ÉTICA(S) DE LA INTELIGENCIA ARTIFICIAL Y DERECHO CONSIDERACIONES A PROPÓSITO DE LOS LÍMITES Y LA CONTENCIÓN DEL DESARROLLO TECNOLÓGICO

eTHICS(S) OF ARTIFICIAL INTELLIGENCE AND LAW. CONSIDERATIONS ON THE LIMITS AND CONTAINMENT OF TECHNOLOGICAL DEVELOPMENT

Fernando H. Llano Alonso
Universidad de Sevilla
https://orcid.org/0000-0001-7589-4166

Fecha de recepción: 19-2-24
Fecha de aceptación: 20-3-24

A María Dolores García Cossío (1938-2024)
In memoriam

Resumen: *El presente trabajo se centra en la ética de la IA y en los principios que la inspiran. Antes de las normas que regulan la IA es necesario conocer las cuestiones éticas relacionadas con el desarrollo, la implementación y el uso responsable de los sistemas de IA. El concurso de estos fundamentos éticos es necesario para proteger los derechos y la dignidad de las personas, garantizar la equidad en el acceso y tratamiento de datos, minimizar los sesgos y riesgos asociados con los algoritmos, promover la transferencia en las decisiones automatizadas y auspiciar la confiabilidad en la tecnología, el beneficio humano y el bienestar social. También se propone el estudio de la ética de la IA desde un punto de vista omnicomprensivo que integre tanto la ética aplicada como la metaética de la IA.*

Abstract: *This paper focuses on the ethics of AI and the principles behind it. Prior to the rules governing AI, it is necessary to understand the ethical issues related to the development, implementation and responsible use of AI systems. These ethical foundations are necessary to protect the rights and dignity of individuals, ensure fairness in data access and processing, minimise biases and risks associated with algorithms, promote transferability in automated decisions,*

and promote trustworthiness in technology, human benefit and social welfare. It is also proposed to study the ethics of AI from an all-encompassing point of view that integrates both applied ethics and metaethics of AI.

Palabras clave: sesgos, transparencia algorítmica, explicabilidad, responsabilidad, ciberseguridad, solidez, metaética de la IA, IA generativa, modelos del lenguaje conversacionales.

Keywords: biases, algorithmic transparency, explainability, accountability, cybersecurity, robustness, AI meta-ethics, generative AI, conversational language models.

1. INTRODUCCIÓN

A finales del mes de marzo de 2023, más de medio millar de académicos, ingenieros expertos en inteligencia artificial (en adelante, IA) y empresarios de compañías tecnológicas firmaron una carta abierta en la que advertían de los "profundos riesgos para la sociedad y la humanidad" que plantean los modelos avanzados; en consecuencia, proponían a los laboratorios de IA una pausa o *impasse*, al menos durante seis meses, en el entrenamiento de sistemas de IA más potentes que el modelo de lenguaje de OpenAI GPT4[1].

Esta carta fue publicada por *The Future of Life Institute*, un instituto que se presenta a sí mismo como una organización cuya principal misión consiste en dirigir la tecnología transformadora hacia el beneficio de la vida y alejarla de los riesgos extremos a gran escala: la IA, la Biotecnología y las armas nucleares. Según los autores de este texto, los sistemas de IA con inteligencia competitiva humana pueden representar un cambio significativo en la historia de la vida en la Tierra, por eso, conforme a los Principios de IA de Asilomar, se sostiene que la IA avanzada debe "planificarse y gestionarse con el cuidado y los recursos adecuados"[2].

Para algunos tecnólogos especializados en el estudio del riesgo de extinción que entraña la IA, pecan de ingenuidad quienes esperan que empresas como OpenAI, Google o Meta van a detener sus investigaciones, renunciando así a los beneficios y las ventajas que les proporciona el desarrollo de sus

[1] Entre los signatarios de esta carta se encuentran: Elon Musk (CEO de SpaceX, Tesla y Twitter); Steve Wozniak (Cofundador de Apple); Yuval Noah Harari; Evan Sharp (Cofundador de Pinterest); Craig Peeters (Director ejecutivo de Getty Images); Emad Mostaque (Director ejecutivo de Stability AI); o Valerie Pisano (Presidenta y Directora ejecutiva de MILA).

[2] *https:// futureoflife.org/open-letter/pause-giant-ai-experiments/*

modelos; pero aun suponiendo que esta hipótesis pudiera llegar a ser real, se preguntan estos críticos, ¿cómo respondería el gobierno chino o sus empresas, por poner un ejemplo, ante ese compás de espera de sus competidores? Probablemente, aprovecharía la oportunidad para acelerar en el desarrollo de los sistemas avanzados de IA[3].

Lejos de traer un invierno tecnológico para la IA generativa, lo cierto es que la carta solicitando la pausa de los laboratorios de IA sirvió para avivar el proceso de avance tecnológico en los meses siguientes. En efecto, laboratorios de IA y empresas tecnológicas como OpenAI han apostado por lanzar al mercado nuevas versiones de modelos de aplicaciones conversacionales más avanzadas y con mayor capacidad.

En los meses posteriores a la publicación de esta carta por *The Future of Life Institute* se produjeron avances significativos en el diseño y desarrollo de nuevos modelos de IA en los principales gigantes tecnológicos (Big Techs), así por ejemplo: Google desplegó *Gemini*, un nuevo modelo de lenguaje (multimodal, en tres versiones: Ultra, Pro y Nano) que podrá rivalizar con la aplicación conversacional Chat GPT-4; Amazon también participó en esta carrera tecnológica al anunciar la inversión de 4000 millones de dólares en Anthropic una startup estadounidense de IA generativa que se ha convertido en uno de los más importantes proveedores mundiales de modelos básicos de IA, y que se presenta a sí misma como un "destacado defensor" de la implementación responsable de IA; también Elon Musk, solo cuatro meses después de haber firmado la petición de pausar la IA, en aras de la supervivencia de la humanidad, lanzó xAI, un chatbot para competir con ChatGPT de OpenAI y Bard de Google.

