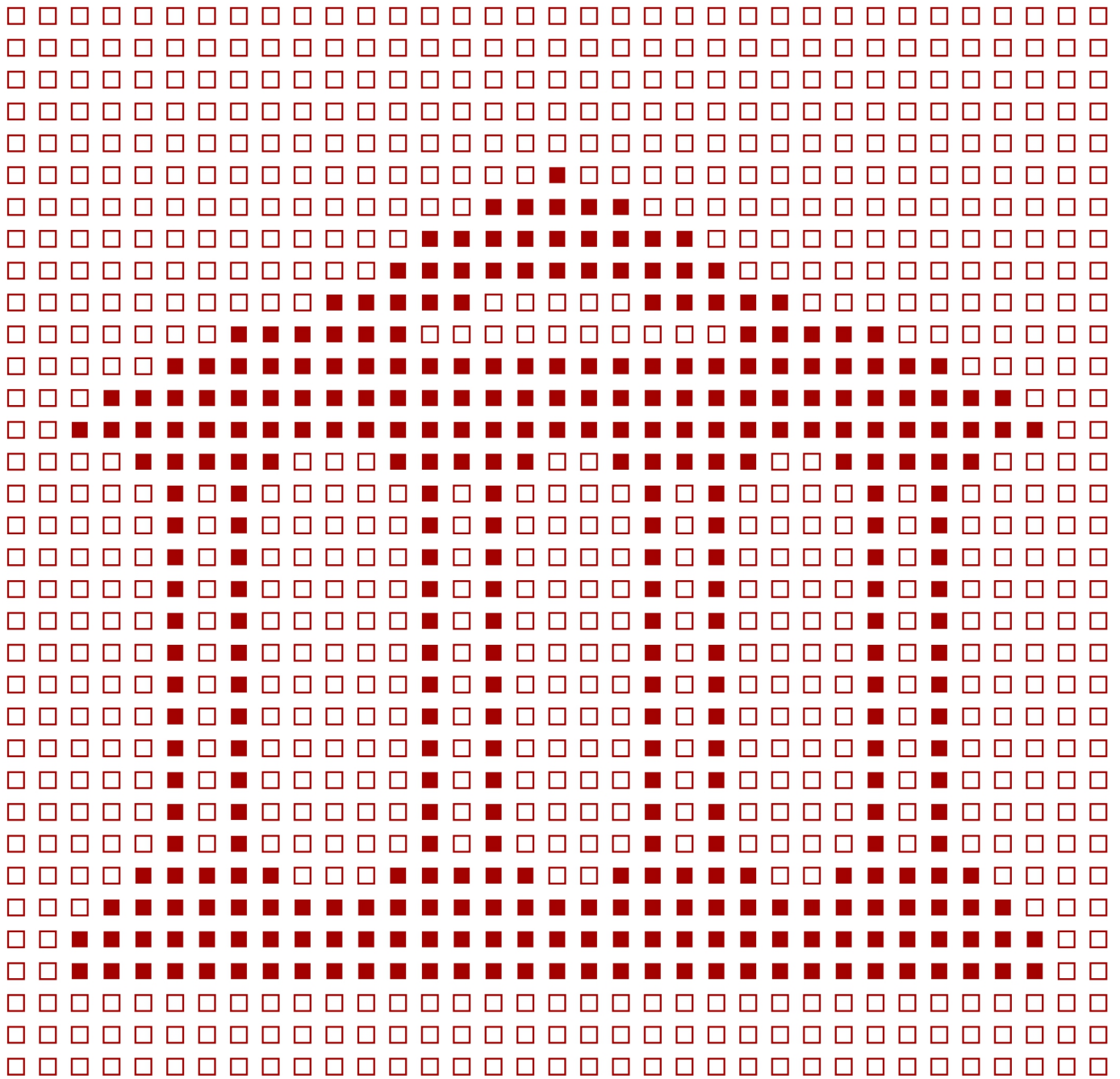




IAcracia

Cómo la inteligencia artificial
gobierna nuestra vida

Contribuciones al III Congreso de Ingeniería y Filosofía
Universidad de Oviedo, 2025



Universidad de Oviedo

IACRACIA: CÓMO LA INTELIGENCIA ARTIFICIAL GOBIERNA NUESTRA VIDA

CONTRIBUCIONES AL III CONGRESO DE INGENIERÍA Y FILOSOFÍA
UNIVERSIDAD DE OVIEDO, 2025

BEATRIZ RAYÓN VIÑA (Dir.)

COORDINACIÓN EDITORIAL: JORGE COQUE Y PABLO REVUELTA SANZ

DISEÑO DE PORTADA: RAMÓN RUBIO GARCÍA





Reconocimiento-No Comercial-Sin Obra Derivada (by-nc-nd): No se permite un uso comercial de la obra original ni la generación de obras derivadas.



Usted es libre de copiar, distribuir y comunicar públicamente la obra, bajo las condiciones siguientes:



Reconocimiento – Debe reconocer los créditos de la obra de la manera especificada por el licenciador:

Edición: Beatriz Rayón Viña, Jorge Coque Martínez y Pablo Revuelta Sanz (2025) *Iacracia: cómo la inteligencia artificial gobierna nuestra vida*. Oviedo. Universidad de Oviedo.

La autoría de cualquier artículo o texto utilizado del libro deberá ser reconocida complementariamente.



No comercial – No puede utilizar esta obra para fines comerciales.



Sin obras derivadas – No se puede alterar, transformar o generar una obra derivada a partir de esta obra.

© 2025 Universidad de Oviedo

© Lxs autorxs

Algunos derechos reservados. Esta obra ha sido editada bajo una licencia Reconocimiento-No comercial-Sin Obra Derivada 4.0 Internacional de Creative Commons.

Se requiere autorización expresa de las personas titulares de los derechos para cualquier uso no expresamente previsto en dicha licencia. La ausencia de dicha autorización puede ser constitutiva de delito y está sujeta a responsabilidad.

Consulte las condiciones de la licencia en: <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode.es>

Servicio de Publicaciones de la Universidad de Oviedo

ISNI: 0000 0004 8513 7929

Edificio de Servicios - Campus de Humanidades
33011 Oviedo - Asturias
985 10 95 03 / 985 10 59 56
servipub@uniovi.es

<https://publicaciones.uniovi.es/>

ISBN: 978-84-10135-83-3

ÍNDICE

Introducción. A modo de síntesis. IAcracia: hacia la sumisión algorítmica

Beatriz Rayón Viña..... 5

I. JUSTICIA EN LA ERA DE LA INTELIGENCIA ARTIFICIAL

Inteligencia artificial policial y judicial, con especial atención al sistema VioGén

Miguel Ángel Presno Linera..... 10

IA para simular las prestaciones de los y las jueces en la mejora del rendimiento de las oficinas judiciales italianas

Paolo Inturri 28

¿Qué decisiones no deberíamos dejar en manos de la inteligencia artificial?

Daniela Morales Lopes..... 40

¿Qué hacer frente al peak data?

Manuel García Domínguez y Lorién Jiménez Martínez 51

Hacia una inteligencia artificial responsable: propuesta para un Centro de Excelencia en IA Responsable

Jaime Morán..... 63

II. EL USO DE LA IA EN EL CAMPO DE LA SALUD

Participación ciudadana e Inteligencia artificial en el ámbito sanitario

Alejandra Rivas Carrero 76

El impacto psicológico de la IA

Carolina Tomás Franco 88

Sobre la cuestión de la Inteligencia Artificial y la soledad en relación con la tercera edad

Candela Carolina Fernández Álvarez..... 97

III. LECTURAS ÉTICAS DE LA IA

IA, narrativa y fascismo

Pablo Revuelta Sanz.....	106
--------------------------	-----

Hacia una IA feminista: una crítica a la veracidad de las imágenes deepfake

Irene Adán Sánchez-Infantes.....	120
----------------------------------	-----

Relaciones y lazos con la IA, efectos en las relaciones interpersonales

Miguel Enrique Naredo Rojas	131
-----------------------------------	-----

Responsabilidad e inteligencia artificial

Patricia Fernández Fernández	142
------------------------------------	-----

La importancia ética de las inteligencias artificiales

Juan Villazón	151
---------------------	-----

MATERIALES DIDÁCTICOS

Test de Touring. Taller práctico.

Pablo Revuelta Sanz, Beatriz Rayón Viña, Susana Valvidares Fernández	162
--	-----

<i>Agradecimientos</i>	168
------------------------------	-----

Introducción

A modo de síntesis. IAcracia: hacia la sumisión algorítmica

Beatriz Rayón Viña

Grupo IF

Universidad de Oviedo

Si bien no es una tecnología nueva, los sistemas de IA han cobrado relevancia en un breve espacio de tiempo. Este fenómeno ha llevado a replantear las actividades en diferentes escenarios, particularmente, laborales, informacionales, de seguridad y en la salud, convirtiéndose también en una fuente de preocupación y desconcierto entre la ciudadanía. Por ello, la implementación social de los sistemas basados en IA ofrece promesas, pero también desafíos. En el ámbito de la salud, por ejemplo, facilita desde diagnósticos más precisos y accesibles que permiten abaratar costos y ampliar la medicina de precisión, hasta nuevas terapias psicológicas basadas en modelos de lenguaje. Sin embargo, también surgen preocupaciones éticas y sociales, como los efectos negativos en la salud mental de adolescentes vulnerables, o los sesgos perjudiciales en herramientas de diagnóstico que pueden discriminar a ciertos colectivos. El enfoque principal de este congreso ha sido plantear cuestiones de implementación en ámbitos como la ética, la justicia y la salud. ¿Qué criterios y principios debemos usar para determinar qué no puede -o debe- hacer una IA? ¿Y quién es responsable

de lo que hacen estos sistemas? Determinar qué es lo justo con respecto al desarrollo de las IA es una cuestión muy complicada. ¿Hasta qué punto deberíamos aplicar el principio de precaución? ¿Qué riesgos implican las bases de datos que perpetúan sesgos? Por otra parte, a nivel de I+D+i, ¿tiene sentido limitar la IA en algunas regiones como la Unión Europea mientras otros países avanzan sin restricciones? Asimismo, a la hora de manejar información delicada, ¿debemos permitir el uso de sistemas de IA opacos, como "cajas negras"? ¿Cómo gestionamos los riesgos de desinformación amplificados por la IA? Y si la IA "piensa" y "crea", ¿debería poder ganar un concurso artístico? ¿Debemos pensar en los futuros derechos de las IA como tecnologías potencialmente autónomas? Estas y otras preguntas similares son las que se buscó plantear para reflexionar y seguir cuestionándonos en este ámbito.

Aunque podamos decir que la IA no habla, no piensa o no siente, las personas sí hablamos con ella, nos ayuda a pensar, a estructurar nuestro pensamiento, y nos produce alegría si funciona bien para lo que necesitamos y frustración cuando no nos da los resultados deseados. Al tratarse de una tecnología inscrita en nuestras sociedades, es pertinente analizar sus posibles virtudes y defectos, ya que afecta al tejido sociopolítico y jurídico, a nuestra salud, a nuestras economías y a nuestro medio ambiente. La IA, que se ha popularizado en los últimos años con la llegada de la inteligencia artificial generativa de manera masiva, no ha de ser entendida como una tecnología neutral, sino como una tecnología profundamente disruptiva que afecta a múltiples cuestiones, que han de ser analizadas y se ha de buscar todas las maneras posibles de educarnos y divulgar sobre ella.

Bajo la perspectiva del eje de justicia, el punto de partida ha de ser comprender que, efectivamente, las decisiones técnicas tienen consecuencias, y que necesitamos mucha pedagogía para entender y decidir qué hacemos con esta herramienta. Durante el congreso se nos invitó a reflexionar sobre la responsabilidad de las personas usuarias, pero también sobre los peligros inherentes de las cajas negras, la falta de explicabilidad del algoritmo y si la supervisión humana es suficiente para implementar la IA en sistemas sociales

como el de la justicia, la salud o el trabajo. La Ley de IA de la UE, analizada en este congreso, nos plantea muchas cuestiones sobre los siguientes pasos a dar como sociedad para que esta herramienta no se convierta en un problema. Algunos comunicantes nos expresaron que el peligro no es la IAcracia —es decir, que la IA se pueda o vaya a volver una suerte de *Terminator* que nos extermine—, sino los malos usos que podamos darle las personas. Sin embargo, generamos lazos con la tecnología, y podemos cuestionarnos sobre las responsabilidades y las culpas de una herramienta que se presenta como autónoma y que podemos dejar que nos sustituya en nuestras decisiones, de forma o no adecuada. Esto nos lleva de nuevo a la cuestión de la responsabilidad y la culpa, cuestiones que vuelven a conectarse con las cajas negras y la falta de explicabilidad. La IA, además, reproduce muchos sesgos que tenemos como personas y que nos gustaría eliminar, pero hemos visto que herramientas como el activismo pueden ayudar a concienciar sobre estas problemáticas. También se presentaron ejemplos prácticos de la IA aplicada a sistemas judiciales y los distintos dilemas éticos, e incluso problemas de inconstitucionalidad, que pueden surgir. Desde una perspectiva tecnooptimista, los problemas que genera la IA podrán ser solucionados con más tecnología y con profesionales excelentes que puedan desarrollar una buena herramienta. Sin embargo, también se presentaron problemáticas como las repercusiones medioambientales y la correspondiente necesidad de perspectivas de decrecimiento. Otro impacto analizado ha sido el problema de transformarnos en una sociedad de la posverdad, amplificada por la increíble capacidad de esta tecnología de acelerar y ampliar la creación y difusión de irrealidades, con el consiguiente riesgo de generar una nueva narrativa que moldee nuestra realidad.

Aplicando la IA a la salud, hubo un consenso en que la IA debe ser vista como una herramienta para la medicina, pero nunca como un sustituto del personal sanitario con buena formación. Se planteó la implementación de la IA en salud bajo el prisma de la obligación bioética de la beneficencia, y que debemos usar esta tecnología para el diagnóstico si sabemos que va a ayudarnos a obtener uno mejor. Sin embargo, también se puso sobre la mesa la necesidad de la comprensión humana y contextual de los diagnósticos, y la necesidad de crear

comunidades de pares extendidas que aborden de manera completa los problemas de salud. Se abordó asimismo la perspectiva feminista en la IA, necesaria y transversal a todas las cuestiones que plantea el grupo IF, tratada en el caso de la salud desde la reflexión sobre los cuidados.

A su vez, los trabajos del aula de extensión, de excelente calidad, trajeron al frente cuestiones como el edadismo y el aislamiento por la excesiva tecnologización de la sociedad, y un escenario paradójico de jueces artificiales que puedan, de forma recurrente, juzgar a otros jueces artificiales.

Desde la perspectiva técnica, la IA fue presentada como una herramienta de automatización que nos puede ayudar a conseguir tiempo libre, pero también hablamos del impacto de las TIC en la atención y del extractivismo emocional asociado a pasarnos horas en redes sociales. También hablamos de software libre y de soberanía tecnológica, de responsabilidad social corporativa extendida a la inteligencia artificial, y de la posibilidad de utilizar datos sintéticos para mejorar los resultados de esta herramienta; aunque también se señalaron sus limitaciones al no disponerse de una regulación insuficiente.

En definitiva, el III Congreso IF: IAcracia, hacia la sumisión algorítmica ha contenido una pluralidad de puntos de vista, disciplinas y debates muy enriquecedores para este gran tema que es la expansión de la inteligencia artificial en el tejido social que nos rodea en la actualidad, algunos de los cuales han sido recogidos en este volumen y se pueden leer a continuación.

I. JUSTICIA EN LA ERA DE LA INTELIGENCIA ARTIFICIAL

Inteligencia artificial policial y judicial, con especial atención al sistema *VioGén*¹

Miguel Ángel Presno Linera

Universidad de Oviedo

RESUMEN

La algoritmización y, en especial, la inteligencia artificial (IA) están cada vez más presentes en el ámbito de las investigaciones policiales y en el de la Administración de Justicia, tanto en los procesos penales como en los demás; en particular, en España se están usando sistemas predictivos para tratar de prevenir la reiteración de los casos de violencia de género (*VioGén*) y como apoyo en la concesión de permisos penitenciarios en la Administración catalana (*Riscanvi*). No estamos, en rigor, ante sistemas de IA si atendemos a la definición de la misma que ha adoptado el reciente Reglamento europeo de inteligencia artificial, que requiere capacidad de autoaprendizaje, pero no es descartable que acaben transformándose en sistemas de IA o que sean reemplazados por sistemas de esa naturaleza; lo que, por los potenciales efectos que su aplicación de los sistemas de IA puede tener en el Estado de Derecho, las libertades individuales y el derecho a la tutela judicial efectiva, los convertiría en lo que el citado Reglamento europeo de inteligencia

¹ Este trabajo es uno de los resultados del Proyecto PID2022-136548NB-I00 “Los retos de la inteligencia artificial para el Estado social y democrático de Derecho”. <https://retina-der.uniovi.es/>

artificial califica como de alto riesgo; lo que comporta el sometimiento a una serie de requisitos obligatorios relativos a la calidad de los datos, la documentación y trazabilidad, la transparencia, la supervisión humana, la exactitud, la ciberseguridad y la solidez. Pues bien, en las páginas siguientes analizaremos, por una parte, qué se entiende jurídicamente por inteligencia artificial conforme al Reglamento europeo; por otra, realizaremos un análisis crítico del sistema *VioGén* y concluiremos con los riesgos que plantea el uso de la IA en la Administración de justicia.

PALABRAS CLAVE: inteligencia artificial, inteligencia artificial judicial, inteligencia artificial policial, sistema *VioGén*

1. Presentación: ¿de qué hablamos jurídicamente cuando hablamos de inteligencia artificial?

Como recuerda Asunción Gómez-Pérez (2023, p. 20), a finales de agosto de 1955, John McCarthy, Marvin Minsky, Nathaniel Rocheste y Claude Shannon propusieron que diez integrantes de un equipo de investigación llevaran a cabo, durante un periodo de dos meses, un estudio sobre inteligencia artificial en el Dartmouth College. Plantearon demostrar la siguiente conjetura: “cualquier aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, estar tan precisamente descrita que se puede construir una máquina para simularla. Se intentará descubrir cómo hacer que las máquinas utilicen el lenguaje, formen abstracciones y conceptos, resuelvan tipos de problemas hasta ahora reservados a los humanos y se mejoren a sí mismas” (McCarthy et al., 1955)².

A partir de este momento fundacional de la IA, se han proporcionado numerosas definiciones –se han contabilizado hasta 55– desde las ópticas política, investigadora e industrial (Samoili et al., 2020, p. 11), que, en general, aluden al desarrollo de sistemas que imitan o reproducen el pensamiento y obrar humanos,

² Pueden verse al respecto el libro de Mitchell (2024, p.27 y ss.) y el discurso de ingreso de la profesora Asunción Gómez-Pérez en la Real Academia Española con el título *Inteligencia artificial y lengua española*, <https://www.rae.es/sites/default/files/2023-05/Discurso%20de%20ingreso%20de%20Asuncion%20Gomez-Perez.pdf> (fecha de consulta: 2 de junio de 2025).

actuando racionalmente –en el sentido de hacer lo correcto en función de su conocimiento– e interactuando con el medio. La IA pretende sintetizar o reproducir los procesos cognitivos humanos, tales como la percepción, la creatividad, la comprensión, el lenguaje o el aprendizaje (Russell & Norvig, 2008, p. 1 y ss.).

La dificultad de ofrecer una definición acabada de la IA se presenta también en el ámbito jurídico, como se puede comprobar leyendo las diferentes versiones que se han ido ofreciendo durante el proceso de aprobación del Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (en adelante, RIA), en cuyo texto se ha conceptualizado como un sistema basado en máquinas y diseñado para funcionar con distintos niveles de autonomía, que puede mostrar capacidad de adaptación tras su despliegue y que, para objetivos explícitos o implícitos, infiere, a partir de las entradas que recibe, salidas tales como predicciones, contenidos, recomendaciones o decisiones que pueden influir en entornos físicos o virtuales (artículo 3). En esta definición cabe descomponer varios elementos, tal y como hacen las Directrices de la Comisión Europea sobre la definición de sistemas de inteligencia artificial, de 6 de febrero de 2025, donde se habla de (1) un sistema basado en máquinas; (2) que esté diseñado para operar con diferentes niveles de autonomía; (3) que pueden exhibir adaptabilidad después del despliegue; (4) que, para objetivos explícitos o implícitos; (5) infiere, a partir de la entrada que recibe, cómo generar salidas (6) como predicciones, contenido, recomendaciones o decisiones (7) que pueden influir en entornos físicos o virtuales.

No todos estos elementos son igualmente importantes ni tienen que estar de forma necesaria presentes en todas las fases para definir a un sistema como de IA; destacaremos dos que sí lo son: la inferencia y la autonomía, que, aunque están muy relacionadas, pueden diferenciarse.

Respecto de la primera, en el considerando 12 del RIA se puede leer que “[e]sta capacidad de inferir se refiere al proceso de obtención de los resultados, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir

en los entornos físicos y virtuales, y a la capacidad de los sistemas de IA para derivar modelos o algoritmos, o ambos, a partir de entradas o datos”.

En palabras de las Directrices, la primera categoría de técnicas de IA mencionada en ese considerando es la de los “enfoques de aprendizaje automático que aprenden de los datos cómo alcanzar determinados objetivos”. Esta categoría incluye una gran variedad de enfoques que permiten a un sistema “aprender” y que, como se especifica en las Directrices, son el aprendizaje supervisado, el no supervisado, el autosupervisado y el aprendizaje por refuerzo. La segunda categoría de técnicas mencionadas en el considerando 12 del RIA son “enfoques basados en la lógica y el conocimiento que infieren a partir del conocimiento codificado o la representación simbólica de la tarea que se va a resolver”.

Pues bien, es esta capacidad de inferencia de un sistema de IA (es decir, su capacidad para generar resultados como predicciones, contenidos, recomendaciones o decisiones que pueden influir en entornos físicos o virtuales) la clave para lograr su autonomía; lo que, como se explica en las Directrices del 6 de febrero, consiste en que los sistemas de IA están diseñados para funcionar con “cierto grado de independencia de las acciones respecto a la intervención humana y de capacidades para funcionar sin intervención humana”.

La referencia a “cierto grado de independencia de acción” en el considerando 12 del RIA excluye los sistemas diseñados para funcionar únicamente con la plena participación e intervención humana manual, que puede ser directa –por ejemplo, a través de controles manuales– o indirecta –por ejemplo, a través de controles automatizados basados en sistemas que permiten a los humanos delegar o supervisar las operaciones del sistema–. En suma, todos los sistemas que están diseñados para funcionar con cierto grado de independencia de las acciones humanas cumplen la condición de autonomía en la definición de un sistema de IA.

2. Inteligencia artificial policial, con especial atención al sistema VioGén

2.1. Presentación

Es bien conocido, en primer lugar, que los sistemas de IA ya se están aplicando en el ámbito de las investigaciones policiales para tratar de anticiparse a la comisión de posibles delitos y, en su caso, adoptar medidas preventivas limitativas de la libertad personal, bien sea atendiendo a criterios geográficos, como *PredPol*, *CompStat*³..., sistemas muy frecuentes en Estados Unidos o en ciertas circunstancias personales, familiares, como el español *VioGén*⁴.

