Ethics of AI*, Oxford: Oxford University Press, 672-683.
TURNER, Jacob (2019): *Robot Rules. Regulating Artificial Intelligence*, Londres: Palgrave Macmillan.
WALSH, Toby (2017): *Android Dreams*, Londres: Hurt&Co.
WATSON, David S. y Luciano FLORIDI (2020): «The explanation game: A formal framework for interpretable machine learning», *Synthese*, 198, 9211-9242.

**Fecha de recepción: 1 de julio de 2024.**
**Fecha de aceptación: 20 de octubre de 2024.**