La súbita expansión de la IA generativa y el interés generalizado que se ha despertado entorno a los chatbots y otras herramientas de IA generativa, ha despertado un cierto sentimiento combinado de temor y neoludismo en muchos sectores profesionales que ven su futuro laboral comprometido ante la hipótesis (cada vez más verosímil) de que los robots empiecen a sustituirles. Un caso emblemático de esta desconfianza hacia la IA generativa es el de la huelga convocada el 2 de mayo de 2023 por el Sindicato de Guionistas de Estados Unidos (WGA, por sus siglas en Inglés), que representa a cerca de 12.000 escritores que trabajan en el mundo del espectáculo (teatro, cine y

[3] M. SIGMAN y S. BILINKIS, *Artificial. La nueva inteligencia y el contorno de lo humano*, Debate, Barcelona, 2023, p. 192.

televisión), y entre cuyas principales reivindicaciones estaba, precisamente, la de la regulación de la IA y la tecnología entrenada con conjunto de datos.

Muestra de lo fundada que está la inquietud de los guionistas y escritores de Hollywood, es un estudio encargado por la propia OpenAI sobre el impacto el mercado laboral de los grandes modelos del lenguaje (LLM, por sus siglas en Inglés), publicado el 17 de marzo de 2023, que situaba a los escritores en la categoría de profesionales "totalmente expuestos" ante la IA generativa. Según este informe, un LLM es capaz de reducir en un 50% el tiempo que tarda un humano en terminar su tarea, por ejemplo, un guión de cine o una novela[4]. La batalla de la productividad de los guionistas humanos frente a la IA generativa estaría perdida, pero no así el debate en torno a la originalidad y el crédito de los escritores, de tal forma que se reconozca al menos que ninguna empresa de entretenimiento esté autorizada legalmente para obligar a los creadores de contenidos que utilicen software de IA –por ejemplo, ChatGPT– para realizar su tarea.

Cinco meses después del inicio de la huelga convocada por el sindicato de guionistas, éstos consiguieron que en el quinto punto el acuerdo de negociación entre la WGA y los estudios de Hollywood se reconociera que "la IA no puede escribir ni reescribir material literario, y el material generado por la IA no se considerará material de origen", ya que una IA no puede considerarse un escritor[5].

Pese a la nula eficacia de la carta abierta en la que los signatarios pedían una pausa de la IA generativa a la vez advertían el riesgo potencial que supone para la humanidad el desarrollo de sus modelos más avanzados, solo dos meses más tarde, en mayo de 2023, un grupo de ingenieros, investigadores y altos ejecutivos de la industria de la IA volvieron a expresar su preocupación por la tecnología de IA que estaban desarrollando, sobre todo en relación con los modelos grandes de lenguaje, el tipo de sistema utilizado por ChatGPT y otros chatbots. Para los firmantes de esta declaración, la IA podría llegar a ser algún día tan poderosa como para ser utilizada para difundir desinformación y propaganda, para eliminar millones de puestos de trabajo tradicionalmente realizados por seres humanos, e incluso convertirse algún día en una auténtica amenaza existencial para la humanidad. Por ello, los signatarios decidieron hacer pública una declaración, publicada por el *Center for AI Safety*, en la que se instaba a:

[4] *https://openai.com/research/gpts-are-gpts*
[5] *https://www.wgacontract2023.org/the-campaign/summary-of-the-2023-wga-mba*

> *Mitigar el riesgo de extinción de la inteligencia artificial debería ser una prioridad mundial junto a otros riesgos a escala social, como las pandemias y las guerras nucleares*[6].

En esta frase, tan escueta como impactante, se alude a la necesidad de que haya un consenso acerca del peligro que se cierne sobre la humanidad, pero también a la dificultad de trasladar esta preocupación a acciones concretas y a ponerse de acuerdo sobre qué formas concretas podría tomar ese peligro y qué medidas podrían adoptarse para protegernos de las amenazas que entraña la IA[7].

Tanto la carta de los tecnólogos y expertos en IA de marzo, como la declaración de los líderes de la IA de mayo, ponen de manifiesto una inquietud compartida ante los millones de personas que, cada vez con mayor frecuencia, recurren a los chatbots como modo de entretenimiento, acompañamiento y/o aumento de la productividad, a medida que la tecnología subyacente mejora a un ritmo vertiginoso.

Para garantizar que la IA sea segura para la humanidad, tres altos ejecutivos de OpenAI: Sam Altman, Greg Brockman e Ylia Sutskever, han revelado que es *concebible* que la IA obtenga habilidades extraordinarias que superen a los humanos en la próxima década, lo cual nos ofrecerá la posibilidad de un futuro más próspero, pero para llegar a ese horizonte tan prometedor será preciso que sepamos gestionar el riesgo que comporta la IA. Para hacer viable este propósito, estos altos directivos han propuesto la creación de un organismo de control o agencia regulatoria internacional encargada de supervisar la IA general (en adelante, IAG). Esta agencia internacional de la IA sería comparable a la que existe para inspeccionar la producción de la energía atómica (el Organismo Internacional de la Energía Atómica); ahora bien, mientras que para producir la tecnología nuclear es necesario hacer grandes inversiones y

---

[6] *https://www.safe.ai/statement-on-ai-risk*: entre los firmantes de esta declaración se encuentran directores ejecutivos de empresas tecnológicas de IA, como Demis Hassabis (Google DeepMind), Sam Altman (OpenAI), Daniela y Dario Amodei (Anthropic) o Bill Gates; también hay ingenieros e investigadores, como Ray Kurzweil y Rob Pyke (Google); y profesores de Ciencias de la Computación, como Bart Selman (Universidad de Cornell), Vincent Conitzer y Philip Torr (Universidad de Oxford), James Mickens (Universidad de Harvard), Yoshua Bengio (Universidad de Montreal-Montreal Institute for Learning Algorithms), o Geoffrey Hinton (Universidad de Toronto), y otros muchos, entre los 350 firmantes iniciales de esta declaración, que ha seguido sumando apoyos hasta el presente.