El problema surge cuando estos sistemas se apoyan en datos que pueden reflejar, de manera intencionada o no, sesgos en función de cómo se registran los delitos, qué delitos se seleccionan para ser incluidos en el análisis o qué herramientas analíticas se utilizan, pudiendo generar una retroalimentación en la que, al menos en no pocas ciudades de Estados Unidos, la geografía –las zonas donde se concentra la vigilancia policial para prevenir delitos o reaccionar rápidamente ante ellos– puede operar, en palabras de O’Neil (2018. p. 110), como “un valor sustitutivo altamente eficaz para la raza”⁵.

Estas situaciones se pueden presentar también en ámbitos donde opera el derecho administrativo sancionador (Díaz González, 2020, p. 181 y ss.), como el tributario o el laboral. En relación con este último,

si en el pasado [la inspección] se ha focalizado en un determinado tipo de empresas –donde lógicamente se habrán encontrado incumplimientos– y esos son los datos que se suministran a la herramienta, basándose en esos datos se considerará que estas empresas son las que deben ser inspeccionadas en el futuro. Esto, a su vez, implicará mayor número de fraudes encontrados en esas empresas, confirmando el sesgo y cerrando el círculo. Por esta razón, es tan importante que

³ <https://www.predpol.com/> y <https://www1.nyc.gov/site/nypd/stats/crime-statistics/compstat.page> (a 2 de junio de 2025).

⁴ <https://www.interior.gob.es/opencms/es/servicios-al-ciudadano/violencia-contra-la-mujer/> (a 2 de junio de 2025).

⁵ Civio Transparencia ofrece una exposición de los sistemas usados por la policía local en diferentes ciudades españolas en <https://civio.es/transparencia/2025/06/03/la-policia-local-de-valencia-comienza-a-usar-ia-predictiva/> (a 2 de junio de 2025).

los datos suministrados sean completos, de calidad y sin sesgos manifiestos (Todolí Signes, 2020, p. 333).

2.2. El sistema *VioGén*.

En España, y en el marco de la Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género, el Gobierno aprobó una serie de actuaciones urgentes para luchar contra este tipo de violencia, entre las que destaca la elaboración de un protocolo de valoración de riesgo de la mujer víctima para su uso por parte de las Fuerzas y Cuerpos de Seguridad. Como resultado, el Ministerio del Interior creó y puso en marcha en julio de 2007 el Sistema de Seguimiento Integral de los casos de Violencia de Género (Sistema *VioGén*), dotándolo de formularios informatizados para practicar y administrar las evaluaciones de riesgo de la mujer víctima, así como de las funcionalidades precisas para llevar a cabo el seguimiento de dichos casos y la implementación de las medidas de seguridad y protección policial acordes con los niveles de riesgo resultantes. La última actualización se ha llevado a cabo a través de la Instrucción número 4/2019, de la Secretaría de Estado de Seguridad, por la que se establece un nuevo protocolo para la valoración policial del nivel de riesgo de violencia de género, la gestión de la seguridad de las víctimas y el sistema de seguimiento integral de los casos de violencia de género.

Desde la entrada del sistema en funcionamiento y hasta el 31 de mayo de 2025 se han evaluado 867.472 casos de violencia de género: del total de los casos registrados había, en esa fecha, 102.575 activos, es decir, con seguimiento policial. De ellos, 44.613 eran sin riesgo apreciado, 43.235 con riesgo bajo, 13.734 con riesgo medio, 978 con riesgo alto y 15 con riesgo extremo (Estadísticas Ministerio del Interior, 2025)⁶. En conjunto estamos hablando del mayor sistema del mundo en ese ámbito (Fundación Éticas, 2022, p. 12)⁷.

Esta herramienta predictiva no es, si asumimos la definición ya vista del RIA, “inteligencia artificial”, pues no tiene autonomía y capacidad de

⁶ <https://www.interior.gob.es/opencms/pdf/servicios-al-ciudadano/violencia-contra-la-mujer/estadisticas/2025/Datos-a-31052025-1.pdf> (a 2 de junio de 2025). Al respecto, pueden verse los estudios de Presno Linera (2023), San Martín Segura (2023) y Martínez Garay et al. (2024).

⁷ <https://dev.eticasfoundation.org/es/la-fundacion-eticas-realiza-una-auditoria-externa-e-independiente-del-sistema-viogen/> (a 2 de junio de 2025).

autoaprendizaje, sino que es “un sistema actuarial que utiliza modelos estadísticos para inferir el riesgo que puede correr una víctima (tanto de agresión como de homicidio), así como su evolución en base a un conjunto de indicadores que han sido determinados y posteriormente evaluados por un grupo de expertos”. No obstante, podría considerarse un sistema de IA en un sentido “impropio” y no parece estar descartada la incorporación de un algoritmo de autoaprendizaje.

1.-HISTORIA DE VIOLENCIA EN LA RELACIÓN DE PAREJA	Respuestas		
Indicador 1: Violencia psicológica (vejaciones, insultos y humillaciones)	SI	NO	N/S
1.1 Intensidad de la violencia psicológica	Leve	Grave	Muy grave
Indicador 2: Violencia física	SI	NO	N/S
2.1 Intensidad de la violencia física	Leve	Grave	Muy grave
Indicador 3: Sexo forzado	SI	NO	N/S
3.1 Intensidad de la violencia sexual	Leve	Grave	Muy grave
Indicador 4: Empleo de armas u objetos contra la víctima	SI	NO	N/S
4.1 Arma blanca 4.2. Arma de fuego 4.3. Otros objetos	SI	NO	N/S
Indicador 5: Existencia de amenazas o planes dirigidos a causar daño a la víctima	SI	NO	N/S
5.1 Intensidad de las amenazas	Leve	Grave	Muy grave
5.2 Amenazas de suicidio del agresor	SI	NO	
5.3 Amenazas de muerte del agresor dirigidas a la víctima	SI	NO	
Indicador 6: En los últimos seis meses se registra un aumento de la escalada de agresiones o amenazas	SI	NO	N/S
2.-CARACTERÍSTICAS DEL AGRESOR			
Indicador 7: En los últimos seis meses, el agresor muestra celos exagerados o sospechas de infidelidad	SI	NO	N/S
Indicador 8: En los últimos seis meses, el agresor muestra conductas de control	SI	NO	N/S
Indicador 9: En los últimos seis meses, el agresor muestra conductas de acoso	SI	NO	N/S
Indicador 10: Existencia problemas en la vida del agresor en los últimos seis meses	SI	NO	N/S
10.1 Problemas laborales o económicos	SI	NO	
10.2 Problemas con el sistema de justicia	SI	NO	
Indicador 11: En el último año el agresor produce daños materiales	SI	NO	N/S
Indicador 12: En el último año se registran faltas de respeto a la autoridad o a sus agentes	SI	NO	N/S
Indicador 13: En el último año agrede físicamente a terceras personas y/o animales	SI	NO	N/S
Indicador 14: En el último año existen amenazas o desprecios a terceras personas	SI	NO	N/S
Indicador 15: Existen antecedentes penales y/o policiales del agresor	SI	NO	N/S
Indicador 16: Existen quebrantamientos previos o actuales (cautelares o penales)	SI	NO	N/S
Indicador 17: Existen antecedentes de agresiones físicas y/o sexuales	SI	NO	N/S
Indicador 18: Existen antecedentes de violencia de género sobre otra/s pareja/s	SI	NO	N/S
Indicador 19: Presenta problemas un trastorno mental y/o psiquiátrico	SI	NO	N/S
Indicador 20: Presenta ideas o intentos de suicidio	SI	NO	N/S
Indicador 21: Presenta algún tipo de adicción o conductas de abuso de tóxicos (alcohol, drogas y fármacos)	SI	NO	N/S
Indicador 22: Presenta antecedentes familiares de violencia de género o doméstica	SI	NO	N/S
Indicador 23: El agresor tiene menos de 24 años	SI	NO	N/S
3.-FACTORES DE RIESGO / VULNERABILIDAD DE LA VÍCTIMA			
Indicador 24: Existencia de algún tipo de discapacidad, enfermedad física o psíquica grave	SI	NO	N/S
Indicador 25: Víctima con ideas o intentos de suicidio	SI	NO	N/S
Indicador 26: Presenta algún tipo de adicción o conductas de abuso de tóxicos (alcohol, drogas y fármacos)	SI	NO	N/S
Indicador 27: Carece de apoyo familiar o social favorable	SI	NO	N/S
Indicador 28: Víctima extranjera	SI	NO	
4.-CIRCUNSTANCIAS RELACIONADAS CON LOS MENORES			
Indicador 29: La víctima tiene a su cargo menores de edad	SI	NO	N/S
Indicador 30: Existencia de amenazas a la integridad física de los menores	SI	NO	N/S
Indicador 31: La víctima teme por la integridad de los menores	SI	NO	N/S
5.-CIRCUNSTANCIAS AGRAVANTES			
Indicador 32: La víctima ha denunciado a otros agresores en el pasado	SI	NO	N/S
Indicador 33: Se han registrado episodios de violencia lateral recíproca	SI	NO	N/S
Indicador 34: La víctima ha expresado al agresor su intención de romper la relación hace menos de seis meses	SI	NO	N/S
Indicador 35: La víctima piensa que el agresor es capaz de agredirla con mucha violencia o incluso matarla	SI	NO	N/S

Tabla 1: Modelo de impresión del Formulario VFR5.0-H para la Valoración Forense del Riesgo

Según el protocolo, en ningún caso una mujer denunciante abandonará las dependencias policiales sin haber sido valorada ni sin que, en su caso, se le hayan

asignado las medidas policiales de protección que correspondan al nivel de riesgo resultante.

El formulario incluye 5 dominios con 35 indicadores de riesgo. Cada ítem se valora como "presente" y "no presente". De este modo, la recogida de información está estandarizada en todo el país.

El formulario incluye 5 dominios con 35 indicadores de riesgo. Cada ítem se valora como "presente" y "no presente". De este modo, la recogida de información está estandarizada en todo el país.

Quien recibe la denuncia cumplimenta el formulario y, sin mostrar todavía el resultado, el Sistema *VioGén* aplica el primer algoritmo y calcula el riesgo de reincidencia que presenta el caso en ese momento; inmediatamente después, calcula el riesgo de feminicidio con el segundo algoritmo. Si considera que el riesgo real es más alto que el calculado por el sistema, el agente actuante puede elevarlo, pero si lo considera más bajo no puede reducirlo. En caso de que aparezca riesgo mortal, se ha dispuesto que se incremente en un nivel el riesgo de reincidencia, que es el que se muestra finalmente al personal policial, junto con una alerta de que el caso es de especial interés, para que se pueda adecuar la protección policial a las características del caso concreto. Además, esta alerta se refleja en una diligencia en el atestado policial, que se envía al Juzgado y a la Fiscalía competentes, para conocimiento de la singularidad del caso y por si estimaran pertinente que las personas implicadas sean evaluadas cuanto antes por profesionales de la psicología o de la medicina forense, quienes podrían profundizar en las circunstancias del caso y proponer nuevas medidas protectoras. El resultado se comunicará a la Autoridad Judicial y Fiscal en forma de Informe automatizado que genera el propio Sistema.

Cada uno de los niveles de riesgo llevará aparejadas medidas policiales para la protección y seguridad de las víctimas, que serán de aplicación obligatoria e inmediata. Así, por ejemplo, si el riesgo es "alto" y en caso de no haberse podido localizar todavía al agresor, se insistirá a la víctima, para su más efectiva protección, en la posibilidad de traslado a un centro de acogida, a la casa de un familiar o domicilio distinto, y se llevará a cabo un control frecuente y aleatorio en

su domicilio y lugar de trabajo y, si procede, en los centros escolares de las personas menores a su cargo a la entrada y salida. Respecto del agresor, se hará un control aleatorio de sus movimientos y contactos esporádicos con personas que frecuente o de su entorno. Si el riesgo se califica como “extremo”, se dará protección permanente de la víctima hasta que el mismo agresor o sus circunstancias dejen de ser una amenaza inminente y, si es procedente, se hará vigilancia a la entrada y salida de sus hijos en los centros escolares. Respecto al agresor, se hará un control intensivo de sus movimientos hasta que este deje de ser una amenaza inminente para la seguridad de la víctima. Esas medidas se adaptarán a las circunstancias concretas del caso, de manera que sean de aplicación personalizada e individual y se comunicarán a la víctima. Si, tras la primera actuación judicial, se acordara alguna medida de alejamiento/orden de protección, esta será comunicada expresamente tanto a víctima como a agresor. Dicha comunicación incluirá la explicación concreta de la medida acordada, alcance y consecuencias de su quebrantamiento.

Una vez explicado de manera sucinta el funcionamiento del sistema *VioGén*, es importante mencionar también alguna de las críticas y objeciones que se han formulado al mismo; así, en primer lugar, y tras un análisis de esta herramienta, López-Ossorio, González Álvarez y Andrés Pueyo (2016, p. 6) explicaban en 2016 que “la sensibilidad o identificación correcta del riesgo de violencia cuando existe la reincidencia fue del 85%, y la capacidad del instrumento para descartar el riesgo cuando no se dio reincidencia, o especificidad, fue del 53,7%”. El valor predictivo negativo sería del 98.5% y el positivo del 8,6%. Con los ajustes de 2019 se estima una sensibilidad del 81% y una especificidad del 61%.

Es preciso aclarar, utilizando las palabras de la profesora Martínez Garay (2014, p. 28), que la sensibilidad es la capacidad de un instrumento de predicción para detectar a las personas que sí reincidirán; la especificidad es un valor complementario al anterior: es la capacidad del instrumento para detectar correctamente a los que no reincidirán. Ambas categorías son complementarias en el sentido de que, cuanto mayor es una de ellas, generalmente menor es la otra: cuanto más amplios sean los criterios para clasificar a una persona como peligrosa, mayor será la sensibilidad (menos peligrosos se “escaparán” del diagnóstico),

pero menor será la especificidad, porque aumenta la probabilidad de incluir como peligrosas a personas que en realidad no lo son (“falsos positivos”). Si los criterios para clasificar a alguien como peligroso son muy estrictos, tendremos menos fallos de este segundo tipo (pocos no-peligrosos serán erróneamente considerados peligrosos), pero habrá personas que sí iban a delinquir en el futuro que se nos habrán quedado fuera del diagnóstico (“falsos negativos”).

Pues bien, si las cosas son así resulta que el sistema *VioGén* fallaría bastante en especificidad, pues, con arreglo al estudio de 2015, casi la mitad de las personas (46,3%) habrían sido diagnosticadas incorrectamente como peligrosas (“falsos positivos”), cantidad que bajaría en 2019 al 39%.

El valor predictivo negativo de *VioGén* –los casos en los que no se advirtió riesgo y, efectivamente, no hubo agresiones– sería muy alto (más del 98%) y muy bajo el positivo –supuestos en los que se pronosticó agresiones y las hubo– (solo el 8,7%); lo que parece lógico, pues, como recuerdan Martínez Garay y García Ortiz (2022, p. 168),

las estimaciones de riesgo de reincidencia se diferencian de las realizadas en otros contextos en que la ocurrencia del evento estimado no es independiente del resultado de la valoración. Si predecimos que hará buen tiempo el fin de semana, puede que ocurra o no, pero nuestra predicción no habrá influido en ello. Sin embargo, cuando se estima el riesgo de violencia, se toman medidas como consecuencia de esas valoraciones (imponer o no una medida cautelar, etc.), medidas que influyen sobre la propia situación valorada. Así, si como consecuencia de una valoración de riesgo alto se adoptan medidas para minimizarlo y estas son eficaces, el evento, contrariamente a lo esperado, no se producirá. Podrá parecer que la estimación fue “equivocada”, y no necesariamente es así (Martínez Garay y García Ortiz, 2022, p. 168).

A este respecto, Martínez Garay y García Ortiz señalan que

este no es un problema que resuelvan la estadística ni los algoritmos, porque es una cuestión político criminal, que presupone una decisión sobre qué es preferible: ¿restringir la libertad de muchas personas que en realidad no hubieran delinquido después o renunciar al control penal sobre personas que van a seguir

cometiendo delitos? Cuando se programa un algoritmo para ayudar a hacer predicciones, alguien ha tomado esta decisión y ha decidido situar los umbrales de discriminación en unos puntos concretos. Y debería estar en condiciones de defender esa decisión ante los afectados por ese algoritmo y ante la opinión pública. En este punto la transparencia aparece como una cuestión fundamental (Martínez Garay y García Ortiz, 2022, p. 165).

En segundo lugar, y como explican González Álvarez, Santos Hermoso y Camacho Collados,

las técnicas de policía predictiva se basan en el análisis de datos históricos, es decir, casos que llegan a conocimiento de los cuerpos policiales. Es por esto que los algoritmos que se generen serán específicos para esos casos, y permitirán predecir casos que muestren características similares o sigan un mismo patrón. El problema... es que tanto en la violencia de género, como en la violencia doméstica en general, muchos casos no llegan a denunciarse, y en consecuencia no forman parte de los registros policiales históricos. Esto plantea otra reflexión importante, y es que, con estos casos no denunciados pueden suceder dos cosas: 1) que sean similares a los casos que sí denuncian, por lo que las herramientas podrían ser aplicables; o 2) puede que tengan una serie de características distintivas que, en parte, expliquen el por qué no se denuncia, y las herramientas de predicción no sirvan (González Álvarez et al., 2020, p. 34).

Y esa ausencia de denuncia previa puede atribuirse, en buena medida, a la violencia que el agresor ejerce sobre la mujer y al miedo que tal situación genera (véase la sentencia del Tribunal Supremo nº 247, de 24 de mayo de 2018, FJ 2).

En tercer lugar, este sistema parece generar un sesgo de “autoridad tecnológica” o de “automatización” si, como se ha venido diciendo (González Álvarez et al., 2018, p. 37), hasta en el 95% de los casos los agentes mantienen la puntuación de riesgo asignada automáticamente por el algoritmo, a pesar de que pueden estimar que existe en riesgo superior al que predice *VioGén* y si tal cosa no se hace por una confianza casi automática en las predicciones del sistema, se estará incumpliendo el propio protocolo; lo que, en su caso, podría dar lugar a la atribución de diferentes tipos de responsabilidad. Por otra parte, puede existir el riesgo de que, precisamente, ante el temor a incurrir en alguna responsabilidad, se

tienda a elevar de forma casi automática el nivel de riesgo que ha pronosticado el sistema *VioGén*.

Para evitar el peligro de un sesgo de automatización, *VioGén* podría articularse como una herramienta predictiva que ofreciera una evaluación en paralelo a la que tendría que efectuar quien tramite la denuncia, debiendo tenerse en cuenta otros posibles factores y sin conocer el pronóstico del sistema antes de haber hecho el suyo propio.

Finalmente, hay que mencionar la falta de transparencia del sistema: aunque, como hemos visto, los indicadores de riesgo son públicos, no lo son los datos sobre cómo se combinan, qué relevancia tiene cada uno en el resultado final e, incluso, y por citar un ejemplo, por qué la condición de extranjera de la víctima, sin otro tipo de consideraciones, es un factor a tener en cuenta. En suma

No se puede acceder a ningún dato o información más allá de lo producido por los expertos que participaron en la definición del sistema... El sistema no ha sido evaluado ni auditado de forma independiente y tampoco involucra a las destinatarias del mismo, que nunca han sido consultadas sobre el sistema, ni en su fase de diseño ni posteriormente durante las diferentes decisiones sobre cómo modificarlo (Fundación Éticas, 2022, p. 34).

3. Inteligencia artificial judicial

Muy relacionado con lo dicho está el recurso a la IA en el ámbito de justicia –IA judicial– para, por ejemplo, apoyar la toma de decisiones sobre prisión provisional o libertad condicional. Así, se está utilizando desde 2009 en Cataluña, como parte del proceso de evaluación previo a la concesión de permisos carcelarios, el sistema *RisCanvi*, que calcula el riesgo de reincidencia de los presos. En primer lugar, se usa una versión corta (*RisCanvi Screening*), que opera a través de 10 apartados (si existe un historial violento, edad a la que delinquiró por primera vez, drogadicciones, alcoholismo, recursos económicos, apoyo familiar) que ofrece dos resultados: riesgo bajo y riesgo alto; en el segundo caso se emplea una versión ampliada formada por 43 factores de riesgo agrupados en las áreas criminológica, personal y biográfica, social y familiar, clínica y de personalidad (nivel educativo, distancia del centro penitenciario a la residencia habitual, cargas familiares ,etc.). A

estos 43 factores se les suma la valoración del riesgo futuro de aparición de cuatro conductas: violencia autodirigida, violencia intrainstitucional, reincidencia violenta y quebrantamiento de condena (Karimi-Haghighi y Castillo, 2021).

De acuerdo con el RIA, deben clasificarse como de alto riesgo determinados sistemas de IA destinados a la administración de justicia y los procesos democráticos, dado que pueden tener efectos potencialmente importantes para la democracia, el Estado de Derecho, las libertades individuales y el derecho a la tutela judicial efectiva y a un juez imparcial. En particular, a fin de hacer frente al riesgo de posibles sesgos, errores y opacidades, procede clasificar como de alto riesgo aquellos sistemas de IA destinados a ser utilizados por una autoridad judicial o en su nombre. Estos sistemas sirven para ayudar a las autoridades judiciales a investigar e interpretar los hechos y el Derecho, así como a aplicar la ley a unos hechos concretos. También deben considerarse de alto riesgo los sistemas de IA destinados a ser utilizados por los organismos de resolución alternativa de litigios con esos fines, cuando los resultados de los procedimientos de resolución alternativa de litigios surtan efectos jurídicos para las partes. La utilización de herramientas de IA puede apoyar el poder de decisión de los jueces o la independencia judicial, pero no debe sustituirlas: la toma de decisiones finales debe seguir siendo una actividad humana.

Parece mantenerse una diferencia sustancial entre la función de apoyo o ayuda y la función decisoria de estos sistemas que, en todo caso, al ser sistemas de alto riesgo, estarán sometidos a supervisión humana por mandato del artículo 14 RIA para “prevenir o reducir al mínimo los riesgos para la salud, la seguridad o los derechos fundamentales que pueden surgir cuando se utiliza un sistema de IA de alto riesgo conforme a su finalidad prevista o cuando se le da un uso indebido razonablemente previsible”.

No obstante, la diferencia entre ayuda, decisión y la propia eficacia de la supervisión humana resultan, cuando menos, discutibles, si tenemos en cuenta la constatada existencia de un sesgo de automatización, que admite el propio RIA cuando requiere que las personas físicas a quienes se encomiende la supervisión humana deban “ser conscientes de la posible tendencia a confiar automáticamente

o en exceso en los resultados de salida generados por un sistema de IA de alto riesgo (“sesgo de automatización”), en particular con aquellos sistemas que se utilizan para aportar información o recomendaciones con el fin de que personas físicas adopten una decisión (artículo 14.3.b).

En este sentido, Ben Green (2022), al estudiar las políticas de supervisión humana en sus interacciones con los algoritmos, encontró que estas políticas adolecen de dos fallas significativas: primera, las evidencias sugieren que las personas no pueden proporcionar una adecuada supervisión; segunda, la suposición incorrecta de que una supervisión humana efectiva legitima el uso de algoritmos defectuosos e irresponsables.