[7] M. SIGMAN y S. BILINKIS, *Artificial. La nueva inteligencia y el contorno de lo humano*, cit., p. 169.

acceder a materiales tan difíciles de obtener, como el uranio enriquecido, gran parte de la tecnología de la IA es de código abierto, con lo que no hay restricciones significativas para que cualquier persona en el mundo pueda llegar a elaborar una bomba atómica informática en su propia casa[8].

Al margen del debate suscitado por los tecnólogos a raíz de su pronunciamiento en aras de la pausa y la cautela en el desarrollo de la investigación de la IA, con la IAG como objetivo en el horizonte, una de las principales conclusiones que se desprenden de esta controversia científico-tecnológica es la coincidencia, por parte de quienes están a favor del desarrollo sin límites de la IA, como de aquellos que mantienen una actitud más recelosa o prudente ante el avance de la tecnología de la IA, es que nos encontramos ante un proceso de transformación capaz de comprometer el futuro de la humanidad debido a su potencial desestabilizador[9].

El *homo tecnologicus* se encuentra ante una nueva ola tecnológica, impulsada por una tecnología de uso general con profundas implicaciones sociales, que le obligará a enfrentarse, en las décadas venideras, a las cuestiones más fundamentales que nunca se haya planteado nuestra especie. A propósito de los efectos transformadores de la nueva ola tecnológica, Mustafa Suleyman y Michael Bhaskar, sostienen en un libro reciente que, en la actualidad, nos encontramos ante un problema de *contención de la tecnología de la IA*, en cualquier fase de su desarrollo y distribución, y que, si en última instancia no se puede impedir que la tecnología prolifere, ni controlar su onda expansiva de consecuencias no deseadas (tanto si son positivas como negativas), al menos debería limitarse. Se trata, en definitiva, de un enorme reto ético-jurídico al que nuestra especie deberá enfrentarse más pronto que tarde.

> *Durante la mayor parte de la historia, el reto de la tecnología consistía en la creación y la liberación de su poder. Ahora es al revés: el reto radica en la contención del poder desatado y en garantizar que siga prestándonos servicio tanto a nosotros como al planeta*[10].

En suma, la necesidad desarrollar una IA al servicio del bienestar de la humanidad exige, de un lado, el establecimiento de un marco normativo

---

[8] Ibid., p. 194.

[9] B. CHRISTIAN, *The Alignment Problem. How Can Artificial Intelligence Learn Human Values?*, Atlantic Books, London, 2021, pp. 328-329.

[10] M. SULEYMAN y M. BHASKAR, *La ola que viene*, trad. esp. C. Fernández, Debate, Barcelona, 2023, p. 58.

adecuado que regule la economía de los datos y se adapte a una IA en permanente evolución; en este sentido, en el seno de la Unión Europea se ha ido conformando un auténtico *corpus iuris digitalis* que tiene ante sí el reto de aprovechar las ventajas de las tecnologías de IA para ganar en competitividad económica, e impulsar la prosperidad y el bienestar de sus ciudadanos, en un entorno neotecnológico seguro, fiable y compatible con los valores y principios que inspiran la Carta de los Derechos Fundamentales de la Unión Europea. Precisamente estos fueron los fundamentos que inspiraron el acuerdo de la Unión Europea, alcanzado el 8 de diciembre de 2023, tras aprobar el parlamento, la Comisión y el Consejo la primera Ley de IA de la historia. Este importante logro, ha supuesto un verdadero hito en la historia del derecho y de la ciencia, además ha servido para colocar a Europa en un lugar destacado para desempeñar un papel de liderazgo a nivel mundial en este ámbito (me refiero al llamado "efecto Bruselas"[11]), que servirá también para generar confianza en los usuarios de las aplicaciones de IA, para establecer mecanismos de control a las aplicaciones de IA de alto riesgo, para reforzar la seguridad, y garantizar los derechos de las personas y las empresas[12].

Desde un punto de vista jurídico, se han producido avances significativos en la regulación de una IA al servicio de la humanidad, sobre todo en el ámbito de la Unión Europea, que cuenta con la primera ley integral del mundo sobre IA y que, como parte de su estrategia digital, vela por la seguridad, la transparencia, la eticidad y el control humano de los sistemas de IA. Sin embargo, no puede soslayarse el hecho de que, aunque el bienestar de la humanidad sea uno de los principios que guían la normativa europea sobre la IA, es urgente e inaplazable el planteamiento de un debate ético a propósito del impacto y los efectos negativos producidos por los sistemas artificiales, así como la falta de asunción de responsabilidades por parte de quienes comercian y se enriquecen con dichos sistemas sin respetar códigos deontológicos ni normas de conducta que permitan diferenciar lo bueno de lo malo, y que favorezcan lo primero y eviten lo segundo[13].

---

[11] A. BRADFORD, "The Brussels Effect", *Northwestern University Law Review*, vol. 107, 2012, pp. 1-67.

[12] F. H. LLANO ALONSO, *Homo ex machina. Ética de la inteligencia artificial y Derecho digital ante el horizonte de la singularidad tecnológica*, Tirant lo Blanch, Valencia, 2024, p. 181.

[13] A. BIRHANE-J. VAN DIJK, "Robot Rights? Let´s Talk about Human Welfare Instead", *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES ´20)*, Association for Computing Machinery, New York, 2020, pp. 207-213; S. DEGLI-ESPOSTI, *La ética de la inteligencia artificial*, CSIC-Catarata, Madrid, 2023, p. 100.

2. LA IA FUERTE Y EL ALINEAMIENTO DE SUS OBJETIVOS CON LOS VALORES HUMANOS

Las leyes de la robótica de Isaac Asimov parten de una presunción: si el código ético de las máquinas y los valores humanos no estuvieran correctamente alineados, una IAG acabaría destruyendo, racional y consecuentemente, nuestro ecosistema y, con ello, la humanidad entera[14]. Por eso mismo, como advirtiera Norbert Wiener –considerado el fundador de la cibernética y un influyente defensor de la automatización– si para lograr nuestros objetivos utilizamos un agente mecánico (como un robot u otra máquina inteligente) deberíamos asegurarnos antes de hacerlo de que el objetivo que perseguimos con dicho agente es el que de verdad deseamos[15].

A esta hipótesis distópica, en la que una IAG no se encuentra alineada con la humanidad, apuntó Isaac Asimov cuando enunció, en 1941, las tres leyes de la robótica, publicadas un año más tarde en un relato corto titulado: *Runaround*, que ha inspirado no solo a otras generaciones ulteriores de escritores de ciencia ficción, sino a científicos cognitivos: Minsky (2006), ingenieros informáticos: Barfield (2015), y juristas: Sartor (1993) y Pagallo (2013), entre otros.

Las tres leyes clásicas de la robótica prohíben a los robots causar daño a los humanos, les obligan a obedecer las órdenes dadas por los seres humanos, y les ordenan preservar su propia existencia. Estas tres leyes establecen los siguientes principios:

1. *Un robot no hará daño a un ser humano ni, por inacción, permitirá que un ser humano sufra daño.*

2. *Un robot debe cumplir las órdenes dadas por los seres humanos, a excepción de aquellas que entren en conflicto con la primera ley.*

3. *Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la primera o con la segunda ley.*

Con posterioridad estas leyes originales fueron modificadas y desarrolladas por otros novelistas de ciencia ficción, e incluso el propio Asimov añadiría una cuarta ley (ley cero), en *Robots and Empire* (1985), para un imagi-

[14] S. DEGLI-ESPOSTI, *La ética de la inteligencia artificial*, cit., pp. 74-75.

[15] N. WIENER, *Cybernetics or Control and Communication in the Animal and the Machine*, The Technology Press-John Wiley and Sons-Inc. Hermann et Cie Editeurs, Cambridge (Massachusetts)-New York, Paris, 1948; *The Human Use of Human Beings: Cybernetics and Society*, Doubleday & Co., Garden City (New York), 1950.

nario universo futurista y distópico en el que las máquinas han asumido el gobierno del planeta tras desplazar a los hombres. Esta cuarta ley responde al siguiente enunciado:

4. *Un robot no puede dañar a la humanidad o, por inacción, permitir que la humanidad sufra daños.*

En cuanto a esta cuarta ley, contempla la hipótesis de que un robot pueda llegar a incluso a actuar contra un ser humano para evitar que haga daño a sus congéneres. Como señala Jacques Pitrat, esta ley cero es una ley pensada para proteger a la humanidad de su peor enemigo: el hombre[16]. A este respecto, en el relato *The Evitable Conflict* (1950) Asimov imagina un mundo controlado por las máquinas, pues éstas han llegado a la conclusión de que la única manera de satisfacer el mandato de la primera ley es conspirar contra los hombres para salvarles de sí mismos. El coordinador de ese gobierno planetario en el que las máquinas regulan la economía mundial se llama Stephen Byerley, un robot humanoide dotado de conciencia artificial que vela por el bien de la humanidad.

La hipótesis del conflicto entre máquinas inteligentes y personas, tantas veces llevada a la novela distópica y el cine de ciencia ficción, en la que la IA podría tomar decisiones contrarias al interés de los seres humanos, se plantea el llamado "problema del alineamiento": ¿cómo conseguir que las máquinas inteligentes no se revuelvan contra sus creadores? A este respecto, como han advertido Sigman y Bilinkis, ese potencial enfrentamiento no requeriría necesariamente una conciencia o maldad intrínseca por parte de la IA.

> *Es suficiente con que perdamos el control sobre algo sumamente poderoso, y capaz de operar a gran escala sobre el mundo, por un error en la función valor, por alguna omisión en las limitaciones que se le impongan, por alguna propiedad emergente inesperada o por la posibilidad que tienen de modificar los objetivos que les damos*[17].

Como puede comprobarse tras la lectura de los dos relatos de Asimov que hemos tomado como ejemplo, *Runaround* y *The Evitable Conflict*, la literatura de ciencia ficción ha anticipado situaciones e hipótesis que han pasado a formar parte de la experiencia jurídica de la era cibernética y precisan una específica regulación: la ética de la IA, la autonomía de las máquinas y su im-

[16] J. PITRAT, *Artificial Beings: The Concience of a Conscious Machine*, John Wiley & Sons Inc., Hoboken (New Jersey)-London, 2009, p. 184.

[17] M. SIGMAN y S. BILINKIS, *Artificial. La nueva inteligencia y el contorno de lo humano*, cit., p. 177.

pacto en el ámbito de los derechos humanos, la conciencia artificial, la personalidad electrónica o la responsabilidad civil y penal por los daños causados por la IA y la robótica[18].

A la vista de la complejidad de las leyes de la robótica, de los dilemas éticos de difícil solución que éstas presentan, de las consecuencias de los sesgos implícitos en los datos de entrenamiento de los algoritmos y de los desafíos que entraña el problema del alineamiento, sería oportuno abordar estas cuestiones dentro de un marco ético que limite y oriente las acciones y decisiones automatizadas de las máquinas inteligentes al interactuar con las personas. Pero, ¿cómo consensuar un código ético aplicable a las máquinas, cuando ni siquiera hemos logrado ponernos de acuerdo entre las personas a la hora de compartir un marco ético universal, ni unos mismos valores morales que orienten la conducta de los individuos? Y por otra parte, ¿con qué conjunto de valores debería alinearse la IA y qué estatus legal y ético debería tener?

La cuestión sobre el alineamiento de los valores y comportamientos de los sistemas de IA con los valores humanos se trató, precisamente, en la Conferencia de Asilomar sobre IA beneficiosa, celebrada en enero de 2017 en California y organizada por el *Future of Life Institute*. Este evento ético-tecnológico contó con la presencia de más de un centenar de expertos e investigadores que se reunieron para debatir y consensuar principios de IA. Esta conferencia concluyó con la formulación de veintitrés principios que suponen, para algunos, un valioso punto de referencia para los debates sobre ética de la IA, mientras que, para otros, carecen de una orientación específica sobre cómo aplicarlos a la práctica. En todo caso, aunque no sean suficientes, se pueden considerar como principios necesarios para iniciar abordar los complejos retos éticos que surgen en torno a la IA.