Lo decidido por el Tribunal Supremo de Wisconsin en el famoso caso *State v. Loomis*, de 10 de marzo de 2017, supone un buen motivo para la reflexión: Eric Loomis fue acusado de cinco delitos por su presunta intervención en un tiroteo efectuado desde un vehículo, y llegó a un acuerdo con la Fiscalía para aceptar los dos últimos cargos y rechazar el resto de acusaciones. El juez de instancia aceptó la conformidad y, al concretar la pena, tuvo en cuenta, entre otras cosas, el resultado de la evaluación de Loomis mediante COMPAS, una herramienta estructurada que valora el riesgo de reincidencia y las necesidades criminógenas del sujeto.

Según el sistema COMPAS, Loomis presentaba un riesgo alto tanto de reincidencia en general como de reincidencia violenta. El juez lo condenó a 6 años de prisión y otros 5 de supervisión postpenitenciaria, y en la sentencia explicó que “a la hora de valorar los distintos factores, excluyo la suspensión de la condena [probation] debido a la gravedad del crimen y porque su historial, su historial en materia de supervisiones judiciales, y las herramientas de valoración del riesgo que se han aplicado, indican que usted presenta un riesgo extremadamente alto de reincidir”⁸.

Loomis recurrió ante el Tribunal Supremo de Wisconsin alegando que el uso de COMPAS para determinar la pena había vulnerado su derecho a un proceso con todas las garantías por tres razones: en primer lugar, se había violado su derecho a

⁸ <https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690> (a 2 de junio de 2025).

ser condenado sobre la base de información fiable y precisa, ya que la naturaleza de secreto comercial del algoritmo con el que funciona COMPAS le había impedido conocer cómo se habían calculado las estimaciones de riesgo y, por tanto, refutar la validez científica del nivel de riesgo que dicha herramienta le había asignado; en segundo lugar, se había vulnerado su derecho a obtener una sentencia individualizada; y en tercer lugar, se había tenido en cuenta indebidamente su género para determinar la pena (Martínez Garay, 2018, pp. 485-502).

Por lo que aquí importa, el Tribunal Supremo estatal concluyó que el hecho de no haber podido saber cómo funciona el algoritmo que calcula los niveles de riesgo, considerado por la empresa creadora como un secreto esencial para su negocio, no vulneró el derecho que tiene todo acusado a ser condenado sobre la base de información fiable y exacta, porque Loomis sí tuvo acceso tanto a los resultados de riesgo que COMPAS le asignó como a la información con la que estos se calcularon, que consiste, básicamente, en datos objetivos sobre el historial criminal del acusado. En segundo lugar, para el Tribunal Supremo, las valoraciones de riesgo aportan al juez información que, unida al resto de circunstancias concurrentes, contribuye a que la condena se ajuste lo más posible al caso concreto, lo que los jueces deben hacer cuando manejen estas herramientas es aplicar a sus resultados un margen de discrecionalidad, poniéndolos en el contexto del resto de circunstancias particulares propias del caso y del individuo concretos.

Pues bien, como señala Martínez Garay (2018, p. 497), aunque de la sentencia deriva que el resultado de la valoración de riesgo sólo puede servir como corroboración de otros indicios, eso no necesariamente impedirá que en la práctica pueda funcionar justo al contrario: que una vez conocido el nivel de riesgo estimado por la herramienta, el juez elabore su razonamiento condicionado –inconscientemente– por dicha cifra, a la que amoldará su valoración del resto de circunstancias del caso concreto para que encajen con ella. En segundo lugar, el aspecto más sorprendente del caso es que el Tribunal Supremo estatal aceptase que las sentencias puedan basarse, en parte, en valoraciones de riesgo realizadas mediante herramientas comercializadas por compañías privadas, cuyo funcionamiento no se revela al público ni al acusado porque se considera un secreto de empresa. Si el acusado tiene derecho a ser condenado sobre la base de

información fiable y exacta, no puede ser suficiente con facilitarle los resultados de su valoración de riesgo y a los factores con los que se calcularon esas puntuaciones, pero sin permitirle saber cuánto peso se ha dado a cada factor, si todos ellos o sólo algunos han influido al final en su puntuación, ni cómo se han combinado.

BIBLIOGRAFÍA

Díaz González, G. M. (2020). “Algoritmos y actuación policial: la policía predictiva”, en Huergo Lora, A. (dir.) y Díaz González, G. M. (coord.). *La regulación de los algoritmos*, Cizur Menor: Aranzadi.

Fundación Éticas (2022). *Auditoría externa del sistema VioGén*. Barcelona.

Gómez-Pérez, A. (2023). *Inteligencia artificial y lengua española* [Discurso de ingreso en la RAE]. <https://www.rae.es/sites/default/files/2023-05/Discurso%20de%20ingreso%20de%20Asuncion%20Gomez-Perez.pdf>

González Álvarez, J. L.; López Ossorio, J. J.; Urruela, C.; Rodríguez Díaz, M. (2018). “Integral Monitoring System in Cases of Gender Violence. VioGén System”. *Behavior & Law Journal*, 4(1), 29-40. <https://doi.org/10.47442/blj.v4.i1.56>

González Álvarez, J. L.; Santos Hermoso, J.; Camacho Collados, M. (2020). “Policía predictiva en España aplicación y retos futuros” en *Behavior & Law Journal*, vol. 6 (1), 26-41. <https://doi.org/10.47442/blj.v6.i1.75>

Green, B. (2022). “The Flaws of Policies Requiring Human Oversight of Government Algorithms”. *Computer Law & Security Review*, 45. <https://doi.org/10.1016/j.clsr.2022.105681>

Karimi-Haghighi, M.; Castillo, C. (2021). “Efficiency and Fairness in Recurring Data-Driven Risk Assessments of Violent Recidivism”. Proceedings of the 36th Annual ACM Symposium on Applied Computing (SAC '21). Association for Computing Machinery, New York, NY, USA, 994–1002. <https://doi.org/10.1145/3412841.3441975>

- López Ossorio, J. J.; González Álvarez, J. L.; Andrés Pueyo, A. (2016). "Eficacia predictiva de la valoración policial del riesgo de la violencia de género". *Psychosocial Intervention*, 25(1), 1-7.
- McCarthy, J; Minsky, M. L.; Rochester, N.; Shannon, C. E. (1955): *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>
- Martínez Garay, L. (2014). "La incertidumbre de los pronósticos de peligrosidad: consecuencias para la dogmática de las medidas de seguridad". *Indret: Revista para el Análisis del Derecho*, 3, 2-77.
- Martínez Garay, L.; García Ortiz, A. (2022). "Paradojas de los algoritmos predictivos utilizados en el sistema de justicia penal". *El Cronista del Estado Social y Democrático de Derecho*, (100), 160-173.
- Martínez Garay, L. (2024). Three predictive policing approaches in Spain: VIOGÉN, RISCANVI and VERIPOL: Assessment from a human rights perspective. *Publicacions De La Universitat De València*.
- Mitchell, M. (2024). *Inteligencia artificial: Guía para seres pensantes*. Madrid: Capitán Swing.
- O'Neil, C. (2018). *Armas de destrucción matemática. Cómo el big data aumenta la desigualdad y amenaza la democracia*, Madrid: Capitán Swing.
- Presno Linera, M. Á. (2023). "Policía predictiva y prevención de la violencia de género: el sistema VioGén". *IDP. Revista de Internet, Derecho y Política*, (39), 1-13. 10.7238/idp.v0i39.416473
- Russell, S., Norvig, P. (2004). *Inteligencia Artificial: un enfoque moderno*. Madrid: Pearson Education.
- Samoili, S., Lopez Cobo, M., Gomez Gutierrez, E., De Prato, G., Martinez-Plumed, F. and Delipetrev, B. (2020). *AI WATCH. Defining Artificial Intelligence*. Luxemburgo: Publications Office of the European Union. <https://doi:10.2760/382730>

San Martín Segura, D. (2023). "Prevención algorítmica de la violencia de género: la discrecionalidad policial como decisión tecnológica en el contexto de VIOGÉN". *Estudios Penales Y Criminológicos*, 44(Ext.), 1-35.

Todolí Signes, A. (2020). "Retos legales del uso del big data en la selección de sujetos a investigar por la Inspección de Trabajo y de la Seguridad Social". *Revista Galega de Administración Pública*, (59), 313-337.

IA para simular las prestaciones de los y las jueces en la mejora del rendimiento de las oficinas judiciales italianas

Paolo Inturri
Universidad de Catania

RESUMEN

El artículo ilustra las potencialidades del uso de modelos de simulación en los procesos de toma de decisiones relacionados con la organización de las oficinas judiciales, en la perspectiva de implementar el derecho a un proceso sin dilaciones indebidas.

Con este fin, la discusión se centra en la descripción de los resultados del proyecto *Just Smart* de la Universidad de Catania concerniente al desarrollo de un modelo de simulación del rendimiento de una sección civil de tribunal, como presupuesto para plantear las potencialidades y los puntos críticos de un posible uso de la inteligencia artificial en el sector a la luz de la normativa del Reglamento de inteligencia artificial.

PALABRAS CLAVE: Derecho a un a un proceso sin dilaciones indebidas; modelos de simulación; inteligencia artificial; Reglamento de inteligencia artificial

1. Introducción

Entre los principios que impregnan la estructura de la administración de justicia europea, el derecho a un proceso sin dilaciones indebidas es de especial importancia. Su reconocimiento formal debe atribuirse al art. 6 del Convenio Europeo de Derechos Humanos (CEDH), que establece el derecho de toda persona «a que su causa sea oída [...] dentro de un plazo razonable [...]».

En el ordenamiento jurídico italiano, del cual parte el presente análisis, el derecho se introdujo mediante la Ley n.º 848/1995, de ratificación y ejecución del CEDH, y con la Ley constitucional n.º 2/1999 que lo incorporó a la Constitución (art. 111, ap. 2), y se implementó principalmente mediante la Ley n.º 89/2001 (“Ley Pinto”), que previó el derecho a una indemnización equitativa por un proceso con dilaciones indebidas (art. 2, ap. 2 *bis*).

Más generalmente, la actualización del derecho ha sido abordada de diversas maneras en el ordenamiento italiano, con intervenciones que afectan tanto al procedimiento como a la organización de las oficinas judiciales. La doctrina jurídica, naturalmente centrada en el análisis de las reformas procesales, ha dedicado poca atención a las organizativas.

Entre estas últimas, se debe prestar especial atención a la introducción del “*Ufficio per il processo*” (U.P.P.) (art. 50, Decreto-Ley n.º 90/2014, convertido, con modificaciones, por la Ley n.º 114/2014), es decir, un personal auxiliar de los magistrados, con funciones accesorias respecto a las puramente jurisdiccionales (Servizio Studi della Camera dei deputati, 2014, p. 196).

La inclusión de la duración razonable del proceso dentro de las reformas horizontales del PNRR ha supuesto una oportunidad para potenciar el mencionado U.P.P. (Servizio Studi della Camera dei deputati, 2022), por una parte, enriqueciendo su personal mediante la inclusión de la figura profesional de los “*addetti*” (Decreto-Ley n.º 80/2021, convertido con modificaciones por la Ley n.º 113/2021); por otra, regulando su marco normativo conforme a una disciplina unitaria (Decreto Legislativo n.º 151/2022) (Silvano, 2024).

Ante la discrecionalidad organizativa dejada a las distintas oficinas judiciales por la normativa, el “*Consiglio Superiore della Magistratura*” ha emitido unas directrices para el reconocimiento de buenas prácticas adoptadas, si bien sin contar con análisis previos sobre el impacto de las diversas posibles políticas organizativas (Consiglio Superiore della Magistratura, 2024).

En la literatura no faltan ejemplos de modelos de simulación utilizados precisamente para predecir los efectos de las decisiones organizativas en la duración de los procesos, cuya utilización parece cada vez más factible debido a la creciente digitalización y datificación de los procedimientos judiciales.

En este contexto se sitúa el presente trabajo, que ofrece una panorámica de la literatura sobre el uso de modelos de simulación en la administración de justicia (apartado 2) y un análisis detallado del desarrollo de uno de estos en el marco del proyecto “*Giustizia Smart: Strumenti e Modelli per ottimizzare il lavoro dei giudici*” (“*Just Smart*”) de la Universidad de Catania, como premisa para hipotetizar el uso de la inteligencia artificial (IA) en el sector y abordar los problemas jurídicos relacionados con el Reglamento (UE) 2024/1689, de 13 de junio de 2024, Reglamento de inteligencia artificial (RIA) (apartado 4), a lo que seguirán las conclusiones del autor (apartado 5).

2. Modelos de simulación y administración de justicia

La justicia se encuentra en un proceso de digitalización y datificación que, abstractamente, permitiría innovar sus procesos de toma de decisiones mediante el procesamiento de datos. Por ello, no es casual que diversos estudios se centren en los posibles usos de una técnica clave en este ámbito: la simulación.

Así, simulaciones basadas en modelos de agentes se han empleado para evaluar el impacto de las políticas en el sistema judicial inglés y galés, con el objetivo de racionalizar la revisión del gasto (Boyle, Guerin, Pratt, & Kunkle, 2003). Asimismo, mediante simulaciones basadas en modelos de eventos discretos, se ha predicho el impacto de un determinado programa del Ministerio de Justicia neerlandés en la duración de los procesos civiles en materia de derecho mercantil (Van Donge, 2017). De modo análogo, simulaciones basadas en el empleo de

modelos multi-agente han sido utilizadas para prever los efectos de las decisiones del presidente del Tribunal de Livorno en la duración de los procesos civiles (Di Martino, Esposito, & Colucci Cante, 2021). Finalmente, con simulaciones basadas en modelos bayesianos, se han estimado los efectos de posibles reformas institucionales dirigidas a agilizar el trabajo del Tribunal de Justicia de la Unión Europea (Fjelstul, Gabel, & Carrubba, 2022).

3. El proyecto Just Smart

Tomando como punto de partida la literatura analizada, el proyecto Just Smart se ha centrado en el desarrollo de un modelo de eventos discretos para simular, en particular, el impacto en la duración de los procesos civiles de las decisiones de los presidentes de las secciones judiciales relativas a la organización del U.P.P.

Los siguientes subapartados están dedicados a la exposición de las fases de desarrollo y utilización del modelo: limpieza y preparación de datos (subapartado 3.1.); programación, verificación y validación (subapartado 3.2.); análisis de escenarios (subapartado 3.3.).

3.1. Limpieza y preparación de datos

El modelo se ha desarrollado a partir de los datos relativos a los procesos definidos de la cuarta sección civil del Tribunal de Catania en el período comprendido entre el 1 de enero de 2017 y el 31 de diciembre de 2022.

En la fase de limpieza de datos se procedió a la eliminación de los datos superfluos respecto a la finalidad del análisis y de los “outliers” (procesos cautelares ante causam, procedimientos de corrección de error material, procesos monitorios).

Posteriormente, se llevó a cabo la preparación de datos a través del software Minitab®. En primer lugar, se creó un vector de probabilidad de ocurrencia de las materias de los procesos, cuya función es generar de forma aleatoria los expedientes a asignar basándose en el histórico de la sección. Luego, dado que no todos los y las jueces presentes en la sección tramitan todas las materias de su

competencia, sino que se remite al presidente la elección respecto a la política de asignación de cada controversia (art. 47 quáter, Real Decreto n.º 12/1941), se ideó una matriz probabilística de elegibilidad, en la que las filas identifican a los y las jueces y las columnas las materias tratadas por la sección: los valores de la matriz serán 1 si el o la juez trata una determinada materia, de lo contrario serán 0. Además, se realizó un análisis estadístico (test de "Anderson Darling") para asignar la distribución de probabilidad estocástica para la generación aleatoria de procesos por materia. Finalmente, se definió el parámetro de la capacidad de los y las jueces: la cantidad máxima de procesos que cada integrante de la sección puede tramitar.

3.2. Planificación, verificación y validación

La programación del modelo de simulación se llevó a cabo en Matlab®. El modelo replica el diagrama de flujo de trabajo de una sección judicial italiana: (1) registro de un proceso en la lista; (2) asignación del proceso a los y las jueces; (3) tramitación de los procesos en paralelo hasta su conclusión (Tabla 1).

Step 1:	Leggo i dati di input (es: N = numero giudici, M = numero materie, $Elig$ = eligibility materie)
Step 2:	Inizializzo le variabili (es: $Disp_g$ = disponibilità giudice, M_g = statistiche giudice)
Step 3:	Seleziono numero di repliche, R_{max}
Step 4:	For $i = 1 : R_{max}$
Step 5:	Genero una coda iniziale di fascicoli, WIP , per ogni giudice
Step 6:	For $j = 1 : M$
Step 7:	Genero un fascicolo e assegno l'id della materia
Step 8:	Assegno il fascicolo ad un giudice che può processarlo secondo la matrice di eligibility, $Elig$
Step 9:	Assegno un tempo di processamento
Step 10:	Calcolo tempo di inizio di processamento e tempo di definizione del fascicolo
Step 11:	Calcolo numero di fascicoli processati dal giudice
Step 12:	End
Step 13:	Calcolo le performance del processo
Step 14:	End
Step 15:	Calcolo le medie delle performance sul numero totale di repliche

Tabla 1: Pseudocódigo del modelo de simulación

Una vez finalizada la programación del modelo, se procedió a verificar su exactitud y a la depuración de errores (*de-bugging*), teniendo en cuenta que, debido a la naturaleza estocástica de las simulaciones, es necesario repetirlas un

elevado número de veces para evitar que el “*ruido de fondo*” influya en sus resultados.

Al término de la fase de verificación, se procedió a la validación del modelo comparando el rendimiento real de las oficinas judiciales con el estimado por el modelo de simulación a través del indicador “M_g” (cantidad de procesos resueltos por cada juez) y el “Flow time” (tiempo que tarda el o la juez en llevar un proceso desde el registro hasta el estado de resolución).

3.3. Análisis de escenarios

Una vez validado el modelo, se simularon dos escenarios con cuatro niveles diferentes de productividad incremental del UPP (del 5%, 10%, 15% y 20%) para identificar la estructura organizativa que más redujera los tiempos promedio de los procesos civiles.

En el primer escenario, se simuló una asistencia de la U.P.P. únicamente a los y las jueces con menor rendimiento (Tabla 2), mientras en el segundo, se simuló una asistencia indiscriminada de la U.P.P. a todos los y las jueces de la sección (Tabla 3). En particular, este último resultó ser el escenario en el que se logra la reducción más significativa de los tiempos de los procesos.

<i>Trib. di CT, 4ª Sez. civ.</i>		
	<i>Flowtime</i>	<i>Relative Percentage Deviation</i>
Stato attuale	1062	0.00%
5%	963	9.32%
10%	943	11.21%
15%	924	12.99%
20%	742	30.13%

Tabla 2: Escenario 1

<i>Trib. di CT, 4ª Sez. civ.</i>		
	<i>Flowtime</i>	<i>Relative Percentage Deviation</i>
Stato attuale	1062	0.00%
5%	933	12.15%
10%	884	16.76%
15%	835	21.37%
20%	786	25.99%

Tabla 3: Escenario 2

3.4. La inteligencia artificial y la administración de justicia en el RIA

Los resultados del proyecto Just Smart evidencian que el empleo de técnicas de simulación en los procesos de toma de decisiones resulta de particular interés también en la administración de justicia.

Además, no pueden ignorarse las potencialidades de la IA, tal como, por lo demás, se ha planteado en la literatura (Di Martino, Esposito, & Colucci Cante, 2021). Sin embargo, su posible empleo debe tener necesariamente en cuenta el RIA.

En particular, una IA entra en el ámbito de aplicación del RIA solo cuando pueda definirse como

un sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras el despliegue, y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales (art. 3, ap. 1, n. 1, RIA).

En resumen, es necesario que el sistema cumpla siete requisitos, tal como se interpretan en las directrices de la Comisión Europea (European Commission, 2025):

- 1) *Automatización* – desarrollo y ejecución mediante componentes de *hardware* y *software*;
- 2) *Autonomía* – operabilidad con niveles variables de independencia respecto a la intervención humana;
- 3) *Adaptabilidad* – la capacidad (posible, pero no necesaria) de autoaprendizaje y de adaptar su comportamiento;
- 4) *Finalidad* – la consecución de objetivos explícitos (codificados directamente por el desarrollador en el sistema) o implícitos (deducidos del conjunto de datos de entrenamiento o de las interacciones ambientales);
- 5) *Deducción* – inferir cómo generar resultados a partir de los datos de entrada, de manera similar a los enfoques *knowledge-based* o de *machine learning*, y a diferencia de los sistemas *model-based* utilizados para la ejecución automática de operaciones (optimización matemática, procesamiento básico de datos, enfoques heurísticos clásicos, sistemas de predicción tradicionales);
- 6) *Predicción, creación, recomendación, decisión* – la generación de resultados como predicciones (estimación de un valor desconocido a partir de valores conocidos), contenidos (materiales, por ejemplo, texto, imágenes, vídeo, música), recomendaciones (sugerencias específicas y adaptadas a los datos de *input*) o decisiones (conclusiones o elecciones producidas en el entorno circundante sin intervención humana);
- 7) *Influenciabilidad* en entornos físicos o virtuales.

A decir verdad, la amplitud de la definición permite hipotetizar el empleo de numerosas tipologías de IA y, al mismo tiempo, excluir del ámbito de aplicación del reglamento a todos los simuladores mencionados en el apartado 2.

Una vez constatado que estamos ante una IA que se encuadra dentro de la definición del RIA, es necesario comprender en qué categoría de riesgo se clasifica. De hecho, el reglamento modula el régimen normativo de la IA en función del nivel de peligro para los derechos fundamentales, resultando en una cuádruple división en sistemas que conllevan: un riesgo inaceptable y, por tanto, prohibidos (Capítulo II); un riesgo alto y, por consiguiente, admisibles solo si cumplen ciertos requisitos

(Capítulo III); un riesgo limitado y, por tanto, sujetos a obligaciones de transparencia específicas (Capítulo IV); un riesgo bajo y, por consiguiente, voluntariamente subordinables a las prescripciones de los códigos de conducta (Capítulo X).

Excluyendo que los sistemas de IA utilizados para simular los tiempos de la justicia puedan encuadrarse dentro de las prácticas prohibidas (art. 5), parece necesario verificar si los mismos pueden considerarse de alto riesgo, para lo que es necesario que cumpla dos condiciones (art. 6, ap. 1, RIA):

- a) que el sistema de IA esté destinado a ser utilizado como componente de seguridad de un producto que entre en el ámbito de aplicación de los actos legislativos de armonización de la Unión enumerados en el anexo I, o que el propio sistema de IA sea uno de dichos productos, y
- b) que el producto del que el sistema de IA sea componente de seguridad con arreglo a la letra a), o el propio sistema de IA como producto, deba someterse a una evaluación de la conformidad de terceros para su introducción en el mercado o puesta en servicio con arreglo a los actos legislativos de armonización de la Unión enumerados en el anexo I.

Dado que las IA de la tipología en examen no cumplen dichos requisitos, solo queda verificar si las mismas no se encuadran en la presunción contemplada en el artículo 6, apartado 2, que califica expresamente como "de alto riesgo" las IA empleadas en el ámbito de la administración de justicia, es decir, aquellos sistemas destinados a ser utilizados por una autoridad judicial, o en su nombre, para ayudar a una autoridad judicial en la investigación e interpretación de hechos y de la ley, así como en la garantía del cumplimiento del Derecho a un conjunto concreto de hechos, o a ser utilizados de forma similar en una resolución alternativa de litigios (Anexo III, n. 8, let. a), RIA) es decir cuando los resultados de los procedimientos de resolución alternativa de litigios surtan efectos jurídicos para las partes (considerando 61, RIA).

A la luz de dicha definición, debería concluirse que las IA de qua no se encuadran en la presunción del artículo 6, ap. 2, ya que no están dirigidas a la

investigación e interpretación de los hechos y del derecho o a la aplicación de la ley a un conjunto concreto de hechos (Inturri, Fichera, & Costa, 2025). A lo sumo, podrían entrañar un riesgo moderado y estarán sujetas a las obligaciones de transparencia (art. 50, ap. 1, RIA).

A pesar de la clasificación del RIA, parecería que las IA en cuestión representan un riesgo más que moderado respecto, en particular, al principio de independencia judicial. De hecho, estas IA, al igual que los modelos analizados en el apartado 2, simulan los tiempos de la justicia a partir del rendimiento de los y las jueces individuales. En otras palabras, si la finalidad directa de estas IA es simular los tiempos de los procesos, la indirecta es perfilar a los y las jueces. Una perfilación que genera tensiones con la independencia de los y las jueces individuales, tanto en su dimensión externa, respecto a otros poderes del Estado y, en particular, a los Ministerios de Justicia, como en la interna, respecto a las presiones de los compañeros en funciones directivas o de autogobierno (Pizzorusso, 2019, p. 500).

En definitiva, el análisis permite vislumbrar las limitaciones del RIA para lograr una protección efectiva de los valores jurídicos fundamentales del Estado de derecho.

4. Conclusión

El presente estudio ha puesto de manifiesto la creciente importancia del principio de la duración razonable del proceso y cómo este puede garantizarse de forma racional utilizando modelos de simulación para prever el impacto de las decisiones organizativas de las oficinas judiciales en la duración de los procesos.

A tal fin, el trabajo se ha centrado en la exposición de los resultados del proyecto Just Smart de la Universidad de Catania, en el marco del cual se ha desarrollado un modelo de simulación de eventos discretos para evaluar el impacto en la duración de los procesos de las decisiones organizativas del U.P.P. en la cuarta sección civil del tribunal de Catania.

Los resultados del proyecto y el análisis de la literatura en la materia plantean las potencialidades del empleo de la IA en el sector. Al mismo tiempo, el

uso de una tecnología similar en la administración de justicia plantea importantes cuestiones jurídicas a la luz de la entrada en vigor del RIA.

Más en detalle, el RIA permite abstractamente hipotetizar el empleo de diversos sistemas de IA para simular la duración de procesos. Sin embargo, estas herramientas implicarían, a lo sumo, un riesgo moderado, al no ser subsumibles en la categoría de sistemas de alto riesgo. Esta circunstancia parece ser fuente de tensiones con el principio de independencia de la judicatura y, más en general, pone de manifiesto las limitaciones de protección del RIA.

En conclusión, el análisis deja abiertas futuras investigaciones relativas al desarrollo concreto de sistemas de IA del tipo planteado y a todas las problemáticas jurídicas consiguientes.

BIBLIOGRAFÍA

- Boyle, S., Guerin, S., Pratt, J., & Kunkle, D. (2003). *Application of agent-based simulation to policy appraisal in the criminal justice system in England and Wales*.
- Consiglio Superiore della Magistratura. (2024). Circolare sulla formazione delle tabelle di organizzazione degli uffici giudicanti per il quadriennio 2026/2029.
- Consiglio Superiore della Magistratura. (2024). Linee guida per l'Ufficio per il Processo ex art. 50 del decreto legge 24 giugno 2014, n. 90 – Monitoraggio delle esperienze, modelli organizzativi e buone prassi, www.cms.it.
- Di Martino, B., Esposito, A., & Colucci Cante, L. (2021). Multi agents simulation of justice trials to support control management and reduction of civil trials duration. *Journal of Ambient Intelligence and Humanized Computing*.
- European Commission. (2025). Commission Guidelines on the definition of an artificial intelligence system. 6 february 2025.

- Fjelstul, J. C., Gabel, M., & Carrubba, C. (2022). The timely administration of justice: using computational simulations to evaluate institutional reforms at the CJEU. *Journal of European Public Policy*.
- Inturri, P., Fichera, S., & Costa, A. (2025). La disciplina dei sistemi di intelligenza artificiale per l'amministrazione della giustizia nel Regolamento (UE) 2024/1689. *Lavoro Diritti Europa* (1).
- Pizzorusso, A. (2019). *L'ordinamento giudiziario*. Napoli: Editoriale Scientifica.
- Servizio Studi della Camera dei deputati. (2014). Documentazione per l'esame dei progetti di legge, Schede di lettura n. 196/1, 30 giugno 2014, www.camera.it.
- Servizio Studi della Camera dei deputati. (2022). Le misure per la Giustizia nel Piano nazionale di ripresa e resilienza, 27 settembre 2022, www.camera.it.
- Silvano, C. (2024). L'organizzazione amministrativa della giustizia in Italia tra esigenze di efficienza e garanzia di giusto processo: il caso dell'Ufficio per il processo. 26 giugno 2024. federalismi.it.
- Van Donge, W. (2017). *Modelling the Public Sector. Using Discrete Simulation Modelling (DSM) for the analysis of throughput times of court cases in the Justice system of the Netherlands*. 12 maggio 2017.

¿Qué decisiones no deberíamos dejar en manos de la inteligencia artificial (IA)?

Daniela Morales Lopes

Universidad de Oviedo

RESUMEN

El avance de la IA ha impulsado su aplicación en diversos ámbitos, incluyendo la toma de decisiones en sectores como la justicia, la sanidad y las finanzas, pero también en tareas cotidianas. Sin embargo, la creciente autonomía de estos sistemas nos lleva a tanto sobre su uso tácito como institucionalizado y, especialmente, sobre qué decisiones no deberíamos delegar en la IA. Para ello, partimos del fenómeno que Sánchez y Torrijos (2024) han llamado *ilusión de la decisión*. Este concepto describe el riesgo de percibir los resultados generados por sistemas algorítmicos como decisiones objetivas, infalibles y neutrales. Esta percepción puede conducir a una delegación acrítica del juicio humano en modelos de IA que no “comprenden” ni “deciden”, sino que generan estimaciones probabilísticas basadas en datos que, por otra parte, presentan sesgos. Por tanto, es importante distinguir entre la apariencia de inteligencia que proyectan los modelos generativos y sus limitaciones reales, como la falta de explicabilidad y la posibilidad de error. Frente a estos desafíos, se propone avanzar hacia una integración de la IA basada en marcos de transparencia técnica y operativa, auditorías para el reconocimiento y mitigación de sesgos, ofrecer formación

técnica y cognitiva para profesionales que utilicen la IA institucionalmente, educar a la ciudadanía en el uso de esta nueva tecnología y garantizar una valoración transdisciplinar del diseño e impacto de los sistemas de IA. El objetivo es contribuir al debate sobre la IA y la toma de decisiones poniendo el foco en la responsabilidad humana.

PALABRAS CLAVE: inteligencia artificial (IA), toma de decisiones, ilusión de la decisión, transparencia, educación ciudadana, justicia

1. Introducción

La integración de la inteligencia artificial (IA) en la toma de decisiones ha despertado gran interés social y académico. A medida que estas tecnologías amplían su presencia en distintos sectores como la sanidad, la justicia o las finanzas, surgen interrogantes sobre su papel en estos procesos que, tradicionalmente, habían estado reservados al juicio humano. Por este motivo, consideramos pertinente abrir un espacio de reflexión articulado en torno a las siguientes cuestiones: ¿Existen límites razonables al uso de la IA en la toma de decisiones? ¿Qué criterios pueden ayudarnos a pensar más claramente cuándo y cómo es deseable el uso de estos sistemas?

En primer lugar, conviene distinguir entre el uso tácito y el uso institucionalizado de la IA. Por uso institucionalizado nos referimos a sistemas de IA que resuelven tareas en determinados sectores laborales. Son usos que van acompañados por protocolos específicos y por supervisión profesional, como los sistemas de diagnóstico en hospitales, herramientas de evaluación de riesgo financiero o la optimización de rutas en logística (Balbaa y Abdurashidova, 2024). El uso tácito ocurre cuando las personas interactúan con estos sistemas en el ámbito privado para mejorar su desempeño o experiencia en todo tipo de actividades, desde redactar un correo formal o explorar posibles argumentos para afrontar una conversación difícil, hasta planificar rutinas de ejercicio personalizadas.

Una dificultad importante que plantea el uso de la IA para la toma de decisiones es que, dada la creciente sofisticación de los modelos generativos de lenguaje para simular la comprensión humana, las personas pueden no llegar a ser

plenamente conscientes de las limitaciones de dichos sistemas (Dentella et al., 2024). Herramientas como ChatGPT (OpenAI, 2024), Claude (Anthropic, 2024), Gemini (Google, 2024) o Copilot (Microsoft, 2024) no se presentan como una mera ayuda técnica, sino como voces que aconsejan, orientan y ofrecen múltiples soluciones ante un mismo problema, lo que puede generar una confianza excesiva.

2. La ilusión de la decisión

Sánchez y Torrijos (2024) denominan ilusión de la decisión al fenómeno de atribuir a los sistemas algorítmicos una autoridad incuestionable, considerando sus resultados como respuestas absolutas. La ilusión de la decisión se alimenta de la percepción de que los sistemas de inteligencia artificial, considerados popularmente como herramientas tecnológicas avanzadas, son más objetivos que las personas a la hora de tomar decisiones. Sin embargo, esta visión ignora que los sistemas de IA, pese a su aparente capacidad de respuesta, en su estado de desarrollo actual, no “deciden” ni “comprenden” (Echeverría y Almendros, 2024).

A diferencia de los seres humanos, que basan sus decisiones en un proceso cognitivo que incluye la intuición y el conocimiento tácito, los sistemas de IA operan a partir de datos y patrones, sin que podamos reconocer en ellos ninguna forma de comprensión subjetiva. Los resultados producidos por los sistemas de IA no son categóricos, sino estimaciones probabilísticas que dependen de los datos con los que estos modelos son entrenados. Estos resultados están condicionados por patrones identificados en vastos volúmenes de información, lo que significa que operan en unos márgenes de confianza.

Esta idea de no dejarse envolver por la ilusión de la decisión se ve reforzada si tomamos en consideración la complejidad que encierra el funcionamiento de estos modelos. La explicabilidad se refiere a la comprensión y justificación del proceso mediante el que un sistema de IA llega a determinadas conclusiones. Sobre todo en los modelos basados en aprendizaje profundo, podemos observar la característica de “caja negra”: la complejidad de las redes neuronales impide que tanto desarrolladores como usuarios puedan rastrear claramente el razonamiento que subyace a una salida (Adadi y Berrada, 2018). Esta incapacidad de comprender

con certeza cómo y por qué se llega a una determinada conclusión limita nuestras posibilidades de cuestionar o corregir errores.

Por otra parte, los sistemas de IA tampoco son necesariamente más neutrales. Al igual que ocurre en las dinámicas de interacción humana, estos modelos presentan diversos sesgos. Su entrenamiento depende de datos que reflejan los prejuicios existentes en las estructuras sociales y culturales de los contextos en los que se recogen. La calidad de los datos de entrenamiento conduce a sesgos relacionados con el género, la raza o la clase social, entre otros. Un claro ejemplo de esto son los sesgos raciales observados en algunos programas de reconocimiento facial cuando los sistemas de IA no incorporan suficiente diversidad en los datos (Buolamwini, 2019).

3. Implementación de la IA en procesos de toma de decisiones

A la luz de lo anterior, en vez de preguntarnos si la IA debería decidir en nuestro lugar —ya hemos visto, por ahora esto es tan solo una ilusión— quizás sería más pertinente que nos cuestionemos bajo qué condiciones las recomendaciones de estos sistemas pueden formar parte legítima de una cadena de toma de decisiones híbrida en la que las personas son el último eslabón (Kim, 2020). No se trata de rechazar de manera tajante el valor que estas tecnologías pueden aportar, sino de reflexionar cuidadosamente sobre los factores que deben tenerse en cuenta al decidir en qué medida implementarlas en los distintos sectores laborales y en otras actividades cotidianas.

4. Transparencia técnica y operativa

Un primer aspecto a tener en cuenta en la reflexión sobre la incorporación de la IA en la toma de decisiones es la transparencia (Huang, Zhang, Mao, y Yao, 2023). La falta de explicabilidad de los modelos dificulta que las personas comprendan cómo se toman las decisiones, lo que despierta en ellas preocupación y debilita su confianza en estas herramientas (Akbar et al., 2024; Osasona et al., 2024). Es el caso de los diagnósticos médicos hechos por sistemas de IA, donde la falta de explicabilidad puede hacer que los pacientes sientan inseguridad y no confíen en los resultados obtenidos.

A continuación, vamos a distinguir dos sentidos complementarios en los que entendemos la transparencia: “transparencia técnica” y “transparencia operativa”. Respecto a la transparencia técnica, una de las prioridades en el diseño de modelos de inteligencia artificial debería ser el avance en las ciencias de la explicabilidad o técnicas de Inteligencia Artificial Explicable (XAI). Este campo busca desarrollar métodos que permitan interpretar y justificar las decisiones de los sistemas algorítmicos, generando explicaciones comprensibles para las personas usuarias sin comprometer la precisión del modelo. Entre las técnicas más utilizadas se encuentra LIME (Local Interpretable Model-agnostic Explanations), que permite explicar decisiones individuales de modelos complejos mediante representaciones simplificadas y comprensibles localmente (Adadi y Berrada, 2018).



Figura 1: Representación gráfica de aspectos a considerar en el debate sobre la integración de la IA en la toma de decisiones

Por otra parte, si el proceso de decisión no es transparente respecto a la rendición de cuentas —transparencia operativa—, asignar responsabilidad en caso de errores o daños se vuelve complejo, como en el caso de accidentes a raíz de la decisión tomada por el sistema de un vehículo autónomo. La transparencia operativa implica clarificar quiénes diseñan estos sistemas de IA, con qué finalidades y bajo qué criterios de responsabilidad lo hacen como garantía de que podemos utilizar un sistema de IA para la toma de decisiones (UNESCO, 2020).

Desde la doble perspectiva de la transparencia técnica y operativa, nos situamos en una mejor posición para plantearnos en qué ámbitos puede ser más problemático el uso de estas tecnologías. Tal es el caso del discurso público automatizado, donde la generación de información errónea por IA puede influir de manera opaca en la deliberación democrática. Pero, también del ámbito militar, en el que la lógica autónoma de la IA introduce un alto nivel de incertidumbre sobre las estrategias y tácticas militares adoptadas por los contendientes. Aquí la preocupación reside en el incremento de la eficiencia para acabar con las vidas enemigas y en la dificultad para atribuir responsabilidades por dichas muertes. En estos escenarios, sobre todo en el último, no solo se compromete la trazabilidad de las decisiones, sino que se erosiona la capacidad de las personas para gestionar las consecuencias (Kissinger, Schmidt y Huttenlocher, 2021).

5. Auditorías para la mitigación de los sesgos

El siguiente aspecto se refiere a la calidad de los datos. Es importante reconocer las limitaciones que tiene la idea de una eliminación total de los sesgos. Pensar que podemos deshacernos de cada sesgo sencillamente contribuiría a alimentar expectativas irreales sobre la objetividad de los sistemas de IA. En su lugar, parece más apropiado dirigir la mirada hacia la mitigación de los sesgos en la medida de lo posible.

La auditoría de sistemas de IA ofrece un marco metodológico interesante para abordar este asunto. Siguiendo a Mökander (2023), podemos distinguir auditorías de tipo técnico (*technology-oriented audits*) —centradas en la evaluación del comportamiento del sistema frente a distintos datos de entrada— y auditorías de procesos (*process-oriented audits*), que examinan la calidad de los mecanismos organizativos y de gestión de los proveedores tecnológicos. La combinación de ambas permite, además de detectar resultados discriminatorios, analizar cuáles podrían ser sus causas estructurales. Dado que los sesgos emergen tanto de los datos como de los procesos de desarrollo, las auditorías permiten intervenir de forma continua a lo largo del ciclo de vida de los sistemas. El objetivo no es certificar que se cumple con unos lejanos estándares de perfección técnica, sino documentar las limitaciones en la calidad de los datos, fomentar la

trazabilidad y promover la responsabilidad operativa. Solo si reconocemos la presencia de sesgos en estos modelos podremos adoptar mecanismos para su mitigación activa.

6. Formación y educación ciudadana

Apostar por la incorporación de sistemas de IA para la toma de decisiones requiere educar a la ciudadanía en un uso responsable. Esto implica formar a los equipos que desarrollan estos sistemas, así como capacitar a las personas usuarias y colectivos afectados para que comprendan el alcance, las limitaciones y las implicaciones éticas de la IA (High-Level Expert Group on AI, 2019).

Cuando la IA se emplea de manera institucionalizada, el personal debería adquirir las habilidades técnicas necesarias, pero también competencias cognitivas para formular las preguntas pertinentes y para interpretar esos resultados. Para que la interacción persona -IA de buenos resultados, además de que los sistemas proporcionen explicaciones, es importante aumentar el esfuerzo cognitivo de las personas usuarias. Si las explicaciones son procesadas de forma superficial, se corre el riesgo de que la presencia de una justificación se confunda con una garantía de fiabilidad o certeza. Determinar cuándo debe presentarse la asistencia de la IA y qué tipo de información debería ofrecer al público usuario es crucial para que este pueda evaluar críticamente la ayuda recibida. Investigaciones recientes (véase Gajos y Mamykina, 2022) sugieren que es más conveniente mostrar primero la explicación antes que la predicción, con el fin de que el análisis de la información no se vea condicionado por un resultado anticipado. También se han propuesto sistemas interactivos que permitan a la persona explorar, mediante el lenguaje natural, el razonamiento detrás de una decisión del modelo, lo que mejoraría la comprensión de los resultados y la confianza en el propio sistema (Steyvers y Kumar, 2023).

El uso invisible o tácito de la IA plantea retos no menos sutiles. A medida que estas tecnologías se vuelven más accesibles, especialmente a través de interfaces conversacionales, se pierde trazabilidad sobre cómo se emplean y en qué medida influyen en la toma de decisiones de quienes las utilizan. Ante esta situación, podría pensarse que lo único que queda es esperar un uso responsable

de estas tecnologías. Sin embargo, esas esperanzas pueden reforzarse si se educa al público en inteligencia artificial. La ciudadanía debería estar concienciada de que, pese a la capacidad para simular el lenguaje humano de estos modelos, su funcionamiento sigue siendo limitado. En este contexto, además de los cursos y programas de educación digital, adquiere relevancia la transparencia —tanto en su vertiente técnica como operativa— por parte de las corporaciones responsables del desarrollo en IA. Conviene que estas instituciones anuncien de forma explícita en sus plataformas que los sistemas no son infalibles. Además de figurar en los “términos y condiciones”, debería presentarse un avis recurrente —como el recordatorio que ofrece ChatGPT (OpenIA, 2024): “ChatGPT puede cometer errores. Verificar la información es importante”—.

7. Colaboración transdisciplinar

Finalmente, es imprescindible apostar por una colaboración transdisciplinar que integre múltiples perspectivas, como las de profesionales de la filosofía, sociología o psicología. La incorporación de estas personas expertas en los equipos de desarrollo de sistemas de IA enriquece la comprensión de los dilemas que despiertan los avances en este campo. Las experiencias profesionales de estas personas pueden resultar muy valiosas a la hora de plantear las preguntas adecuadas sobre los fines, los valores y los riesgos asociados a la adopción de estos sistemas en procesos de toma de decisiones, así como para evaluar el impacto que esta implementación tiene en la sociedad. Los debates sobre los retos tecnológicos en los que participan las voces de distintos actores son una gran herramienta para orientar la innovación tecnológica hacia una dirección socialmente más justa. Para ello, sería importante incorporar diversidad —de género, clase y otras dimensiones— en el desarrollo de los modelos de IA (Floridi et al., 2018).

8. Conclusiones

A lo largo de este trabajo hemos explorado las complejidades que plantea el uso de la IA en la toma de decisiones. Aunque estas tecnologías nos ofrecen nuevas posibilidades, también nos sitúan ante riesgos significativos. Resulta legítima, entonces, la preocupación por su impacto, pero sería un error considerar que las decisiones que toman los sistemas de IA están completamente desvinculadas de la

acción humana. Cada decisión automatizada es, en última instancia, la consecuencia de decisiones humanas anteriores: qué datos utilizar, desde qué valores se ha diseñado el sistema y qué voces han participado en este diseño. Desmontar la ilusión de la decisión implica reconocer la responsabilidad humana que se esconde tras el velo de la complejidad algorítmica. Informar con honestidad sobre su funcionamiento y dotar a la ciudadanía de competencias técnicas y cognitivas es clave para evitar caer en una delegación ciega del juicio.

No todos los ámbitos son igualmente aptos para la implementación de sistemas de IA. El discurso público y el terreno militar representan casos donde, más allá de la conveniencia técnica, habría que repensar sus implicaciones sociales y políticas. En ambos escenarios, se erosiona nuestra capacidad de asumir las consecuencias de nuestras decisiones y se difumina la posibilidad de exigir responsabilidades. En lugar de concebir la IA como un sustituto del juicio humano, es preciso entenderla como una herramienta que amplifica nuestras capacidades, pero que requiere ser encuadrada dentro de marcos que velen por una integración lo más responsable posible. Con esta propuesta, no pretendemos agotar el debate, sino seguir fomentando el intercambio de perspectivas sobre el papel de la IA en la toma de decisiones. Lo que está en juego aquí no es simplemente el desarrollo de una tecnología, sino el tipo de sociedad que perseguimos construir con ella.