El décimo principio de la Conferencia de Asilomar versa sobre la alineación de valores y reza así:

> *Los sistemas de IA altamente autónomos deben ser diseñados de manera que sus objetivos y comportamientos puedan ser asegurados para alinearse con los valores humanos a lo largo de su operación*.

Como vemos, este principio se refiere fundamentalmente a los sistemas de IA autónomos, capaces de tomar decisiones automatizadas y llevar a cabo

[18] U. PAGALLO, *The Laws of Robots. Crimes, Contracts, and Torts*, Springer, Dordrecht-Heidelberg-New York-London, 2013, p. 23.

acciones sin intervención humana. En otras palabras, con este principio lo que se pretende es garantizar que los objetivos y comportamientos de los sistemas de IA autónomos estén diseñados de forma que se alineen con valores humanos, para que no actúen al margen de las normas establecidas, ni produzcan daños o lesiones a los seres humanos, tal y como prescribe la primera regla de Asimov.

Pero, ¿cómo conseguimos que los sistemas de IAG superinteligentes estén alineados con los valores humanos y sigan la intención humana sin rebelarse? En julio de 2023 OpenAI publicó un documento de investigación, firmado por Ilya Sutskever y Jan Leyke, codirectores del equipo *Superalignment*, en el que explicaban los resultados de su investigación sobre un modelo de supervisión automatizado que sea capaz de guiar el comportamiento de otro mucho más inteligente (este proceso es el que se utiliza, por ejemplo, por OpenAI para ajustar su modelo del lenguaje estrella en 2023: Chat GPT-4)[19].

El ajuste algorítmico a GPT-4 que está probando el equipo *Superalignment* permitiría,a la postre, que el modelo más potente e inteligente siga la guía del modelo más débil y menos inteligente, pero sin reducir por ello su rendimiento. En todo caso, los investigadores de OpenAI admiten que este novedoso método de superalineamiento no garantiza aún que el modelo más fuerte se comporte a la perfección, aunque puede servir como un punto de referencia del que partirán futuras investigaciones sobre el control de la IAG[20].

3. LA ÉTICA DE LA IA Y LA AMPLIACIÓN DE LOS PRINCIPIOS DE LA BIOÉTICA

En el anterior epígrafe nos ocupamos del alineamiento de los valores y comportamientos de los sistemas de IA con los valores humanos. Esta operación de alineación es conveniente realizarla desde el momento en que, a través de la IA, ha sido posible automatizar tareas y delegar al software o a máquinas inteligentes el desarrollo de determinadas funciones.

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 huma-

[19] *www.wired.com/story/openai-ilya-sutskever-ai-safety/*
[20] R. GÓMEZ PÉREZ, *El dilema de la IA*, Rialp, Madrid, 2023, p. 92.

no en plena era de la automatización. En definitiva, esta es la 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[21].

El objetivo de este enfoque ético es doble: por un lado, promover la concordancia o alineamiento entre las intenciones de las distintas partes 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 decisivos en relación con su ámbito de actuación. Este segundo objetivo (el de la vigilancia contra 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, en la medida que propicia el mantenimiento de un alto 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 sino, al contrario, es la receta para su para su éxito a largo plazo y para cosechar todos los beneficios que promete[22].

Sin embargo, para que haya una alineación entre las tecnologías y las expectativas éticas, éstas deben estar claramente definidas. Establecer 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 mencionar la gran labor ético-jurídica realizada a través de las declaraciones internacionales sobre los principios éticos de la IA. En este sentido, Luciano Floridi[23] ha destacado algunas que,

[21] J. SCHAICH BORG-W. SINNOTT-ARMSTRONG-V. CONITZER, *Moral AI and How We Get There*, Penguin Random House, Milton Keyes-Dublin, 2024, pp. 190-191.

[22] S. QUINTARELLI, *Intelligenza Artificiale. Cos´è davvero, come funziona, che effetti avrà*, Bollati Boringhieri, Milano, 2020, p. 84.

[23] L. FLORIDI, *Etica dell´intelligenza artificiale. Sviluppi, opportunità, sfide,* Raffaelo Cortina Editore, Milano, 2022, pp. 94-95.

a su juicio, serían las más influyentes hasta ahora en la definición del marco ético-jurídico de la IA; estas serían las tres principales:

- Los *Principios de Asilomar*, un conjunto de 23 principios desarrollados bajo los auspicios del Future Life Institute, en colaboración con los participantes en la Conferencia Asilomar que se celebró 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, y por supuesto, también 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 todos tengan acceso al uso y el beneficio de la IA).
- La *Declaración de Montreal para un desarrollo responsable de la Inteligencia Artificial* (2018) en la que se enuncian una serie de diez principios que sientan las bases para fomentar la confianza de la sociedad en los sistemas de IA. El sexto principio que se proclama en esta declaración 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 el 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 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; i) sostenibilidad.

A diferencia de los Principios de Asilomar y de la Declaración de Montreal, en la Declaración de la EGE se hace referencia expresa en el apartado d) a los sesgos discriminatorios y se 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". De ahí que al final del primer párrafo de este cuarto parágrafo se proclame 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 en la creación de un marco jurídico específico sobre IA. En este sentido, el Parlamento Europeo aprobó una resolución, el 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. En dicha resolución, que marcó un hito en el comienzo de la normativa de la Unión Europea sobre IA, se 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*[24].

A propósito de las declaraciones de principios de la IA, Floridi ha advertido que existe una suerte de hipertrofia de principios éticos que se suceden cada vez que una organización aprueba una declaración aún a riesgo de incurrir en innecesarias repeticiones o superposiciones de principios proclamados en otras declaraciones precedentes, lo cual también conduce a la confusión y la ambigüedad.

[24] 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)), parágrafos 1 y 20.