BIBLIOGRAFÍA

- Adadi, A., y Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Akbar, M. A., Khan, A. A., Mahmood, S., Rafi, S., y Demi, S. (2024). Trustworthy artificial intelligence: A decision-making taxonomy of potential challenges. *Software: Practice and Experience*, 54(9), 1621–1650. <https://doi.org/10.1002/spe.3216>
- Anthropic. (2024). Claude [Large language model]. <https://www.anthropic.com>

- Balbaa, M., y Abdurashidova, M. (2024). The impact of artificial intelligence in decision making: A comprehensive review. *EPRA International Journal of Economics, Business and Management Studies*, 11, 27–38. <https://doi.org/10.36713/epra15747>
- Buolamwini, J. (2019, February 7). Artificial intelligence has a problem with gender and racial bias. *TIME*. <https://time.com/5520558/artificial-intelligence-racial-gender-bias/>
- Dentella, V., Günther, F., Murphy, E., et al. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, 14, 28083. <https://doi.org/10.1038/s41598-024-79531-8>
- High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. European Commission. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- Echeverría, J., y Almendros, L. (2024). Inteligencias artificiales colectivas: u europeo. *Cuadernos del GESCyT*, (1). Universidad de la República.
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... y Schafer, B. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Gajos, K. Z., y Mamykina, L. (2022). Do people engage cognitively with AI? Impact of AI assistance on incidental learning. En *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)* (pp. 1–13). ACM. <https://doi.org/10.1145/3490099.3511138>
- Google. (2024). Gemini [Multimodal AI assistant]. <https://deepmind.google/technologies/gemini/>
- Huang, C., Zhang, Z., Mao, B., y Yao, X. (2023). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, 4(4), 799–819. <https://doi.org/10.1109/TAI.2022.3194503>

- Kim, H. S. (2020). Decision-making in artificial intelligence: Is it always correct? *Journal of Korean Medical Science*, 35(1), e1. <https://doi.org/10.3346/jkms.2020.35.e1>
- Kissinger, H. A., Schmidt, E., y Huttenlocher, D. (2021). *The age of AI and our human future*. John Murray Publishers.
- Microsoft. (2024). Copilot [AI-powered assistant]. <https://www.microsoft.com/en-us/microsoft-copilot>
- Mökander, J. (2023). Auditing of AI: Legal, ethical and technical approaches. *Digital Society*, 2(49). <https://doi.org/10.1007/s44206-023-00074-y>
- OpenAI. (2024). ChatGPT [Large language model]. <https://chat.openai.com>
- Osasona, F., Amoo, O., Atadoga, A., Abrahams, T., Farayola, O., y Ayinla, B. (2024). Reviewing the ethical implications of AI in decision-making processes. *International Journal of Management y Entrepreneurship Research*, 6, 322–335 <https://doi.org/10.51594/ijmer.v6i2.773>
- Steyvers, M., y Kumar, A. (2024). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, 19(5), 722–734. <https://doi.org/10.1177/17456916231181102>
- Torrijos, C., y Sánchez, J. C. (2023). *La primavera de la inteligencia artificial: creatividad y lenguaje en una nueva era tecnológica*. Los Libros de la Catarata.
- UNESCO. (2020, October 19). Does AI make better decisions than humans? Thinking ethics of AI [Video]. YouTube. <https://www.youtube.com/watch?v=2E7l1hdjHsg>

¿Qué hacer frente al *peak data*?

Manuel García Domínguez

Universidad Carlos III de Madrid

Consejo Superior de Investigaciones Científicas

Lorién Jiménez Martínez

Universidad de Zaragoza

RESUMEN

Desde los años setenta, sabemos que no existen economías de crecimiento constante que no terminen por encontrar límites en este tercer planeta del sistema solar, por lo que en su desarrollo llegará a un punto máximo o *peak*, al cual le seguirá una fase de decrecimiento. Desde esta premisa, se han desarrollado los estudios sobre *peaks* en distintas áreas: *peak* mineral, *peak* energético, *peak* demográfico... Sin embargo, en la mayoría de acercamientos a la inteligencia artificial, se sigue asumiendo que la digitalización es un proceso imparable de crecimiento sin límites. En una síntesis entre las investigaciones en economía biofísica y los estudios en filosofía de la tecnología, este capítulo pretende señalar y superar los errores que comete esta postura, pese a todo, hegemónica.

Para ello, se hará un movimiento iterativo de ida y vuelta entre la noción de límite y el complejo digital. Concretamente, se abordarán dos límites materiales al

desarrollo de éste. Por un lado, el límite en las reservas de minerales y otros recursos necesarios para la existencia y correcto funcionamiento de la infraestructura digital, agotadas por un ejercicio creciente de extractivismo y la impotencia material para la reintroducción de residuos en la cadena de valor. Este primer acercamiento nos permitirá probar la existencia de un cénit o peak digital, al que le seguirá un proceso de decrecimiento, confrontando al paradigma tecnooptimista de una digitalización sin límites. Por otro lado, el choque frontal entre los residuos generados y externalizados por la industria digital (desde las emisiones de carbono a los desechos electrónicos) y los límites planetarios dentro de los cuales se puede desarrollar la vida. Este segundo acercamiento politiza la necesidad de un decrecimiento digital y aleja la posibilidad de actuar como un sujeto automático, señalando la urgencia de una contracción en el uso de las tecnologías digitales y de una disputa política por adelantar el cénit digital antes de que sea demasiado tarde.

Este capítulo cambia así el ámbito del debate de la inteligencia artificial: de un diálogo sobre cómo debiera ser su uso y su relación con los derechos de privacidad o propiedad, a un debate sobre su misma existencia y los conflictos y tensiones entre su infraestructura y el mismo derecho a la vida. Por ello, rompe con la tradición naif que asume la digitalización como presupuesto de sus reflexiones y, en el mismo ejercicio de crítica, propone una superación de esta postura a la luz de la profunda crisis ecosocial en la que nos encontramos.

PALABRAS CLAVE: Decrecimiento, desdigitalización, crisis ecosocial, límites planetarios

A menudo, se dice que para creer que la economía puede crecer indefinidamente en un sistema finito hay que ser un loco o un economista. Hoy en día, añadiríamos a esa pequeña lista todo un conjunto de profesiones, entre las que podríamos incluir a filósofos e ingenieros. Desde hace tiempo, uno no sabe con precisión sobre qué se está hablando en muchos papers científicos o ensayos filosóficos. Cabría preguntarles, como reza aquel meme: «¿tu objeto de estudio está aquí, entre nosotros?». Y es que a menudo la investigación surge sobre una especulación futura sin mayor base que la de un consenso poco fundamentado: el

de una digitalización sin límites o una inteligencia artificial cuya capacidad puede multiplicarse hasta donde queramos.

Resulta habitual que, ante cualquier límite que podamos imaginar, surja un ingeniero pregonando un supuesto artefacto para superarlo y, tras él, un filósofo preocupado por las implicaciones que ese artefacto tendrá en nuestras vidas. Pero ni el ingeniero ni el filósofo suelen señalar los límites materiales de sus promesas futuras (por ejemplo, la disponibilidad de minerales y energía en este planeta para llevarlo a cabo) ni atender a los impactos ecosociales que provocan las tecnologías presentes (por ejemplo, la contribución de los centros de datos a la crisis climática). Pensando en objetos que nunca llegarán, descuidamos los objetos que a día de hoy nos rodean o, dicho de otro modo, “la pregunta dominante, «¿qué construiremos mañana?», nos impide ver las preguntas que deberíamos hacernos sobre nuestra responsabilidad actual por lo que construimos ayer” (Dourish y Mainwaring, 2012, p. 6). Esta postura se sostiene sobre dos errores que merecen ser atendidos y subsanados (Almazán, 2018, p.206):

El primero, no identificar a las fuerzas productivas como fuerzas a su vez destructivas, en tanto que dependen para su expansión de la devastación del planeta, sus habitantes y las formas de vida que se sitúan fuera de su lógica. [...] El segundo [...] consiste en no entender que toda organización material, y en concreto todo avance tecnológico, es radicalmente no neutral, y ver cómo los supuestos logros irrenunciables de la sociedad industrial no son separables de las nocividades del mundo que los ha engendrado.

1.

A mediados de la tercera década del siglo, no podemos seguir considerando las tecnologías industriales como fuerzas solamente productivas, sino como fuerzas productivo-destructivas (Sacristán, 1983). Si comprendemos las tecnologías únicamente en su dimensión productiva, como herramientas que permiten, facilitan o catalizan la transformación de recursos en bienes y servicios, el debate se focalizará en cómo orientar estas herramientas para producir de forma justa. En el caso de la inteligencia artificial, los debates atenderán a cómo evitar sesgos en el entrenamiento de modelos, cómo coser la brecha digital e incluso

cómo educar a los niños en competencias digitales. En todos estos casos, dirán, de haber algún elemento destructivo en la inteligencia artificial, será aquel que se deriva del mal uso o diseño de estas herramientas y, añadirán convencidos: la inteligencia artificial será una herramienta potencialmente positiva si la sabemos encauzar.

Quienes abanderan esta postura suelen reducir la tecnología al objeto técnico o a alguna de sus partes. En el caso de la inteligencia artificial, se comprende únicamente como una aplicación, una web, un programa de software o un asistente virtual. El alto grado de tecnificación del proceso de producción, así como una conexión imperceptible entre los dispositivos digitales que compramos y la infraestructura que los produce, nos ofrecen algunas razones para entender el por qué de esta postura. Pero aun así faltarían otras tantas, relacionadas con las estrategias de producto de las grandes tecnológicas, la ausencia de cobertura mediática e incluso la lógica misma del modo de vida imperial, que

oculta sistemáticamente las condiciones de producción –de la extracción de recursos naturales a las condiciones laborales de los trabajadores– que le permiten externalizar los impactos negativos de las operaciones del capital a regiones periféricas del mundo (Brand y Wissen, 2021, p.9).

Sin embargo, este relato restringe la inteligencia artificial a una parte minúscula de lo que es realmente. Un acercamiento materialista a la inteligencia artificial supone entenderla como una industria global formada por distintos complejos tecnológicos, así como estructuras políticas y militares para sostenerlos: yacimientos mineros, refinerías y fundiciones, fábricas de componentes manufacturados y ensambladores, distribuidores, toda la infraestructura doméstica y de Internet (centros de datos, cables submarinos, torres de telefonía, satélites...), recolectores de e-waste, cementerios electrónicos... Si atendemos a toda la cadena de valor de la inteligencia artificial, encontramos que cada uno de estos procesos no es únicamente productivo, sino eminentemente destructivo (Crawford, 2023). Pongamos un ejemplo poco conocido: el impacto de los centros de datos, cuyo consumo energético, hídrico y mineral , así como la contaminación acústica y el uso abusivo de suelos, sacrifica los territorios de allí donde se asientan, y dificulta cualquier proceso de transición ecosocial.

En términos generales, la destrucción asociada a la digitalización puede clasificarse en dos tipos: destrucción de las reservas de minerales (u otros, como los combustibles fósiles) a través de un rápido proceso de degradación material, y la destrucción de los ecosistemas a un ritmo muy superior al de su recuperación.

2.

Del primer tipo de destrucción se puede sacar una conclusión inmediata: la explotación creciente de reservas finitas, con una tasa de reciclaje que orbita el veinte por ciento, supone que en algún momento habrá escasez de minerales para los componentes y de combustibles fósiles o yacimientos de uranio para su demanda energética, lo que llevará inevitablemente a un proceso de decrecimiento digital. Esto implica que viviremos un peak data o cénit digital en las próximas décadas, al que le seguirá un proceso de progresiva o repentina desdigitalización. Dicho de otro modo, las expectativas de crecimiento tecnológico, sobre la que se basan la mayoría de proyecciones económicas, ingenieriles e incluso filosóficas, no están materialmente fundamentadas: no hay un planeta que pueda sostener, al menos indefinidamente, una inteligencia artificial que controle cada rincón de nuestro mundo (de nuevo, no por su consumo particular, sino por su consumo incrustado, encarnado, integrado...). Como decíamos al comienzo, su objeto de investigación no está ni puede estar entre nosotros, pero lo que sí están ya aquí son las consecuencias ecológicas de lo construido hasta ahora.

Del segundo tipo de destrucción surge un imperativo urgente: no podemos esperar a que ese cénit digital llegue y es que, para entonces, los daños generados por el complejo digital pueden llevar a puntos irreversibles en la destrucción de nuestro medio e incluso a poner en riesgo nuestras mismas condiciones de vida en este nuestro planeta. Los informes de la comunidad científica son claros: cada año de inacción y retardismo ecosocial multiplica los daños que sufrimos y sufriremos (IPCC, 2023): necesitamos una contracción urgente de esta economía digital y digitalizada.

3.

En este punto, suelen surgir voces en todas las partes del espectro político diciendo que la solución a todos estos conflictos está precisamente en la inteligencia artificial, que permitirá la desmaterialización de algunos procesos, diseñará mejores políticas climáticas y, sobre todo, aumentará la eficiencia hasta tal punto que podamos seguir pensando en un crecimiento constante con un consumo decreciente de recursos.

Bajo esta postura tecnosolucionista, la crisis ecosocial no es una crisis, sino una suerte de desafío tecnológico; y, del mismo modo que la aplicación de la tecnología ha sido una de las principales causas de los problemas ecosociales, dirán, también que constituye su principal solución (Morozov, 2016). Sin embargo, esta postura, pese a ser hegemónica, no ha logrado ningún avance en términos históricos en la mitigación de la crisis ecosocial, como vemos con el progresivo empeoramiento de los distintos índices. Es decir, aun siendo la postura más extendida desde que se puso en relieve este conflicto, no hay evidencia empírica del desacoplamiento de la economía con la degradación medioambiental (Parrique et al., 2019) ni de que el aumento de la eficiencia haya derivado en una reducción en el consumo de recursos (Alexander y Rutherford, 2019). En este punto, con el clavo ardiendo entre las manos, dirán que el hecho de que no haya ocurrido no significa que la nueva gran revolución de la inteligencia artificial no vaya a cambiar las cosas. Y aquí, aún confrontando argumentos fundamentados material e históricamente con puras especulaciones tecnológicas, deberemos hacer un esfuerzo por ver por qué los mismos fallos que nos hicieron perder aquellas cinco valiosas décadas se prolongan hacia la inteligencia artificial.

Hay dos críticas que se pueden hacer: una crítica hacia los efectos de un aumento de la eficiencia en modelos de crecimiento y una crítica a que la inteligencia artificial suponga verdaderamente, en términos generales, un aumento de la eficiencia. La primera ha sido ampliamente abordada por Samuel Alexander y Jonathan Rutherford (2019), quienes desmontan la conocida hipótesis Kuznets, sobre la que se basan, directa o indirectamente, todos los defensores de la apuesta tecnosolucionista. Según esta hipótesis, la relación entre la evolución tecnológica y

su impacto ambiental (ligado al consumo de recursos naturales) tiene la forma de una «U» invertida. Este razonamiento, sin embargo, se enfrenta a distintos problemas, desde la absoluta ausencia de evidencia histórica a un marcado carácter especulativo, aunque quizás el más acuciante sea la confusión entre el desacoplamiento relativo y el absoluto.

El primero refiere a la disminución del impacto ecológico por unidad de producción económica y el segundo a la disminución del impacto ecológico general de toda la producción. La contradicción de la eficiencia bajo un paradigma crecentista se da cuando los avances en el desacoplamiento relativo no se invierten en mantener la producción disminuyendo el consumo, sino en aumentar la producción hasta el punto de que “los beneficios de las mejoras de eficiencia se vean anulados parcial e incluso totalmente por el crecimiento del consumo que fue posible gracias a dichas mejoras” (Alexander y Rutherford, 2019, p.14). Por ejemplo, incluso asumiendo que un disco duro contamine menos que diez mil libros (lo cual es un cálculo parcial, aislado, y un mal reflejo de las complejas dinámicas que dan lugar a un disco duro y que el disco duro da lugar en la sociedad, así como del ciclo de vida y la localización de la producción de ambos artefactos, el libro y el disco duro) (Moreno, 2021), esa reducción se podría haber visto anulada por una multiplicación exponencial del número de documentos digitales en nuestras vidas, así como de dispositivos electrónicos y plataformas para leerlos.

Además, en esta y otras comparaciones, se suele caer en la falacia de la necesidad. Por ejemplo, quienes dicen que una videollamada contamina menos que un viaje en avión o que ver una película en streaming tiene menos impacto que comprarla, no dan cuenta que nuestro consumo no es independiente de la accesibilidad de ese consumo, es decir, que las tecnologías digitales no sólo resuelven necesidades, sino también las crean. Estas tecnologías no vienen a imponerse sobre un consumo desenfrenado de películas nuevas o viajes transatlánticos diarios, sino a modelos basados en bienes comunes y ritmos decelerados, como el alquiler de películas en las bibliotecas o videoclubs, o la llamada por teléfono habitual.

Y bueno, incluso poniéndonos a elucubrar mientras el planeta arde, y asumiendo que hemos introducido en la ecuación de la eficiencia todos los procesos productivos y destructivos del complejo tecnológico actual, con su correspondiente datificación del mundo, la acumulación y estructuración de tales datos, la logística global de producción y consumo, los territorios a los que se desechan la mayor parte de los productos electrónicos que se producen.... aún queda una segunda crítica por esgrimir: ese pensamiento es pretermodinámico.

Una vez introducido el sistema económico dentro del medio natural, las distintas leyes de rendimientos decrecientes, cuya consecuencia más amplia es que la tasa de aumento de la eficiencia no puede ser cada año mayor que el anterior, refutan la creencia razonable en un crecimiento ilimitado de la eficiencia (Georgescu-Roegen, 1996). A esto, hemos de sumarle la caída de la tasa de retorno energético, el agotamiento de las reservas accesibles de minerales críticos, la indisponibilidad de recursos hídricos e incluso personales o la entrada en escenarios climáticos inciertos. En esta línea, la imposibilidad de cerrar los ciclos de producción y desecho debido a la degradación de la energía útil, es decir, que no se pueda reciclar todo lo que se usa, nos invita a realizar economías de baja entropía basadas en la capacidad de los ecosistemas de cerrar la brecha entre los distintos ciclos. Por el contrario, el tipo de productos que demanda la inteligencia artificial (bajamente entrópicos) y el tipo de residuos que emite (altamente entrópicos e imposibles de reciclar por los ecosistemas actuales), terminan por incluir un factor decreciente y tendiente a cero independiente de cualquier intervención tecnológica, que trunca los deseos de una eficiencia que supere la unidad...

4.

Véanse las dificultades que encontramos al tratar de debatir estas posturas, y es que bajo la creencia de una omnipotencia virtual de la técnica, siempre se podrá apelar a una inteligencia artificial que haga todo consumiendo cada vez menos incluso llegando a una máquina de movimiento perpetuo, aunque para ello tengan que saltarse las leyes termodinámicas y bioquímicas más básicas. Nuestra sociedad aspira a seguir creciendo ansiosamente a la espera de una

desmaterialización absoluta que no llega (y, siendo honestos, nunca llegará), pero el tiempo se agota y debemos decidir entre un decrecimiento controlado basado en la suficiencia o el tardío decrecimiento brusco resultado de un planeta en llamas. En términos coloquiales: si un familiar enfermara por la extralimitación en el consumo de cualquier sustancia, ¿reduciríamos aquella sustancia y trataríamos los daños generados con los medios que tuviéramos o esperaríamos a algún invento del cual no hay ninguna evidencia como única alternativa para salvar su vida mientras sigue, y sigue, enfermando?

Seamos claros, tras cinco décadas perdidas, no tenemos tiempo que perder y, precisamente, las nuevas tecnologías actúan bajo el paraguas del retardismo y el obstruccionismo climático, desplazando las soluciones hacia un futuro que estamos perdiendo y, entre tanto, manteniendo un modelo económico hecho a la medida de las grandes corporaciones económicas y políticas. Entiéndase que si hemos realizado este ejercicio de debatir con el pensamiento mágico no es porque sea de nuestro agrado, sino porque mientras permanezcamos atascados en esta imagen de las tecnologías digitales, no será posible ninguna transición ecosocial justa y, sobre todo, efectiva. Las tecnologías digitales, la inteligencia artificial, por sus demandas materiales y sus dinámicas internas, tienen los días contados (al menos con la extensión que tienen en la actualidad), pero la transición ecosocial debe proyectarse mucho más allá, pues ese es su objetivo: mantener la vida humana y no humana en este nuestro planeta, el único habitable por ahora y pese a todo.

Ahora bien, y antes de que nos lancen piedras a la cabeza diciendo que la digitalización es clave para el tratamiento actual de ciertas enfermedades o para la distribución de ayudas económicas, parece claro que estas tecnologías no pueden ni deben desaparecer de un día para otro, y será necesaria una lucha política para decrecer de forma justa en términos globales, manteniendo lo que quede en la zaranda después de la crítica. Pero para ello hace falta un diálogo abierto y democrático basado en un pluralismo radical, es decir, en la diversidad de opciones sobre las que decidir y no sólo bajo la aceptación implícita de la digitalización como vector de progreso. Nuestra pequeña aportación a este debate, que servirá a

modo de conclusión, será este breve programa de decrecimiento digital, que puede servir de matriz provocadora sobre la que iniciar un fructífero diálogo:

1. Descentralización de Internet en una infraestructura distribuida y sustitución de los centros de datos hipertecnológicos por tecnologías de red alternativas de baja tecnología (De Decker, 2015).
2. Ralentización de la velocidad de procesamiento y disminución del almacenamiento.
3. Abandono progresivo de aplicaciones que requieren grandes flujos de información constantes, como el streaming, las videoconferencias y los juegos en vivo, y sustitución por registros físicos (videoclubs, discos duros, dispositivos mp3/mp4...).
4. Establecimiento de juicios climáticos a las grandes tecnológicas, responsabilizándolas económica y políticamente de los daños ecológicos y sociales provocados por la masificación interesadamente imprudente de sus productos.
5. Comunitarización de las tecnologías digitales, abandonando la individualización de los móviles, ordenadores o televisores y apostando por modalidades colectivas. Por ejemplo, impulsando locutorios públicos, cosotecas informáticas, salas comunes en vecindarios...
6. Restricción del uso de tecnologías de big data e inteligencia artificial a cuestiones médicas, endureciendo los requisitos para su uso.
7. Eliminación de toda criptotransacción.
8. Aumento del periodo de vida útil de los productos electrónicos y reforma del diseño para facilitar el proceso de reciclaje: formatos modulares, persecución de la obsolescencia programada, reacondicionamiento...
9. Racionalización en el consumo de datos, de la misma forma que se racionalizan otros recursos en momentos de escasez.
10. Recuperación de técnicas humildes de comunicación, socialización y consumo (Almazán et al., 2024), creando un inventario de las técnicas con las que nuestros padres y abuelos satisficieron aquellas necesidades que ahora se resuelven digitalmente.

BIBLIOGRAFÍA

- Alexander, Samuel y Rutherford, Jonathan. (2019). A critique of techno-optimism. Efficiency without sufficiency is lost. En Kalfagianni, Agni; Fuchs, Doris y Hayden, Anders (Eds.). Routledge Handbook of Global Sustainability Governance. Routledge.
- Almazán, Adrián. (2018). La actualidad del ecologismo como propuesta de autonomía. En A. Almazán, C. Madorrán, E. Santiago y J. Riechmann (eds.). Ecosocialismo descalzo, pp. 185-214. Icaria.
- Almazán, Adrián; Garrido, Jesús; López, Pablo; González, Vanessa y González, Luis. (2024). Técnicas humildes para el decrecimiento. Ecologistas en Acción.
- Brand, Ulrich y Wissen, Markus. (2021). Modo de vida imperial. Vida cotidiana y crisis ecológica del capitalismo. Tinta limón.
- Crawford, Kate. (2023). Atlas de IA. Poder, política y costes planetarios de la inteligencia artificial. NED Ediciones.
- De Decker, Kris. (2015). Cómo construir una Internet de Baja Tecnología. Manchester. Low Tech Magazine.
- Dom, Ann; Evans, Joám; Meynen, Nick; Francesco, Diego, y Barczak, Piotr. (2021). El fin de la minería. Seas at Risk.
- Dourish, Paul y Mainwaring, Scott. (2012). Ubicomp's colonial impulse. UbiComp '12, pp. 1-6.
- Georgescu-Roegen, Nicholas. (1996). La ley de la entropía y el proceso económico. Visor.
- IEA (2024). Electricity 2024. Analysis and forecast to 2026. International Energy Agency.
- IPCC. (2023). Climate Change 2023. Synthesis Report. Summary for Policymakers. Intergovernmental Panel on Climate Change.
- Moreno, Félix. (2021). Peak memory. Alicante: Autoedición.
- Morozov, Evgeny. (2016). La locura del solucionismo tecnológico. Katz Editores.