Frente a esta profusión de principios éticos, Floridi nos propone una idea de IA que no debe ser entendida como un nuevo tipo de inteligencia, sino como una forma de actuar sin precedentes. Por eso, de todas las áreas de la ética aplicada, la bioética es la que más se parece a la ética digital en el tratamiento que ésta hace de nuevas formas de agentes, pacientes y entornos[25]. 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 respuesta a la pregunta: "¿Cómo funciona?"), como el sentido ético de responsabilidad (accountability) (como respuesta a la pregunta: "¿Quién es responsable del modo en que funciona?")[26].

Estos cinco principios son válidos tanto para expertos, como por ejemplo diseñadores o ingenieros de productos, como para los no expertos, como pacientes o clientes[27]. A continuación comentaremos, sucintamente y por separado, cada uno de estos cinco principios:

1. Beneficencia

De los cuatro principios tradicionales de la bioética es el más fácil de observar por su claridad. La mayoría de las declaraciones de principios de la IA coinciden en interpretar este término como "bienestar" de los seres humanos y de todas las criaturas sintientes (Declaración de Montreal), y la conciben como un principio-guía de la IA hacia el bien común, el beneficio de la humanidad y del planeta (Conferencia de Asilomar).

2. No maleficencia

Este principio complementa a la IA benéfica en la medida que avisa contra las diversas consecuencias negativas que se derivan del uso abusivo e impropio de las tecnologías de IA, como las violaciones contra la intimidad personal. En este sentido, la Declaración de Montreal sostiene que quienes diseñan y desarrollan los sistemas de IA deben asumir la responsabilidad

[25] L. FLORIDI, *The Ethics of Information*, Oxford University Press, Oxford, 2013.

[26] L. FLORIDI, *Etica dell´intelligenza artificiale. Sviluppi, opportunità, sfide,* cit., p. 96.

[27] D. S. WATSON-L. FLORIDI, "The explanation game: A formal framework for interpretable machine learning", *Synthese*, 2020, pp. 1-32.

correspondiente actuando contra los riesgos derivados de la innovación tecnológica. Sin embargo, en el noveno principio de esta declaración (el de responsabilidad) se advierte que, cuando un sistema de IA provoca daños o perjuicios y se demuestra que el sistema de IA es responsable a pesar de haber sido utilizado según lo previsto, no es razonable culpar a las personas involucradas en su desarrollo y uso.

3. Autonomía

Este principio se relaciona con el anterior en la medida que la autonomía funcional de la máquina inteligente implica una suspensión del control humano en la realización de una determinada tarea automatizada. No parece lícito defender que el usuario sea considerado responsable de los eventuales efectos negativos causados por el funcionamiento del sistema. En otras palabras: la autonomía funcional puede considerarse, como causa de la erosión del control del funcionamiento del sistema, una buena razón para la limitación de responsabilidad de los individuos implicados en su diseño y desarrollo (programadores, constructores, etc…)[28].

Ahora bien, si ningún actor humano puede ser considerado plenamente responsable de los efectos causados por el funcionamiento de la IA, y si no basta tampoco con imputarle la culpa al sistema de IA mismo, ¿cómo se puede resolver 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[29].

4. Justicia

El principio de justicia, pese a su importancia, es el que más acepciones presenta en las diferentes declaraciones de principios de la IA. Mientras que en la Declaración de Montreal se 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"

[28] T. M. POWERS-J. G. GANASCIA, "Autonomy", *The Oxford Handbook of Ethics of AI* (ed. M. D. Dubber-F. Pasquale-S. Das), Oxford University Press, Oxford, 2021, pp. 31-32.

[29] S. QUINTARELLI, *Intelligenza Artificiale. Cos´è davvero, come funciona, che effetti avrà*, cit., p. 90.

(sexto principio); en los Principios de Asilomar (concretamente en el decimoquinto) se alude a la justicia como prosperidad económica compartida; por lo demás, en otras declaraciones el término justicia tiene otros significados (por ejemplo en el sentido de 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[30].

5. Explicabilidad

Para salir de esta confusión debida a la diversidad de significados divergentes del 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 incorporación 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 puzzle 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 se 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 ser informada por un conocimiento previo: cómo actuaría la IA en nuestro lugar y, en tal caso, cómo podrían mejorarse sus prestaciones; y, por último, para que la IA sea justa, necesitamos saber a quién responsabilizar ética o jurídicamente en caso de un resultado grave o negativo, lo que a su vez requeriría una comprensión adecuada de por qué se produjo tal resultado[31].

4. DIVERSOS ENFOQUES Y MODELOS ÉTICOS DE LA IA

La ética de la IA puede abordarse desde diversos enfoques, cada uno de ellos centrado en diferentes aspectos éticos relacionados con el desarrollo, la implementación y el uso de la IA. Entre los distintos enfoques éticos de la IA cabe diferenciar al menos cuatro perspectivas, dependiendo de cuál sea su

[30] L. FLORIDI, *Etica dell´intelligenza artificiale. Sviluppi, opportunità, sfide,* cit., p. 100.
[31] Ibid.

interés o particular objeto de análisis, ya sea el usuario, el diseñador, la sociedad, el dato o el algoritmo.

Tanto la ética de la IA del dato, que se centra en cómo se recopilan, utilizan y gestionan los datos en los sistemas de IA, como la ética que se ocupa del diseño ético de los algoritmos, comparten principios éticos fundamentales para guiar el desarrollo, la aplicación y el uso de la IA de manera ética y responsable. Dicho esto, conviene advertir que, pese a que no existe todavía un consenso universal sobre los principios éticos de la IA, el proceso de confluencia en torno a dichos principios irá avanzando a medida que las organizaciones de expertos en ética, científicos e investigadores, tecnólogos, empresas de IA, comités y grupos de trabajo colaboren para conseguir la formulación de unos mínimos principios y estándares éticos universales a propósito de las tecnologías convergentes (tecnologías como las que han permitido, por ejemplo, la convergencia entre la IA, la robótica o el análisis de datos para el desarrollo de sistemas y vehículos autónomos).

Respecto a los principios-guía comunes a algunos de los códigos éticos para el desarrollo, despliegue y buen uso de la IA cabe mencionar los siguientes:

1. Transparencia algorítmica

Este principio exige la visibilidad, la cognoscibilidad, la auditabilidad y la explicabilidad de los factores intervinientes en las decisiones tomadas con algoritmos a las personas que utilizan, regulan, y son afectadas por los sistemas de IA que emplean dichos algoritmos[32].

La transparencia significa además permitir que las personas comprendan cómo se desarrolla, entrena, opera y despliega un sistema de IA en el dominio de aplicación relevante, de modo que los consumidores, por ejemplo, puedan tomar decisiones más informadas. El principio de transparencia también se refiere a la capacidad de proporcionar información significativa y claridad sobre qué información se proporciona y por qué[33].

[32] A. SANTANGELO, "Equità degli algoritmi e democracia", *DigitCult*, 2020, vol. 5, pp. 21-30 (*http://dx.doi.org/10.53136/979125994120634*); G. VESTRI, "Transparencia algorítmica", en *Diccionario de términos para comprender la transformación digital*, G. VESTRI (Dir.)-M. CASTILLA BAREA (Coord.), Aranzadi, Cizur Menor (Navarra), 2023, pp. 346-348.

[33] Principio 1.3 de la lista de principios de IA de la OCDE, adoptados en mayo de 2019.

2. Equidad

Entendiendo la justicia como equidad, lo que se pretende con este principio es que los algoritmos predictivos que resultan decisivos en la toma de decisiones y que tanto condicionan la vida de las personas (por ejemplo, en un proceso de selección de candidatos en una oferta de trabajo, en la búsqueda de receptores de un trasplante, en la aprobación de préstamos, en la evaluación de riesgos, etc…) sean diseñados evitando la discriminación y los sesgos en la toma de decisiones basadas en datos. En definitiva, el cumplimiento de este principio garantiza un trato justo y equitativo para todos los usuarios[34].

3. Responsabilidad

La generalización de los sistemas de aprendizaje automático basados en algoritmos en muchos sectores ha sacado a la luz el problema de la responsabilidad por "fallo de algoritmo". Los sistemas de aprendizaje automático basados en algoritmos permiten a las máquinas mejorar su rendimiento con el tiempo; sin embargo, en caso de error, las consecuencias para las personas y los bienes pueden ser especialmente graves. La complejidad de estos sistemas, sin embargo, hace difícil comprender a quién se puede responsabilizar de estos daños. A propósito de la responsabilidad de los sistemas expertos basados en IA, algunos expertos en la materia han señalado que

> *la atribución de «sujeto de derecho» a los efectos, en su caso, de imputar el daño ocasionado a un comportamiento o actividad debería referirse a sistemas expertos con capacidad de aprendizaje que actúan de forma completamente autónoma respecto de la persona que responda por ellos. De otra parte, la regla de la responsabilidad proporcional en materia de causalidad en relación con los problemas de incertidumbre que se generan en un entorno operado por sistemas expertos con o sin interacción con humanos debe tenerse en cuenta a la hora de plantearse una posible regulación de la responsabilidad civil por los daños ocasionados*[35].

---

[34] M. MORENO REBATO, "Discriminación algorítmica", en *Diccionario de términos para comprender la transformación digital*, cit., pp. 141-143. S. VANTIN, "Inteligencia Artificial y derecho antidiscriminatorio", en *Inteligencia Artificial y derecho. El jurista ante los retos de la era digital,* F. H. LLANO ALONSO (ed.), Thomson Reuters Aranzadi, Cizur Menor (Navarra), 2021, pp. 367-384.

[35] S. NAVAS NAVARRO, "Sistemas expertos basados en inteligencia artificial y responsabilidad civil", *Diario La Ley*, 13 de diciembre de 2021; en la misma línea, vid., M. MARTÍN

4. Privacidad

Existe una compleja intersección entre la privacidad, la protección de datos y la ética de la IA. En su reciente libro *The Ethics of Privacy and Surveillance*, Carissa Véliz propone una definición mixta de privacidad en la que se hibridan las dos teorías principales entorno a este concepto: de un lado, la teoría del *acceso*, según la cual, la cuestión de la privacidad consiste fundamentalmente, bien en la accesibilidad limitada, o bien en la inaccesibilidad a la información sensible (en este sentido, se entiende que perdemos privacidad cuando la información que afecta a nuestros datos personales es accesible para los demás) por otro lado, la teoría del *control* de privacidad centra su atención en la protección de nuestros datos personales ante la intromisión o injerencia de actores externos[36].

A propósito de la protección de datos, dentro del marco regulatorio europeo, en el artículo 5 del Reglamento General de Protección de Datos (RGPD)[37] se establecen los principios relativos al tratamiento de los datos personales, entre los que cabe resaltar los requisitos de *licitud*, *lealtad* y *transparencia* en el tratamiento de datos (art. 5.1 a), además de los principios de *finalidad* (art. 5.1 b) y *minimización de datos* (art. 5.1 c)[38].

CASALS, "Proportional Liability in Spain", en M. MARTÍN CASALS-D. M. PAPAYANNIS, *Uncertain Causation in Tort Law*, Cambridge University Press, Cambridge, 2016, pp. 43-66.

[36] C. VÉLIZ, *The Ethics of Privacy and Surveillance*, Oxford University Press, Oxford, 2024, pp. 74-75.

[37] Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016.

[38] El tratamiento de datos se entenderá lícito si se cumple al menos alguna de las siguientes condiciones previstas en el art. 6 RGPD: a) Cuando el interesado ha dado su consentimiento para el tratamiento de sus datos personales para uno o varios fines específicos; b) cuando el tratamiento es necesario para la ejecución de un contrato en el que el interesado es parte o para la aplicación a petición de este de medidas precontractuales; c) cuando el tratamiento es necesario para el cumplimiento de una obligación legal aplicable al responsable del tratamiento; d) cuando el tratamiento es necesario para proteger intereses vitales del interesado o de otra persona física; e) cuando el tratamiento es necesario para el cumplimiento de una misión realizada en interés público o en el ejercicio de poderes públicos conferidos al responsable del tratamiento; f) cuando el tratamiento es necesario para la satisfacción de intereses legítimos perseguidos por el responsable del tratamiento o por un tercero, siempre que sobre dichos intereses no prevalezcan los intereses o los derechos y libertades fundamentales del interesado que requieran la protección de datos personales, en particular cuando el interesado sea un niño.