Parrique, Timothée; Barth, Jonathan; Briens, François, Briends; Kerschner, Christian; Kraus-Polk Alejo; Kuokkanen, Anna y Spangenberg, Joachim. (2019). Decoupling Debunked. Evidence and arguments against green growth as a sole strategy for sustainability. European Environmental Bureau.

Sacristán, Manuel. (1983). Entrevista con Manuel Sacristán. *Mientras Tanto*, 16/17, 195–211.

Hacia una inteligencia artificial responsable: propuesta para un Centro de Excelencia en IA Responsable

Jaime Morán

HIBERUS

RESUMEN

La expansión de la inteligencia artificial (IA) en diferentes tipos de organizaciones (públicas, privadas y del tercer sector) plantea retos estructurales de cumplimiento legal, responsabilidad ética y sostenibilidad tecnológica. Para afrontar estos desafíos, desde el presente artículo se propone la creación de un Centro de Excelencia Legal, Normativo, Ético y Medioambiental en IA como una estructura organizativa integrada en las entidades que utilizan o desarrollan sistemas de IA. El objetivo del centro es operar como un punto único de contacto para abordar cualquier cuestión relacionada con el uso responsable de estas tecnologías, apoyando en la interpretación normativa, asesoramiento legal, prevención de riesgos, generación de conocimiento, capacitación interna y relación con organismos reguladores. Este Centro combina la experiencia técnica y jurídica con la necesidad de institucionalizar la gobernanza de la IA, convirtiéndose en un nodo de referencia dentro de la organización. En este documento se describen las

motivaciones, objetivos, alcance y funciones principales de este centro, incluyendo su rol como observatorio normativo, generador de conocimiento y soporte transversal a empleados y proveedores. La implementación de este modelo permite avanzar hacia una inteligencia artificial más segura, transparente, alineada con los valores democráticos y conforme a normativas tales como Reglamento Europeo de IA y otros marcos normativos vigentes.

PALABRAS CLAVE: IA responsable, Gobernanza, Cumplimiento, Ética, Normativa

1. Introducción

El uso de la inteligencia artificial se ha convertido en un componente central de la transformación digital de muchas organizaciones. Desde procesos automatizados de atención al cliente hasta sistemas predictivos aplicados en recursos humanos, salud, seguridad o marketing, los algoritmos desempeñan hoy un papel decisivo en la toma de decisiones que afectan a millones de personas en gran cantidad de áreas. Esta expansión, sin embargo, viene acompañada también de preocupaciones sobre los impactos éticos, legales, sociales y medioambientales de estas tecnologías (Floridi et al., 2018).

Los marcos regulatorios internacionales y europeos, como el Reglamento (UE) 2024/1689 de Inteligencia Artificial, la NIS2, el RGPD o la Directiva de servicios digitales, han introducido nuevas obligaciones para las entidades que desarrollan, implantan o utilizan sistemas de IA. Estas obligaciones incluyen desde requisitos técnicos como la trazabilidad, explicabilidad o control humano, hasta evaluaciones de impacto, documentación exhaustiva y mecanismos de supervisión permanente.

En este contexto, muchas organizaciones carecen de los recursos, el conocimiento o la estructura adecuada para afrontar este nuevo escenario normativo y ético de forma efectiva. Para abordar esta necesidad, desde hiberus se plantea la creación de un Centro de Excelencia Legal, Normativo, Ético y Medioambiental en IA, diseñado para integrarse en el tejido organizativo de empresas, administraciones o fundaciones que necesiten implantar IA de forma responsable y conforme a derecho.

Este centro, en aquellas organizaciones donde está implementado, actúa como infraestructura de soporte y gobernanza, articulando los aspectos técnicos, normativos y éticos en una única unidad especializada, con capacidad para acompañar a los equipos en todas las fases del ciclo de vida de los sistemas de IA.

2. Objetivos y alcance del Centro de Excelencia

El Centro de Excelencia tiene como objetivo general proporcionar a las organizaciones una estructura interna estable que centralice la gobernanza normativa, ética y medioambiental de los sistemas de inteligencia artificial. A diferencia de otros enfoques más fragmentados o reactivos, el centro se concibe como un dispositivo proactivo, transversal y autónomo que acompaña a los equipos desde la planificación estratégica hasta el mantenimiento de los sistemas desplegados.

Entre sus objetivos específicos se destacan los siguientes:

- Actuar como punto único de contacto para cualquier cuestión relacionada con la IA y su regulación.
- Aportar seguridad jurídica a los desarrollos técnicos, evaluando riesgos y garantizando el cumplimiento normativo.
- Construir una base de conocimiento que se mantenga al día sobre regulaciones, buenas prácticas y estándares éticos aplicables al uso de IA.
- Ofrecer soporte continuo a empleados, proveedores y socios estratégicos sobre cuestiones jurídicas y éticas relacionadas con la IA.
- Promover la cultura de la IA responsable dentro de la organización mediante formación, guías y asesoramiento personalizado.
- Asumir un rol de interlocución ante organismos reguladores, manteniendo al día las obligaciones emergentes y facilitando la adaptación a ellas.
- Documentar y auditar el funcionamiento de los sistemas de IA desde un enfoque de transparencia y rendición de cuentas.

El centro no solo tiene una vocación técnica o jurídica, sino que se articula como un espacio de diálogo y aprendizaje continuo que une tecnología, derecho, ética y sostenibilidad.

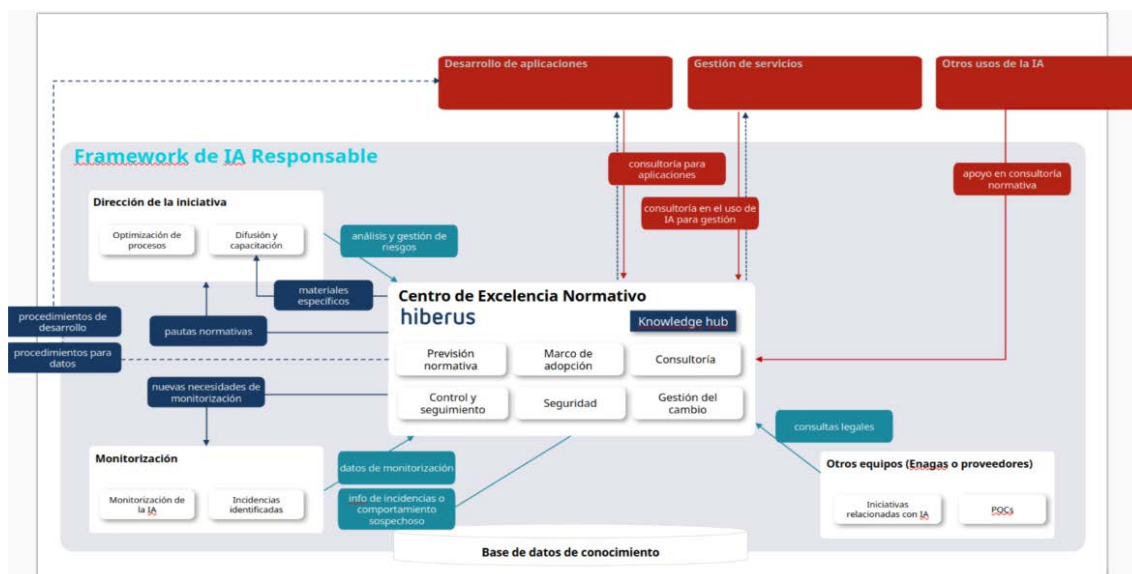


Figura 1: Esquema del Centro de Excelencia Legal, Normativo, Ético y Medioambiental de la IA.

HIBERUS

3. Funciones operativas del centro

El Centro de Excelencia tiene un diseño cuyo fin es cubrir una serie de funciones fundamentales, que responden tanto a las necesidades regulatorias como a las exigencias prácticas de las organizaciones. Estas funciones son complementarias y se articulan en torno a varios ejes estratégicos:

3.1. Observatorio normativo y regulatorio

Una de las principales funciones del centro es actuar como observatorio permanente del ecosistema normativo de la IA. Esto incluye:

- Seguimiento de nuevas regulaciones, estándares, directrices y jurisprudencia a nivel nacional e internacional.
- Análisis de impacto normativo sobre los sistemas en uso o en desarrollo dentro de la organización.
- Contacto directo con agencias reguladoras y autoridades supervisoras como la Agencia Española de Supervisión de la Inteligencia Artificial (AESIA), la Agencia Española de Protección de Datos (AEPD), la Comisión Nacional de los Mercados y la Competencia (CNMC) o la Comisión Europea.
- Elaboración de boletines, informes o alertas normativas adaptadas a los sectores específicos en los que opere la organización.

Esta función permite anticipar cambios regulatorios, reducir el riesgo de incumplimiento y garantizar la alineación constante entre los sistemas tecnológicos y el entorno legal.

3.2. Base de conocimiento compartida

El centro administra una base de datos de conocimiento que reúne información relevante y útil para los equipos que trabajan con IA. Esta base incluye:

- Compendios normativos (Reglamento IA, RGPD, sectoriales...).
- Guías y estándares éticos.
- Casos de uso, buenas prácticas y lecciones aprendidas.
- Documentación interna generada por la propia organización.

Este repositorio está disponible en formato de autoservicio y puede configurarse con distintos niveles de acceso en función de los roles organizativos para aquellos miembros de la organización y sus colaboradores. Su función es facilitar la toma de decisiones informada y fomentar una cultura organizacional centrada en la responsabilidad tecnológica.

3.3. Soporte a empleados y proveedores

El Centro de Excelencia funciona como punto de atención para cualquier empleado o proveedor que tenga dudas o necesite orientación sobre aspectos relacionados con la IA. Este soporte se caracteriza por:

- Atención directa a consultas técnicas, jurídicas o éticas.
- Resolución de dudas sobre proyectos concretos o aspectos generales del marco regulador.
- Emisión de informes o recomendaciones vinculantes en fases críticas de proyectos cuando la organización así lo requiera.
- Revisión de procesos, cláusulas contractuales o prácticas operativas con terceros.

Este servicio contribuye a extender la cultura de la IA responsable más allá del núcleo técnico o jurídico de la organización, involucrando a todas las áreas implicadas en el uso o gestión de sistemas inteligentes.

4. Apoyo a equipos técnicos

Uno de los pilares funcionales del Centro de Excelencia es el acompañamiento directo a los equipos técnicos que desarrollan, entrenan o integran sistemas de inteligencia artificial en sus proyectos. Lejos de posicionarse como un ente externo o fiscalizador, el centro es presentado como un colaborador estratégico que aporta criterios jurídicos, éticos y de sostenibilidad desde las primeras fases del ciclo de vida de cada proyecto.

Este asesoramiento técnico-jurídico se divide así:

- **Diseño preventivo:** inicialmente, el centro está en continua colaboración con los equipos para garantizar que los sistemas integren criterios de transparencia, control humano y no discriminación desde el principio. Se revisan los requisitos funcionales, la procedencia y calidad de los datos, y se anticipan posibles riesgos legales o éticos.
- **Supervisión de pilotos y pruebas de concepto:** en esta fase, el centro evalúa el comportamiento del sistema, identifica desviaciones no previstas y emite recomendaciones de mejora o de suspensión si detecta riesgos significativos para los derechos fundamentales o el cumplimiento normativo.
- **Implantación y mantenimiento:** una vez desplegado el sistema, el centro mantiene una función de seguimiento activo, validando documentación técnica, revisando logs, actualizando criterios legales y garantizando la trazabilidad de decisiones automatizadas.

Este acompañamiento técnico permite a las organizaciones reducir errores, acelerar tiempos de validación y aumentar la robustez jurídica y social de sus sistemas de IA.

5. Seguimiento y mejora continua

El Centro de Excelencia no opera de forma estática ni burocrática. Su diseño incluye un sistema de mejora continua basado en la recopilación y análisis de datos internos sobre el uso real de IA en la organización.

Entre las fuentes de información que alimentan esta mejora se encuentran:

- Inventarios de sistemas con IA implementados.
- Informes de funcionamiento, logs de uso y errores.
- Feedback de usuarios, clientes y personal interno.
- Historial de consultas recibidas y resoluciones emitidas.
- Registro de incidencias legales, técnicas o reputacionales.

Toda esta información se convierte en input para rediseñar procesos, actualizar materiales formativos, ajustar directrices y generar conocimiento interno útil para el conjunto de la organización. Este enfoque permite transformar el cumplimiento normativo en un proceso vivo, que no solo evita sanciones, sino que mejora la eficacia y legitimidad de las soluciones tecnológicas.

6. Integración con planes internos de *Compliance*

El Centro de Excelencia se alinea y refuerza los sistemas de cumplimiento normativo (*Compliance*) ya existentes en la organización. Para ello, adopta como referencia la metodología RACI (Responsible, Accountable, Consulted, Informed), mediante la cual se definen claramente los roles y responsabilidades de cada actor implicado en la gestión de sistemas de IA.

Esta integración implica:

- Definición de competencias para cada unidad (jurídica, técnica, auditoría, riesgos, dirección, etc.).
- Identificación de interlocutores únicos para cada tipo de consulta o intervención.
- Asignación formal de tareas en planes de despliegue, desarrollo o mantenimiento de sistemas IA.

- Creación de matrices de responsabilidades que faciliten el gobierno transversal de la tecnología.

Gracias a esta articulación, el centro no funciona como una isla normativa, sino como parte del ecosistema organizativo, complementando y reforzando la estrategia global de cumplimiento.

7. Proceso de análisis de brechas y gestión del riesgo

Uno de los procedimientos fundamentales del Centro de Excelencia es el análisis de brechas o GAP Analysis, que permite identificar las desviaciones entre el uso actual de la IA y las exigencias legales, éticas y medioambientales aplicables.

Este proceso se compone de varias fases:

- Inventario de sistemas IA: identificación exhaustiva de los sistemas actuales o en desarrollo, con análisis de su funcionalidad, grado de automatización, datos empleados, niveles de riesgo y destinatarios.
- Determinación del marco legal aplicable: se parte del Reglamento Europeo de IA, pero se amplía a otras normas vinculadas (protección de datos, mercados digitales, propiedad intelectual, responsabilidad civil, etc.).
- Evaluación de desviaciones: se revisan criterios como el consentimiento en la recogida de datos, la explicabilidad del sistema, la supervisión humana, la existencia de mecanismos de reclamación o la proporcionalidad del uso, entre otros.
- Valoración del riesgo: se clasifican los riesgos en función de su impacto (económico, reputacional, penal, etc.) y su probabilidad, aplicando la metodología de Compliance propia de la organización.
- Diseño de medidas correctoras: en función de los hallazgos, se proponen actuaciones para mitigar o eliminar los riesgos identificados, ya sea mediante rediseño, refuerzo documental o capacitación adicional del personal.

Además, se revisan aspectos relativos a la legalidad del tratamiento de datos, autoría, etiquetado, calidad, control humano, y todos aquellos que el Reglamento IA define como requisitos de alto riesgo. La Figura 2 representa el

entramado normativo que regula los sistemas de IA, incluyendo la normativa europea y sus interconexiones con otras leyes aplicables.



Figura 2: Esquema del marco legal general

8. Evaluación de riesgos por dominios

La evaluación de riesgos desarrollada por el centro se organiza por dominios temáticos. Esta categorización permite abordar de manera sistemática los distintos focos de riesgo y establecer controles adecuados para cada uno.

Los dominios principales definidos por hiberus incluyen:

1. Discriminación, toxicidad y desinformación
2. Privacidad y seguridad
3. Protección de activos
4. Fraudes y usos indebidos
5. Interacción humano-máquina
6. Impactos socioeconómicos y medioambientales
7. Seguridad, fallos y limitaciones de los sistemas de IA

Para cada dominio se establece un conjunto de puntos de control, que permiten identificar con precisión los elementos críticos. Por ejemplo, en el dominio de privacidad y seguridad, los puntos incluyen lo que recoge la tabla 1.

Estos controles se cruzan con las disposiciones del Reglamento de IA, permitiendo distinguir entre prácticas prohibidas, de alto riesgo o sujetas a vigilancia reforzada.

Privacidad y seguridad	
Introducción de datos en el entrenamiento	Consentimiento/base de legitimación
	Evaluación de Impacto en Protección de Datos
	Transferencias internacionales de datos
	Confidencialidad/acceso a datos por terceros no autorizados
	Exposición de menores de edad
Identificación/evaluación de personas	Identificación biométrica en tiempo real
	Análisis y tratamientos a partir de lo anterior
	Realización de perfilado de usuarios
	Segmentación de usuarios
	Segmentación de usuarios con datos enriquecidos
	Toma de decisiones automatizadas
Generación de contenidos con datos personales (verdaderos o falsos)	Intromisión en honor/imagen
	Suplantación de identidad

Tabla 1: Puntos de control en el dominio de privacidad y seguridad

9. Aplicación práctica y sostenibilidad del modelo

Uno de los principales valores añadidos del modelo propuesto es su vocación de permanencia. El Centro de Excelencia no se plantea como un recurso puntual ligado a un proyecto o contrato, sino como una estructura duradera que forma parte del modelo de gobernanza organizacional.

Su aplicación práctica permite a las organizaciones:

- Cumplir con los requisitos del Reglamento IA desde su entrada en vigor.
- Tener trazabilidad y documentación actualizada de todos sus sistemas de IA.
- Responder a auditorías internas o externas con solvencia.
- Prevenir sanciones, daños reputacionales y conflictos legales.

- Generar confianza entre empleados, ciudadanía y agentes externos.

Además, su sostenibilidad se basa en:

- Su alineación con estrategias de Compliance y transformación digital.
- Su valor añadido como generador de conocimiento y cultura organizativa.
- Su capacidad para adaptarse a distintos tamaños, sectores y niveles de madurez tecnológica.

El modelo puede implantarse de manera modular, comenzando por sectores críticos (sanidad, justicia, contratación pública, etc.) y ampliándose progresivamente al conjunto de la organización.

10. Conclusión

La gobernanza de la inteligencia artificial no puede delegarse únicamente en marcos normativos o certificaciones externas. Es necesario construir estructuras institucionales dentro de las organizaciones que hagan posible el despliegue seguro, ético y legal de estas tecnologías. El Centro de Excelencia Legal, Normativo, Ético y Medioambiental en IA que aquí se propone cumple esa función.

A través de un enfoque integral, transversal y operativo, este modelo permite pasar de la ética declarativa a la acción estructurada. Ofrece herramientas prácticas para que las organizaciones gestionen los riesgos asociados a la IA, cumplan con la normativa vigente, formen a su personal y se preparen para un futuro tecnológico más justo y sostenible.

La inteligencia artificial responsable no es un eslogan: es una arquitectura institucional que debe construirse desde hoy hasta en la organización más pequeña que hace uso de ella.

BIBLIOGRAFÍA

- AEPD, G. (2021). La protección de datos en las relaciones laborales. <https://www.aepd.es>
- Ai, H. (2019). High-level expert group on artificial intelligence. *Ethics guidelines for trustworthy AI*, 6.
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>.
- Parlamento Europeo y Consejo de la Unión Europea. (2024). *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial)*. Diario Oficial de la Unión Europea, L 211.

II. EL USO DE LA IA EN EL CAMPO DE LA SALUD

Participación ciudadana e inteligencia artificial en el ámbito sanitario

Alejandra Rivas-Carrero¹
Universidad de Salamanca

RESUMEN

En la medicina contemporánea, la participación activa del paciente (*patient engagement*) ha devenido fundamental ante los desafíos epistemológicos, éticos y praxeológicos que plantea el uso de la inteligencia artificial (IA) en el diagnóstico y tratamiento de enfermedades. Las metodologías tradicionales de resolución de problemas, como la ciencia básica, la ciencia aplicada y la consultoría profesional, han perdido eficacia en una actualidad marcada por incertidumbres y riesgos derivados del propio avance científico y tecnológico. La sociedad que nos rodea, ampliamente conocida como “sociedad del riesgo”, reclama una metodología que incluya las voces de todos los afectados y las afectadas, que se vea respaldada por una “comunidad de pares extendida” dentro del marco de la ciencia *posnormal*, y el *patient engagement* es una señal de la presencia de dicha comunidad en el ámbito sanitario. Como instrumento mediador entre pacientes y expertos, la IA puede fortalecer los componentes epistemológicos, éticos y praxeológicos de esta

¹ El proyecto que ha dado lugar a estos resultados ha contado con el apoyo de una beca de la Fundación “La Caixa” (ID 100010434). El código de la beca es LCF/BQ/DR24/12080004.

manifestación médica de la comunidad de pares extendida, pero también presenta límites, ya que omite elementos esenciales para la buena práctica médica, como la prudencia, la empatía y la deliberación.

PALABRAS CLAVE: *patient engagement*, inteligencia artificial (IA), sociedad del riesgo, comunidad de pares extendida, ciencia posnormal

1. Introducción

Las grandes transformaciones científico-tecnológicas por las que está pasando la sanidad contemporánea pone en el centro del debate la noción de *patient engagement*, que en castellano se traduciría como participación activa del paciente. Esta noción, que hace referencia a la implicación del público general en determinados procesos sanitarios y al bucle de retroalimentación que se genera entre dicho público y los especialistas (Rowe y Frewer, 2000), plantea interrogantes filosóficas de gran calado: ¿cómo contribuye el paciente al conocimiento médico?, ¿qué implicaciones éticas y praxeológicas tiene su involucramiento, especialmente en un contexto donde la inteligencia artificial (IA) comienza a desempeñar un papel relevante en el diagnóstico, pronóstico y tratamiento de enfermedades?, ¿puede la IA traer consecuencias para la autonomía del paciente, o para las conductas que este adopta en el manejo de su enfermedad?, ¿le hará seguir o no determinados tratamientos, cambiar o no determinados hábitos o hacer hincapié o no en determinados síntomas?

Resulta muy oportuno en este debate rescatar el concepto de dimensión. Las dimensiones de algo, como las dimensiones del espacio euclídeo, aunque puedan disociarse en un plano analítico, son ontológicamente inseparables. Lo mismo se aplica a las dimensiones de la cultura científica. Es habitual en la literatura especializada de los estudios sociales de la ciencia y la tecnología elaborar una diferenciación entre una dimensión epistémica, axiológica y praxeológica para analizar la cultura científica a nivel individual, a pesar de que en la praxis se solapan irremediablemente (Cámara Hurtado & López Cerezo, 2012). Como veremos, las preguntas anteriormente planteadas acerca del *patient engagement* pueden abordarse desde estas tres dimensiones, que, inevitablemente, se entrelazan en la experiencia clínica.

2. La noción de *patient engagement* en la sociedad del riesgo

Para comprender las razones por las que la noción de *patient engagement* está cobrando mucha fuerza, es necesario situarla en el marco de la “sociedad del riesgo”, un término acuñado por el sociólogo alemán Ulrich Beck (1986) para retratar una era en la que hay que gestionar riesgos introducidos por la propia modernidad, por el propio desarrollo científico y tecnológico, y que requieren, por ende, nuevas formas de actuar sobre el mundo.