Con respecto a los principios de *limitación de finalidad* y *minimización de datos*: el primero implica que los datos personales han de ser recogidos con fines determinados, explícitos y legítimos, y que no serán tratados ulteriormente de manera incompatible con dichos fines (a

5. Seguridad

Al igual que sucede con la interrelación entre el principio de responsabilidad y la protección de datos, también existe una interacción entre el principio de seguridad y la solidez de los sistemas de IA.

En efecto, tanto en el apartado 1.4 de la lista de principios de la IA de la OCDE, como en el artículo 15 del Reglamento de Inteligencia Artificial, "solidez" es sinónimo de capacidad de resistir o superar condiciones adversas, incluidos los riesgos de la seguridad digital. Este principio establece además que los sistemas de IA no deben plantear riesgos que no sean razonables para la seguridad, incluida la seguridad física, en condiciones de uso normal o previsible o de uso indebido a lo largo de su ciclo de vida[39].

Hay dos maneras de mantener sistemas de IA sólidos, seguros y protegidos: en primer lugar, mediante su *trazabilidad* y posterior análisis e indagación; y, en segundo lugar, aplicando un *enfoque de gestión de riesgos*.

La trazabilidad puede ser útil para realizar el análisis y la investigación de los resultados de un sistema de IA y es una forma de promover la rendición de cuentas; además, puede ayudar también a prevenir futuros errores y a mejorar la fiabilidad del sistema de IA.

En relación con el enfoque de gestión de riesgos, aplicado a lo largo de todo el ciclo de vida del sistema de IA, éste puede ser de gran utilidad para identificar, evaluar, priorizar y mitigar los riesgos potenciales que pueden afectar negativamente al comportamiento y los resultados de un sistema.

5. A MODO DE CONCLUSIÓN: HACIA UNA METAÉTICA DE LA IA

Como se ha podido comprobar, la ética de la IA es un tipo de ética aplicada a los negocios y la tecnología, pero no deja de ser un subconjunto de la Ética (con mayúsculas). La ética de la IA se ocupa de identificar y dar respuesta a las problemáticas morales que se derivan de la implantación y el

excepción de los fines de archivo en interés público, de investigación científica e histórica o estadísticos que, de conformidad con lo establecido en el art. 89. 1 RGPD, no se considerarán incompatibles con los fines iniciales); el segundo principio exige que los datos personales sean adecuados, pertinentes y limitados a lo necesario en relación con los fines para los que son tratados.

[39] *https://oecd.ai/en/dashboards/ai-principles/P8*

uso de soluciones de IA. La autonomía, la justicia, la explicabilidad, la transparencia de los algoritmos, la equidad, la responsabilidad, la privacidad, la seguridad…, son algunas de las principales áreas de interés y preocupación por parte de expertos tecnólogos, filósofos y juristas, entre otros especialistas en la materia.

Paradójicamente, a pesar de la importancia y la necesidad de la ética de la IA, ésta puede ser utilizada a veces de forma irresponsable o poco ética por algunas grandes empresas o compañías tecnológicas (un caso paradigmático es el de Timnit Gebru, a quien Google contrató en 2020 como codirectora de un grupo de investigación centrado en la IA ética, pero a la que despidió meses después por haber criticado los riesgos éticos de los modelos del lenguaje de la compañía para la que trabajaba).

En general, la tendencia actual de los gigantes tecnológicos (que responden al acrónimo GAFAM: Google, Amazon, Facebook, Apple y Microsoft) es la de ir reduciendo plantilla de sus equipos de ética de la IA (Elon Musk, por ejemplo, despidió a la mitad de los trabajadores de Twitter cuando, en octubre de 2022, adquirió esta red social, incluido su pequeño equipo de IA ética; otro caso revelador fue el de Microsoft, compañía tecnológica que, en marzo de 2023, cerró el equipo de "ética y sociedad" de su división de IA, entre otros motivos porque, según John Montgomery, uno de los vicepresidentes de la división de IA, el objetivo de la compañía era que los modelos conversacionales de OpenAI llegasen a los clientes a la mayor velocidad posible, sin reglas ni restricciones éticas que retrasasen su entrega).

Frente a estos intentos de rentabilizar la ética para blanquear estrategias de mercado (*ethics washing*) con un aparente compromiso de buena praxis y respeto a los principios y las reglas éticas de la IA, pero que en realidad solo sirven para ocultar la intención de dirigir dichas estrategias al aumento de la competitividad de la empresa y a una mejora en su cuenta de resultados, sería oportuno recurrir a una ética de la ética de la IA (es decir, a una *metaética de la IA* que se centre en la fundamentación de los juicios normativos o de valor)[40]. Por eso, más que de una sola ética de la IA, tal vez lo apropiado sea hablar en plural, esto es, de éticas de la IA (la metaética de la IA y la ética aplicada al mundo de las empresas tecnológicas).

Entre las principales teorías éticas en las que se basan los códigos morales que sirven para el diseño y el desarrollo de la IA y la robótica podrían desta-

[40] I. G. R. GALVILÁN, *Robots en la sombra: RPA, robots conversacionales y otras formas de automatización cognitiva*, Anaya, Madrid, 2021.

carse, entre otras, el utilitarismo (Bentham, Mill), el deontologismo (Kant), la ética de la virtud (Aristóteles, Tomás de Aquino), la ética del contrato social (Locke, Rousseau, Rawls). Esta metafísica de la IA será objeto de estudio especializado en un trabajo posterior.

Fernando Llano Alonso
Facultad de Derecho
Universidad de Sevilla
Campus Ramón y Cajal
c/Enramadilla
41018 Sevilla
e-mail: llano@us.es