La metodología que Kuhn caracterizaba como “ciencia normal”, firmemente anclada en la contrastación de hipótesis, en el razonamiento lógico-deductivo, en la neutralidad axiológica y en el gobierno de los “expertos”, es extremadamente limitada a efectos de gestionar los riesgos contemporáneos, de origen predominantemente científico y tecnológico. Nuestras nuevas “circunstancias”, como diría Ortega y Gasset (1965), no se entienden sin la tecnologización, y los riesgos que enfrentamos hoy no escapan a este proceso. Ante tales riesgos, tan desbordados por hechos dudosos, conflictos de intereses y la necesidad de toma de decisiones inmediata, y ante la tendencia actual en el discurso académico de valorizar la participación ciudadana en todas las esferas públicas, cobra mucho sentido el modelo de “ciencia posnormal” de Funtowicz y Ravetz (1993).

Una característica esencial de este modelo es la presencia de una “comunidad de pares extendida”. En tiempos posnormales —en este caso saturados por altos riesgos e incertidumbres de naturaleza científica y tecnológica— las estrategias tradicionales de resolución de problemas (la ciencia básica, la ciencia aplicada y la consultoría profesional) resultan alarmantemente ineficaces. Sobre este escenario la ciencia no puede hablar con una única voz. La comunidad de pares que evalúa la calidad de la investigación y las políticas de ciencia y tecnología no solo debe incluir a especialistas, sino también a personas usuarias de productos tecnológicos, a clientes de servicios sociales y a todos los miembros de la población en riesgo.

El *patient engagement* parece ser un claro indicio de que la ciencia posnormal también ha cruzado las fronteras de la medicina. Así, por ejemplo, en situaciones de alto riesgo e incertidumbre a la hora de tratar enfermedades

crónicas con tratamientos innovadores, como en la realización de trasplantes de células madre en casos del Parkinson (Lindvall, 2015), pacientes, familiares y otros grupos afectados han evaluado el conocimiento científico y tecnológico junto con personas “expertas”. Esta colaboración es particularmente necesaria en el contexto sanitario actual, donde la excesiva especialización fomenta la incertidumbre, la sensación de perderse en un bosque cada vez más intrincado, y donde el riesgo de excesiva medicalización está en auge ante la creciente inundación de la llamada “cascada iatrogénica” (Laporte, 2024). En semejante coyuntura, es preciso reivindicar el *patient engagement* por razones epistemológicas, porque mejora las técnicas de investigación y tratamiento en medicina; por razones éticas, dado que favorece el bienestar del paciente y refuerza su autonomía en las decisiones que toma sobre su propia salud; y por razones praxeológicas, en vista de que juega un papel clave en la mitigación de ciertos riesgos tecnológicos, como aquellos suscitados por los avances en radiología diagnóstica, que a menudo convierten tumores benignos en cánceres (Sitges Serra, 2020), o por el abuso de los psicofármacos, que frecuentemente generan nuevos trastornos mentales en lugar de tratar los anteriores (Pérez Álvarez, 2021). Quizás el secreto del *patient engagement*, lo que le provee de todos estos valores filosóficos, radica en su capacidad de aportar una perspectiva más global del asunto, que no solo recorre los síntomas, los efectos secundarios y las interacciones con otros medicamentos de forma más exhaustiva, sino que también contempla el impacto social, psicológico y funcional que la enfermedad ejerce sobre la vida cotidiana de la persona en tratamiento. Esta mirada holística dota de una riqueza filosófica singular a la participación activa de dicha persona.

3. Las fortalezas de la IA para el *patient engagement*

Hasta cierto punto, la IA podría respaldar la democratización de la ciencia en el ámbito sanitario, tanto desde una escala epistémica como ética y praxeológica. A través de bots conversacionales, por ejemplo, podría servir como instrumento de mediación entre la persona en tratamiento y los conocimientos especializados.

Un *chatbot* de salud es un programa informático diseñado para acceder a bases de datos sanitarias y simular diálogos con sujetos humanos en donde se generan respuestas personalizadas, usualmente operando a partir del procesamiento del lenguaje natural (PLN). Como tal, esta modalidad de inteligencia artificial podría contribuir al cierre de lo que Frickel y Vincent (2007) denominan *knowledge gap*, refiriéndose a la brecha entre lo que la persona en tratamiento conoce y lo que debería conocer para tomar decisiones informadas sobre su salud. Gracias a su capacidad para ofrecer explicaciones claras y distintas adaptadas a las necesidades individuales, los *chatbots* de salud pueden ayudar a esa persona a comprender mejor su condición médica, esto es, su etiología y pronóstico, así como a conocer de forma más íntegra las diferentes opciones de tratamiento disponibles, junto con sus riesgos y beneficios. Además, pueden orientarle sobre estrategias de autocuidado, promoviendo de esta forma una mayor responsabilidad en el manejo de su enfermedad.

En este sentido, la IA no solo fortalece la dimensión epistemológica del *patient engagement*, sino también su dimensión ética y praxeológica. Por supuesto, la adquisición de conocimientos por parte de la persona en tratamiento aguza su comunicación de esa perspectiva global que posee por naturaleza, por ser el quien sufre la enfermedad, pero estas cuestiones cognitivas le permiten, a su vez, gestionar ciertos riesgos sanitarios y experimentar, por consiguiente, más bienestar y autonomía en la toma de decisiones sobre su propia salud.

4. Los límites y desafíos de la IA para el *patient engagement*

No obstante, la inteligencia artificial en salud también presenta obstáculos para la participación activa de la persona en tratamiento. Si queremos entenderlos, debemos recurrir a la distinción filosófica entre “inteligencia artificial débil” e “inteligencia artificial fuerte” de John Searle (1980).

Por un lado, la IA débil alude a máquinas que, como grandes actores capaces de interpretar un papel sin realmente sentirlo en su interior, simulan la inteligencia humana sin mayor consciencia. Un programa de ajedrez, por ejemplo, podría parecer inteligente en la medida en que despliega movimientos óptimos y estrategias complejas, pero no es más que el seguimiento mecánico de reglas

preestablecidas. Las inteligencias artificiales fuertes, en cambio, serían máquinas que van más allá de la mera imitación. No solo simularían la inteligencia humana, sino que verdaderamente poseerían intencionalidad, emociones y entendimiento. Representarían, por lo tanto, la respuesta afirmativa a la pregunta clásica de Turing “¿puede pensar una máquina?”.

A los ojos de Searle, no existe tal cosa como la inteligencia artificial fuerte; solo es posible la débil. Su argumento de la habitación china ilustra muy bien este punto. Se trata de un experimento mental en donde hay que imaginar a una persona que está dentro de una habitación y que sigue instrucciones para manipular símbolos chinos sin realmente comprender el idioma. Esto demuestra que, aunque una máquina parezca entender, en realidad no es así; simplemente procesa símbolos (Searle, 1980).

Así pues, en relación con lo expuesto, la posición por la que aquí se aboga es que, la IA en salud —el uso de *chatbots* para consultas médicas, el uso de aprendizaje profundo para la segmentación de tejidos, para la detección de microaneurismas o para el diseño de medicamentos, el uso de Python para la historia clínica, el uso de robots para intervenciones quirúrgicas, etcétera— es débil toda. ¿Por qué? Porque, a pesar de que evidentemente ejecutan ciertas tareas con éxito, lo hacen sin realmente comprender, sin vivir estados mentales genuinos.

Otra distinción filosófica que resulta muy pertinente a la hora de analizar los límites de la inteligencia artificial para el *patient engagement* es la distinción entre IA general e IA regional (o “estrecha”). El término “inteligencia general” fue acuñado por Charles Spearman en 1904 a partir de sus estudios sobre las capacidades cognitivas humanas y su teoría bifactorial de la inteligencia (Spearman, 1904), mientras que el de IA estrecha (*narrow AI* en inglés) surgió de manera gradual en la literatura científica para diferenciar los sistemas de IA especializados (como asistentes virtuales, motores de búsqueda y software de reconocimiento de imágenes) del ideal de la IA general (Long y Cotner, 2019). El caso es que, hasta el momento, las inteligencias artificiales solamente abordan

funciones específicas sin buscar incorporarlas a un sistema más global, con lo cual son siempre regionales, nunca generales.

Lo mismo se aplica al ámbito sanitario. Los asistentes virtuales, los robots cirujanos y cualquier otra forma de IA en salud, son regionales, porque están diseñados únicamente para ejecutar tareas específicas, bien definidas (como el aprendizaje lingüístico, el análisis de radiografías o la realización de operaciones quirúrgicas). Estas innovaciones tecnológicas carecen de la flexibilidad cognitiva de los humanos, que nos permite operar sobre una gran variedad de realidades técnicas.

Por el contrario, el personal médico en verdad comprende y vive estados mentales genuinos, es capaz de comprender a la persona en tratamiento por semejanza propia, de establecer empatía con ella, y desempeña una multiplicidad de funciones en diversos ámbitos técnicos. De ahí que a la medicina se le pueda catalogar como “ciencia de diseño”, como proponen Cuevas Badallo y Obdulia Torres (2022), porque es a la vez arte y tecnología, es arte en cuanto a la relación entre personas médicas y en tratamiento; y es tecnología en la medida en que se basa en diferentes ciencias para lograr un fin concreto.

Así, por ejemplo, la medicina se basa en las matemáticas, como cuando utiliza la estadística y el concepto de correlación. Es a este nivel formal, cuantitativo, en el que principalmente se mueve la IA, o por lo menos con más potencia. Pero este es un nivel puramente sintáctico, y la medicina es mucho más que eso. La fiebre, el colesterol alto y el insomnio se pueden correlacionar con muchas enfermedades, a buen seguro, pero apoyarse solo en esta información, además de poco eficaz, es peligroso. La medicina basada en la evidencia, que se centra mayoritariamente en datos cuantificables, no se aleja mucho de tal panorama. La confianza excesiva en la estadística, en lo cuantitativo, que podría incluso caracterizarse como una forma de “atrincheramiento tecnológico” (González García et al., 1994), tiende a dar como resultado que se deje de lado tanto lo que la medicina tiene de “arte” como las vivencias de la persona en tratamiento, lo cual conlleva a su vez, mucho margen de error y la excesiva medicalización. Todo ello conduce al fenómeno de *disease mongering* (propagación

de enfermedades, en castellano), a saber, la práctica de extender los criterios diagnósticos de las enfermedades y fomentar de modo intenso su concienciación pública, para aumentar el mercado de tratamientos.

Como es de suponer, otras ciencias sobre las que se basa la medicina son las ciencias naturales. El personal médico constantemente actúa en función de los resultados de investigación en teoría celular, histología, citología, inmunología, etcétera, los que pretenden construir líneas causales de manera íntegra. A la inteligencia artificial, sin embargo, al menos de momento, no se le ha abierto ninguna puerta para poder llevar a cabo este tipo de investigaciones científicas, debido a su condición de débil y estrecha.

Las ciencias sociales y humanas también juegan un papel importante en la medicina. Para empezar, el personal médico en cierto modo aplica también la psicología. La relación con la persona en tratamiento es fundamentalmente psicológica, y si desaparece, desaparece con toda la medicina. A pesar de que hay médicas o médicos que nunca ven a pacientes (porque se dedican a una cuestión sumamente especializada), lo que hacen termina cayendo en manos de otro personal médico que sí interactúa diariamente con pacientes, y es ese personal es el que trata. Es con ese personal médico que las personas en tratamiento pasan por el efecto placebo, el efecto nocebo, etcétera. No hay que olvidar, asimismo, que las enfermedades son procesos que, si bien arrastran una etiología somática, tienen repercusiones psicológicas. O a la inversa, una enfermedad puede ser causada por factores psicológicos, y, sin embargo, venir acompañada por síntomas somáticos.

La IA en salud, esto es, los lenguajes de programación, la cirugía robótica, los relojes inteligentes, etcétera, lamentablemente plantean una amenaza para esa relación, y, por tanto, para el *patient engagement*, cuyo espacio natural coincide con ella. De nuevo, amenazan lo que la medicina tiene de arte. Como artista además de tecnóloga o tecnólogo, el personal médico realiza una gran labor de interpretación de la experiencia del paciente y la moldea (o al menos debería hacerlo), pues su práctica facultativa requiere de la intuición, de la empatía y de una proyección individual que es muy difícil de estandarizar. En este aspecto, la medicina muestra similitudes significativas con respecto al mundo de la

psicoterapia, en donde la implicación de la persona en tratamiento es decisiva en dicho proceso.

Por otro lado, la medicina está impregnada de elementos culturales. Los propios conceptos de salud y enfermedad son culturales: dependen de los modos de vida, costumbres, el desarrollo científico, político, artístico, etcétera, en definitiva, de un contexto histórico en particular. La cultura que rodea al personal médico y la o el paciente va cambiando, lo que induce cambios en los patrones nosológicos: lo que ayer fue enfermedad hoy ya no lo es, y lo que hoy es enfermedad ayer no lo era. Esto es evidente en el caso de los trastornos mentales. La histeria, por ejemplo, ha pasado de ser una enfermedad ginecológica a ser una neurológica, con Charcot (Didi-Huberman, 2005), y posteriormente de ser una enfermedad neurológica a una psicológica, con Babinski y Froment (1918). No obstante, las categorías diagnósticas en el caso del resto de enfermedades (las enfermedades no mentales) también experimentan un dinamismo muy significativo en función de cambios culturales. No podemos olvidar lo sucedido, por ejemplo, en lo que respecta a la polio, la sífilis y la VIH, todas enfermedades que brotaron en un entorno cultural particular. Por lo tanto, es preciso señalar que los sistemas de inteligencia artificial actuales no integran actualmente, en su totalidad, estos factores contextuales que resultan determinantes para la formulación de estrategias de *patient engagement*.

Entre las diferentes caras de la profesión médica, tampoco podemos excluir la de la historiadora o el historiador. Cuando se afronta a la historia clínica para reconstruir la biografía de la persona en tratamiento, el personal facultativo debe emplear la metodología de la historiografía, dado que la historia clínica no puede constituir por sí sola el diagnóstico. De forma análoga a como se manejan reliquias y relatos conjuntamente, el personal médico no solo debe interpretar signos (lo que le dicen las máquinas), sino también síntomas, también las narrativas del paciente. Por desgracia, el estado actual de las ciencias de la computación omite la posibilidad de que la inteligencia artificial maneje una verdadera metodología cualitativa, en donde la voz del paciente también tenga un papel protagonista.

Finalmente, la medicina también lleva la filosofía en sus entrañas. La discusión acerca del uso de la inteligencia artificial en el ámbito sanitario no solo nos remite a la pregunta clásica “¿puede pensar una máquina?”, sino también a la pregunta “¿puede filosofar una máquina?”, dado que en el centro de la medicina está la ética. La ética es el ideal regulador kantiano de la medicina, es lo que guía y organiza la práctica facultativa médica hacia una meta que a menudo no se alcanza. Esa meta es, naturalmente, la de eliminar las enfermedades (en consonancia con la idea de medicina como ciencia de diseño, la medicina también es “arte sanador”), pero una y otra vez resulta inalcanzable debido a que sencillamente no hay cura, y por ende el personal médico debe conformarse con otras aspiraciones éticas, quizá menos pretenciosas, como las de paliar y acompañar. Ahora bien, las inteligencias artificiales que se utilizan actualmente en el ámbito sanitario no están reguladas por la ética de la misma manera que lo está un ser humano. Ciertamente, al nivel sintáctico de las ciencias formales, a saber, de las matemáticas y la lógica, los algoritmos de las inteligencias artificiales pueden aplicar lo que Anscombe (1957) llamaba “leyes éticas”, haciendo referencia a leyes morales universales como el imperativo categórico de Kant y el principio de utilidad de Bentham y Mill. Pero estas leyes en ética normativa son defectuosas, porque no pueden calcular la prudencia, la deliberación, ni las circunstancias concretas, obstruyendo espacios en los que las personas en tratamiento pueden tener un papel activo en la toma de decisiones y en la gestión de riesgos tecnológicos en el marco de su propia salud.

5. Reflexión final

El *patient engagement*, como exponente de la comunidad de pares extendida en el contexto sanitario actual, tiene implicaciones epistemológicas, éticas y praxeológicas que pueden verse alteradas por el uso de la inteligencia artificial para la investigación y tratamiento de enfermedades. Por un lado, como una herramienta más, al igual que la estadística, las ecografías, las imágenes por resonancia magnética, la tomografía axial computarizada, etcétera, la IA podría fortalecer esos atributos epistemológicos, éticos y praxeológicos del *patient engagement*. El problema es cuando la IA pretende sustituir a la práctica facultativa médica, lo cual desembocaría en una medicina reducida a sus componentes formales, arrebatada, por tanto, de componentes psicológicos, éticos y

contextuales, entre otros, imprescindibles para el *patient engagement*. Esto es lo que debemos evitar a toda costa.

BIBLIOGRAFÍA

- Babinski, J. & Froment, J. (1918). *Hysteria or Pithiatism, and Reflex Nervous Disorders in the Neurology of War*. (Trans. J. D. Rolleston). University of London Press.
- Beck, U. (2002). *La sociedad del riesgo: hacia una nueva modernidad*. Paidós Ibérica. Publicado originalmente en 1986.
- Cámara Hurtado, M., & López Cerezo, J. A. (2012). Political dimensions of scientific culture: highlights from the Ibero-American survey on the social perception of science and scientific culture. *Public Understanding of Science*, 21(3), 369-384. <https://doi.org/10.1177/0963662510373871>
- Didi-Huberman, G. (2005). *Invention of hysteria: Charcot and the photographic iconography of the Salpêtrière*. The MIT Press.
- Cuevas Badallo, A., & Torres González, O. (2022). Medicine as a science of design. *Azafea: Revista de Filosofía*, 24, 21-38. <https://doi.org/10.14201/azafea2022242138>
- Frickel, S., & Vincent, M. B. (2007). Hurricane Katrina, contamination, and the unintended organization of ignorance. *Technology in Society*, 29(2), 181-188. <https://doi.org/10.1016/j.techsoc.2007.01.007>
- Funtowicz, S. & Ravetz, J. (1993). Science for the post-normal age. *Futures*, 25(7), 739- 755.
- González García, M.,I., López Cerezo, J.,A., Luján, J.,L., & de Melo Martín, M.^a I. (1994). Las concepciones de la tecnología. *Arbor*, 149(585), 125-145. Recuperado de <https://www.proquest.com/scholarly-journals/las-concepciones-de-la-tecnología/docview/1301393118/se-2>
- Laporte, J. R. (2024). *Crónica de una sociedad intoxicada*. Península.

- Lindvall, O. (2015). Treatment of Parkinson's disease using cell transplantation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1680), 20140370. <https://doi.org/10.1098/rstb.2014.0370>
- Long, L. N., & Cotner, C. F. (2019). A Review and Proposed Framework for Artificial General Intelligence. *2019 IEEE Aerospace Conference*, 1-10. <https://doi.org/10.1109/aero.2019.8742125>
- Ortega y Gasset, J. (1965). *Meditación de la técnica: vicisitudes en las ciencias*. Espasa-Calpe.
- Pérez Álvarez, M. (2021). *Ciencia y pseudociencia en psicología y psiquiatría: Más allá de la corriente principal*. Alianza.
- Rowe, G., & Frewer, L. J. (2000). Public participation methods: a framework for evaluation. *Science, Technology, & Human Values*, 25(1), 3-29.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424. <https://doi.org/10.1017/S0140525X00005756>
- Sitges Serra, A. (2020). *Si puede, no vaya al médico*. Debate.
- Spearman, C. (1904). «General Intelligence», Objectively Determined and Measured. *The American Journal of Psychology*, 15(2), 201. <https://doi.org/10.2307/1412107>

El impacto psicológico de la IA

Carolina Tomás Franco

UNED

RESUMEN

El texto destaca la paradoja humana: somos capaces de crear tecnologías extraordinarias, pero aún no hemos ahondado con la profundidad necesaria en nuestra propia cognición. Queremos crear sistemas inteligentes sin tener, ya no un dominio, sino un conocimiento de cómo funcionan los procesos básicos en el ser humano tales como: la atención, la percepción o el control de impulsos, entre otros.

Como seres humanos, tomamos decisiones con relativa autonomía, ya que estamos dominados/as por aspectos como el ego, la envidia o necesidades de atención; de las que, en muchas ocasiones, ni siquiera somos conscientes.

Bajo estas estructuras que se esconden circunstancias que son aprovechadas por quiénes tienen mejor conocimiento de ellas. Lejos de ser perfectas, aparecieron en nosotros probablemente cumpliendo funciones de supervivencia que nos han llevado hasta aquí. Nos compete en ese sentido tener la humildad de sentirnos limitados y a la vez agradecidos porque lo que hoy puede parecer una tara en su día nos salvó la vida. Pero existen estructuras en nuestro funcionamiento, como pueden ser los sesgos cognitivos, que son aprovechados por el sistema y, muy especialmente, desde las tecnologías creadas en base a inteligencia artificial.

En este momento tan crucial, en el que la información es el bien máspreciado, debemos ser capaces de estar bien informados/as tanto en lo que ocurre en el exterior como en nuestro propio interior, para poder lidiar con todo el bombardeo que trata de captar nuestra atención y moldear nuestras emociones (Castells, 1996-1998). De manera que mantengamos nuestra autonomía de pensamiento y capacidad crítica, reconociendo que hemos mirado durante demasiado tiempo lo que ocurre fuera, y ahora toca aprender a mirar dentro. Para ello, una nueva educación es necesaria, que tenga la valentía de incorporar en la enseñanza qué significa ser persona y cómo trabajar en ese ámbito, al igual que nos enseña cómo cuidar nuestro físico o nuestro intelecto. De esta manera construiremos una estrategia de adaptación cultural ante la revolución tecnológica. Porque no se trata de luchar, ni de negar, ni de ensalzar: se trata de no perdersenos a nosotros mismo en el camino.

PALABRAS CLAVE: autoconocimiento, inteligencia artificial, sesgos cognitivos, educación, pensamiento crítico, dependencia tecnológica

1. La paradoja del conocimiento: dominamos más el mundo exterior que nuestro propio interior

Durante años vamos a la escuela, aprendemos matemáticas, lengua, ciencias... Todas las asignaturas tienen un gran componente de conocimiento del llamado “objetivo”, pero ¿qué sabemos de nosotros mismos? Sabemos cómo funcionan nuestros órganos, qué partes los componen (más o menos) pero nunca nos han enseñado cómo se estructuran nuestras creencias, cómo podemos gestionar nuestros propios pensamientos o emociones, si podemos (o no) trabajarlos de manera consciente o cómo hacerlo.

¿Qué hace que fulano/a (que, seamos sinceros, a veces es un poco cretino/a) se convierta en nuestro amigo/a, y que mengano/a (tan encantador/a) nos resulte simplemente “meh”? ¿Por qué si vamos caminando por el mismo sitio, mi pareja va viendo unas cosas y yo voy viendo otras completamente diferentes? ¿Si atendemos ciertas conductas inapropiadas estamos reforzándolas, pero si no las atendemos estamos generando necesidades no cubiertas y, por lo tanto, carencias emocionales?

¿Cómo funciona realmente mi inteligencia? ¿Y mi memoria? ¿Por qué cuando trato de expresar algo, salen de mi boca cosas que realmente no son lo que yo quería

decir? Estas preguntas sobre nosotros mismos no son nuevas, ni tienen que ver con la aparición de la inteligencia artificial, pero cobran un sentido especialmente urgente en este momento, dado que la digitalización de nuestros procesos cerebrales implicará cambios importantes en ellos, que se adaptarán y... quién sabe en qué se convertirán.

2. La urgencia del autoconocimiento en la era digital

Como seres humanos, desarrollamos cosas increíbles: desde gigantescos edificios o cohetes espaciales hasta máquinas que parecen pensar. Cuántos/as genios/as en la historia habrán creado grandes e impresionantes inventos y a la vez, sido incapaces de hablar con claridad (o simplemente con respeto) a alguien a quién amaban.

Esto nos ha ido haciendo más inteligentes en ciertos aspectos, pero, paradójicamente, menos en otros. Hemos ido delegando nuestras tareas al exterior sin dedicar demasiado tiempo a conocer qué implica exactamente eso para nosotros.

Imagínate que llegas a una ciudad desconocida sin tu aplicación de navegación correspondiente. ¿Sabrías llegar al hotel? ¿Podrías mantener la incertidumbre de sentirte perdido/a y arriesgarte a descubrir cosas nuevas? Si respondiste que sí a esta última pregunta, ¿lo has hecho con la convicción de que en realidad siempre tendrás la aplicación de navegación como respaldo? ¿Has llegado a vivir alguna vez esa experiencia? Quiénes tenemos algunos años ya, si, tuvimos una temporada de vivencia sin GPS; pero a las nuevas generaciones se les ha privado directamente la posibilidad de enfrentarse a esa incertidumbre y, por consiguiente, de desarrollar habilidades en referencia a esto. Estudios recientes demuestran que el uso habitual del GPS impacta negativamente en la memoria espacial durante la navegación autoguiada, afectando especialmente a las habilidades de navegación dependientes del hipocampo (Dahmani & Bohbot, 2020). Además, el uso de navegación paso a paso promueve una forma pasiva de navegación que no apoya el aprendizaje espacial, teniendo un impacto perjudicial en las habilidades de navegación humana y la cognición espacial (Aporta et al., 2021).

Con esta nueva revolución de la IA, somos conscientes de que puede ayudarnos en muchas cosas: grandes tareas pueden ser automatizadas gracias a sistemas muy complejos, grandes cantidades de datos pueden ser analizadas en muy poco tiempo. Pero ¿qué implica todo esto para nuestros sistemas, nuestros cuerpecitos y cerebritos biológicos, nuestra capacidad de emocionarnos o de relacionarnos? ¿Nos ayuda a

mejorar como seres humanos? ¿Y qué significaría “mejorar”? ¿Sentir o relacionarnos más? ¿O quizás menos?

3. Los sesgos cognitivos: huecos en nuestro diseño donde otros plantan sus semillas

Esta falta de autoconocimiento se vuelve especialmente problemática cuando consideramos cómo funciona realmente nuestro cerebro. Para entender mejor estos posibles cambios que pueden ocurrir, necesitamos hablar de algo fundamental: nuestros sesgos cognitivos. ¿Qué sabemos de ellos? ¿Qué son? ¿Cuántos tenemos? ¿Por qué tenemos sesgos? ¿Para qué? ¿Podemos eliminarlos? ¿Tienen utilidad? ¿Somos conscientes de nuestros propios sesgos, o solo vemos los sesgos en el otro?

El sesgo de confirmación (por poner un ejemplo) es considerado una tendencia universal en el ser humano (Nickerson, 1998). Implica que, si creemos tal o cual cosa, nuestro cerebro siempre va a tender a encontrar la información que corrobore nuestra creencia (y la encuentra). Y este sesgo es una de las grandes dificultades a la hora de disolver diferencias entre personas (Tversky & Kahneman, 1974). Porque los cerebros de cada una de ellas con diferentes creencias encuentran justificaciones (que, dicho sea de paso, no suelen ser válidas para quien piensa lo contrario).

Y es cierto que, a veces, la cosa se queda en que yo creo que te toca lavar los platos y tú crees que me toca a mí. Pero la cosa se complica cuando una parte de la población cree que se debe invertir más en armamento, mientras que otra parte de la población considera que esa inversión debe ir a sanidad. Y aquí entran en juego los grandes mecanismos de manipulación. Porque si el algoritmo siempre nos va a dar esa información que corrobora nuestra creencia, la va a reforzar y no vamos a tener una visión global de una situación que nos permita ser críticos. Las “burbujas de filtro” creadas por algoritmos refuerzan nuestras percepciones existentes y limitan nuestra exposición a información diversa (Pariser, 2011). Por otro lado, si no tenemos una creencia todavía bien formada (o bien informada) al respecto, el algoritmo nos la puede moldear. Si tal o cual empresa, quiere que creamos cierta cosa, solo tiene que darle al algoritmo la estructura para que sea eso lo que nos muestre. Si no tenemos las herramientas (o a veces, la energía vital) para contrastar esa información, esa será con la que nos quedemos. De igual modo, si creemos que todo lo que vemos, es todo lo que

hay, nuestra percepción de la realidad se vuelve cada vez más estrecha y manipulable. Pero también puede ocurrir que, simplemente, se busque minimizar la importancia de esos temas, manteniendo la distracción en otros planos, jugando ya no solo con nuestros sesgos, sino con nuestra atención, motivación y percepciones. “Pan y circo”, tan viejo y tan vigente, ya sabéis.

Y es que no es el sesgo de confirmación el único que está en el partido. El sesgo de disponibilidad nos dice que sobrevaloraremos más la información que sea más accesible (Tversky & Kahneman, 1974). El sesgo de anclaje se refiere al hecho de que pesará mucho más la primera información que tengamos disponible (Sherif et al., 1958; Tversky & Kahneman, 1974). ¡Aja! ¿Cuánto crees que esto se utiliza? Puede ser más o menos crucial, (aunque si somos mínimamente críticos, nos hacemos conscientes de que ni siquiera esto es irrelevante) si el primer anuncio del año es de una u otra marca de bebida refrescante. Pero ¿qué pasa cuando es la primera noticia, la portada del periódico, o el primer vídeo recomendado?

También, sabemos que existe el sesgo de autoridad (Milgram, 1963). Si creemos que tal o cual medio de comunicación es más verídico y que tal o cual lo es menos (sea real o no), tenderemos a buscar la información en un lugar o en otro. ¿Cuántas veces nos paramos a comprobar la veracidad de la información que nos llega? ¿Buscamos realmente información de diferentes visiones de la realidad para poder forjar opiniones más globales? ¿De qué depende de que algo se perciba como una autoridad? A veces, implica simplemente ponerse una bata blanca. En otras ocasiones, se percibe como una autoridad en un tema simplemente alguien que tiene muchos likes o quien tiene una especial capacidad para expresarse. A veces, no depende de que lo que se dice tenga realmente veracidad o no, sino simplemente que “parezca” que la tiene.

Y, por si esto no fuera suficiente, podríamos seguir así hasta completar más de 100 sesgos detectados de manera científica por la psicología (Kahneman, 2013). Y digo esto con plena conciencia de que incluso las ciencias pueden llegar a estar sesgadas.

4. Y las semillas plantadas por otros crecen en nuestro interior

Además de las opciones del hecho de que somos manipulables (y manipulados/as, no creamos lo contrario, nadie se escapa de ello en uno u otro ámbito), existen muchos riesgos con respecto a cómo las nuevas tecnologías están invadiendo

nuestro entorno. Ya desde la infancia está muy medido cómo deben estructurarse los colores, los tiempos que deben durar las secuencias, etc. De manera que atrapen tu atención (Heller, 2004; González Conde & Sueiro Domínguez, 2017). A día de hoy, conocer las tecnologías para ciertas generaciones tiene más que ver con hacer scroll infinito que con entender cómo funcionan.

Pero las nuevas tecnologías no impactan únicamente en los procesos más “orgánicos” sino también en aquellos más sociales, más relacionales. Estamos más conectados con el resto del planeta y podemos pasar horas hablando con alguien que vive en la otra parte del mundo, pero somos incapaces dar los buenos días a nuestra vecina en el ascensor. E incluso hay quien prefiere charlar con sistemas tecnológicos que con otros seres humanos (Turkle, 2015). ¿Qué lectura deberíamos sacar de este tipo de aspectos?

5. Y ante este panorama, ¿cómo podríamos lidiar con ello?

Decía Carl Jung que “quien mira afuera sueña, quien mira dentro despierta” (Jung, 1969). Y cuando decía esto, a mi entender, se refería a que tenemos que aprender a conocernos mejor, pero no solo nuestros gustos musicales o nuestro estilo de vestir, sino qué mecanismos guían nuestros impulsos, o nuestra atención, o seleccionan lo que se ha de guardar en nuestra memoria: aprender a reconocernos en nuestros miedos, en nuestras esperanzas y a poner en conciencia nuestro quehacer; tomar posesión de nuestros pensamientos y así poder guiar nuestros sentimientos, reconocerlos en todo nuestro cuerpo, y no solo como productos de una mente sino de un sistema completo.

Se sabe que recordaremos una información u otra en función de nuestra personalidad. Personas con tendencias más depresivas, recordarán los sucesos más tristes mientras que aquellas más optimistas, los más alegres (Ruiz-Caballero & Bermúdez, 1991; Wittekind et al., 2013). Pero tener una tendencia más depresiva u optimista, no depende solo de la voluntad o de las creencias de cada uno/a. Hay una gran parte que existe en nuestra propia fisiología (Beck et al. 2010). ¿Cuánto somos conscientes de las propias “jugarretas” que nos hace nuestro propio cerebro? Se ha visto que esto es reconocible y trabajable, de manera que nuestros patrones fisiológicos se pueden, hasta cierto punto modificar (Vázquez et al., 2010). Pero solo reconociendo, aceptando y con propuestas dirigidas, esto puede ocurrir. Si no, seguiremos dejando que

un piloto automático tome el control (muchas veces con conductas absolutamente autodestructivas incluso) y creemos que tenemos claro lo que hacemos, porque recordemos, si creemos que tenemos el control, vamos a encontrar información (real o no) que nos dé la razón.

6. “La loca de la casa”: entendiendo nuestro cerebro

Una vez tuve una profesora que al cerebro le llamaba “la loca de la casa”, porque, aunque pueda parecer contraintuitivo, nuestro cerebro no piensa, genera pensamientos hacia nuestra parte consciente, que es diferente. Si yo te digo “no pienses en una manzana” ¿qué imagen aparece en tu mente?

Nuestro cerebro está estratégicamente estructurado para ser eficiente y adaptativo, pero no es perfecto, ni mucho menos (Kahneman, 2013). Nuestro cerebro trabaja más con símbolos y emociones que con información puramente racional y esta parte es exclusiva de nuestro cerebro consciente. Pero para que la información sea consciente, antes ha tenido un procesamiento inconsciente y este, es el que genera lo que creemos que generamos de forma autónoma.

Antes comentaba: ¿Para qué sirve un sesgo? Pues probablemente, a lo largo de la historia, para salvarnos la vida. ¿Por qué la primera impresión se fija con más ímpetu? Pues probablemente, porque si vimos una vez a una serpiente morder a alguien y morir a las horas, tu cerebro no quiera recoger una muestra de participantes suficientemente representativa de la población para darle un peso científico al evento y poder concluir bajo un paradigma no refutable que si, que esa serpiente es venenosa.

7. Estrategia de defensa cultural: nuestro autoconocimiento

Y entonces, ¿qué podemos hacer? Quizás debamos comenzar por tener una mirada crítica incluso hacia aquello que pensamos como real o verdadero. ¿Esto lo hago porque realmente quiero o es mi ego dictando sentencia? ¿Realmente quiero hacer eso o es la envidia la que me lleva por un camino del que quizás me arrepienta? ¿Qué busco realmente cuando hago esto, atención, respeto? Igual es que nunca nos han enseñado a trabajar sobre este tipo de tareas autocríticas, reconocerlas, ponerles las palabras adecuadas y verlas con un pelín de objetividad.

Pero también se trata de entender que ese “mirar dentro” del que hablaba Jung, no debería ser solo una reflexión individual. En una sociedad donde la tecnología (y cómo esta se crea), moldea nuestros procesos cognitivos, necesitamos una respuesta colectiva y bien estructurada. Y, desde mi opinión, esa respuesta nace en la educación, una de calidad real, actualizada con la evidencia científica y la psicología de la educación, que entienda cómo aprenden los humanos y no solo qué materias se deben aprender, que tenga la valentía de incluir aspectos absolutamente novedosos como “aprender a ser” y que esté centrada en personas y, para ello, que cuente con los recursos necesarios.

Además, es crucial tener en cuenta cómo somos y cómo nos afecta el mundo que nos rodea. Es importante en este contexto ser consciente de cómo funciona la tecnología. Conocer de dónde viene la apariencia de “súper verdad” o “súper realidad” que nos ofrecen los sistemas IA, los cuáles debemos conocer cómo funcionan y tener en cuenta sus limitaciones. Ser conscientes de que hacernos la vida tan sencilla, solo puede acarrearos falta de capacidades a nosotros, y ventajas competitivas para otros. El autoconocimiento como estrategia social, se convierte así no solo en una herramienta de crecimiento personal, sino es una respuesta de adaptación cultural que debemos defender (Baumeister & Leary, 1995; Csikszentmihalyi, 2020).

BIBLIOGRAFÍA

- Aporta, C., Kritsch, I., Collignon, B., Gearheard, S., Binnie, D., Atwood, D., & Mead, E. (2021). Rethinking GPS navigation: creating cognitive maps through auditory clues. *Scientific Reports*, 11, 7764.
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3), 497-529.
- Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (2010). *Terapia cognitiva de la depresión* (19ª ed.). Desclée de Brouwer.
- Castells, M. (1996-1998) *La era de la información: economía, sociedad y cultura*. (Trilogía) Alianza Editorial.

- Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life*. Basic Books.
- Hachette UK. Dahmani, L., & Bohbot, V. D. (2020). Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Scientific Reports*, 10, 6310.
- González Conde, D., & Sueiro Domínguez, E. (2017). Uso del color en la infancia. *Revista de Estudios e Investigación en Psicología y Educación*, 4, 085-088.
- Heller, E. (2004). *Psicología del color*. GG.
- Jung, C. G. (1969). *The archetypes and the collective unconscious* (2nd ed.) Princeton University Press.
- Kahneman, D. (2013). *Pensar rápido, pensar despacio*. Debate.
- Milgram, S. (1963) Behavioral Study of obedience. *Journal of Abnormal and Social Psychology*, 67(4), 371-378.
- Nickerson, R. S. (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175-220.
- Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristic and biases. *Science*, 185(4157), 1124-1131.
- Vázquez, C. Hervás, G., Hernangómez L., & Romero, N. (2010). Modelos cognitivos de la depresión: una síntesis y nueva propuesta basada en 30 años de investigación. *Psicología Conductual*, 18(1), 139-165.
- Wittekind, C. Terfehr, K. Otte, C. Jelinek, L., Hinkelmann, K., & Moritz, S. (2013). Moodcongruent memory in depression- The influence of personal relevance and emotional context. *Psychiatry Research*, 215(3), 606-613.

Sobre la cuestión de la inteligencia artificial y la soledad en relación con la tercera edad

Candela Carolina Fernández Álvarez

Universidad de Oviedo

RESUMEN

Se reflexiona sobre el impacto que tiene la inteligencia artificial en el mundo actual, subrayando los riesgos que supone y cómo deberíamos manejarlos. Nos centramos en el grupo social de las personas ancianas, analizando los principales problemas que enfrentan y tratamos de ofrecer algunas soluciones. Por último, nos centramos en la problemática de la soledad no deseada en las personas de la tercera edad y analizamos si la inteligencia artificial podría ser una ayuda o un problema añadido. Concluimos que la problemática debe ser abordada analizando sus causas y no tratando de solventar sus consecuencias con tecnologías.

PALABRAS CLAVE: inteligencia artificial, ancianos, soledad, riesgos, tecnologías

1. Introducción

El impacto que la inteligencia artificial tiene sobre la población está cambiando la forma en que nos relacionamos con el mundo y con otros sujetos. Esta cuestión, unida al hecho de que el desarrollo y evolución de la inteligencia artificial es inminente e imparable, hace urgente un debate que nos guíe en nuestras acciones futuras. Parece innecesario para la población común cuestionarse si debemos o no detener el desarrollo de estas tecnologías, ya que esto parece estar fuera de nuestro alcance. Los avances tecnológicos están tan intrincados en nuestras vidas que resulta inimaginable tratar de dar un paso hacia atrás. A pesar de esta descorazonadora realidad, no se debe dejar de lado que el principio de precaución debe ser una de nuestras guías. Las legislaciones que se lleven a cabo y los desarrollos por parte de los conglomerados tecnológicos no deben dejar de lado el hecho que, si no hay unas medidas de protección que nos aseguren que los beneficios del desarrollo de una tecnología van a superar con creces los riesgos, y que estos riesgos, además, respetan la idea de justicia común y los derechos humanos, esta tecnología debe detenerse. Debiendo ser este principio nuestra guía a la hora de desarrollar nuevas tecnologías, las personas de a pie no tienen prácticamente ningún poder de decisión sobre estas cuestiones, incluso para las personas que están inmersas en los proyectos tecnológicos se les dificultaría seguir este principio al pie de la letra. Por esta razón, parece más sensato centrar nuestros esfuerzos en cómo gestionar el impacto que está teniendo y va a tener en nuestras vidas tecnologías tan complejas como la inteligencia artificial.

Para tratar de gestionar el impacto que tienen se debe entender que todas las tecnologías suponen ciertos riesgos. Es inocente pensar que se pueden realizar tecnologías sin riesgo alguno. Desde riesgos que se trasladan temporalmente, riesgos que surgen de la nada, e incluso, riesgos que surgen al remediar otros. El mundo actual no deja espacio a la inacción, es necesario ser conscientes de estos riesgos, de sus causas y consecuencias, y tratar de minimizarlos. Sabiendo que vivimos en sociedades del riesgo (Beck, 2002), y que un mundo sin ellos es una fantasía, ser conscientes de ellos, y actuar en consecuencia es la mejor solución para abordarlos.

A pesar de ser conscientes de esto, nos encontramos en un momento de alarma máxima, donde los avances científicos, que en este caso nos referimos a los avances en inteligencia artificial, son más rápidos que los avances a nivel moral y de legislación. Debido a ello se nos dificulta gestionar estas tecnologías porque no conocemos el impacto que tendrán en un futuro próximo, e incluso cuando conseguimos prever este impacto o creemos conocerlo, ya han nacido nuevas aplicaciones y tecnologías que nos posicionan otra vez en el punto de salida. Nos vemos casi impotentes ante un mundo que parece moverse más rápido que nosotros. El miedo que nos transmiten estas nuevas tecnologías se traduce en desconcierto y desconfianza. Pero no quiere decir que no podamos hacer nada al respecto. Gobiernos, como el de España, intentan solucionar estos problemas y evitar estos miedos mediante legislaciones, más o menos completas. Aun así, estos esfuerzos en legislar parecen no ser suficientes, parecen quedarse atrás continuamente, dejando vacíos legales que causan nuevas problemáticas. Es necesario que las leyes sean garantistas para evitar mayores problemas a posteriori, y que así se eviten al máximo dichos vacíos legales. Esto se podría conseguir mediante la deliberación, no solo de personas expertas, sino también ciudadana. Aunque la legislación sea una parte importante a la hora de gestionar estas tecnologías, no podemos caer en legalismos. La excesiva burocracia y las barreras legales que pueden traer las legislaciones demasiado rígidas pueden incluso frenar los avances, y las soluciones para ciertos riesgos que puedan surgir.

La inteligencia artificial es una tecnología muy nueva y prometedora, si la usamos bien, tenemos en cuenta sus riesgos y actuamos en consecuencia podríamos evitar problemas mayores a posteriori.

2. Inteligencia artificial, la tercera edad y la soledad

La inteligencia artificial, al ser una tecnología nueva, la cual mejora cada día, da lugar a diferentes problemáticas que necesitan de la supervisión de expertos. Si hablásemos de todos los riesgos y problemas que pueden surgir, la lista sería infinita. Centrar la problemática y tratar de ser conscientes de los problemas uno a uno es una mejor estrategia a seguir.

Dentro de los diferentes grupos sociales existen algunos que suelen verse más afectados por las problemáticas sociales. Las personas de bajos recursos, los colectivos migrantes, las personas poco alfabetizadas, y, por supuesto, los ancianos, son algunos de los grupos que suelen verse más afectados por los problemas que pueden dar tecnologías como la inteligencia artificial. Estos grupos suelen sufrir problemas tales como la desinformación, los bulos, la imposibilidad de manejarse con estas tecnologías, incluso problemas más graves como el impacto medioambiental que puede surgir fruto del uso de estas tecnologías. Las problemáticas son diversas, y, muchas veces, sus herramientas para manejar estas situaciones son escasas. Las personas ancianas, además, suelen reunir varias de las características que las hacen más vulnerables. Muchas de ellas no han podido tener una buena escolarización, lo que les hace más vulnerables a la desinformación. Esto se une a que contra mayor sea la edad, mayor es el deterioro físico y mental, lo que les hace, muchas veces, dependientes de otras personas e incapaces de comprender muchos de los avances tecnológicos. Todas estas cuestiones, sumado a que muchas de estas personas viven en el ámbito rural, y, en algunos casos, lejos de sus familias, hacen que sufran de una soledad no deseada. Todo esto crea un cúmulo perfecto para un grupo vulnerable al que la tecnología podría ayudar, o, hacer sufrir más.

Nos encontramos en un mundo cada vez más digitalizado: gestiones que solo se pueden hacer por internet, citas que no son presenciales, QR como carta en los restaurantes o noticias que solo encontramos en la red. Son muchas las dificultades que se le pueden presentar a la gente mayor. Aunque la tecnología está pensada para facilitarnos la vida y permitirnos centrarnos en lo verdaderamente importante, a veces termina generándonos más complicaciones de las que resuelve: páginas oficiales mal hechas, gestiones imposibles o números de teléfono que no te coge nadie. Todas estas problemáticas podrían no serlo tanto si las personas tuvieran una comunidad grande y estable que se ayudase la una a la otra. Sin embargo, el principal problema de las personas más mayores no es el avance tecnológico al que no han sabido adaptarse, sino la soledad que enfrentan en todo el caos actual.

Viendo cómo la inteligencia artificial parece estar ayudando a mejorar nuestras vidas, ocupándose de tareas excesivamente repetitivas para que los humanos podamos centrarnos en las tareas en las que podamos desarrollar nuestra creatividad, nos preguntamos si la inteligencia artificial podría o no ayudar a solventar las problemáticas que enfrentan las personas de edad más avanzada, sobre todo, el problema de la soledad.

3. IA: ¿más problemas o una solución?

La inteligencia artificial, de manera muy sencilla y resumida, es la disciplina que trata de crear máquinas que imiten la inteligencia humana para alcanzar ciertos objetivos. Dentro de esta disciplina podemos diferenciar diferentes subcategorías las cuales tienen unas características determinadas. Eligiendo cuidadosamente cuál usar y cómo hacerlo podríamos solucionar muchos de los problemas que se les presentan a las personas mayores en nuestra sociedad. No solo podríamos solucionar los problemas relacionados con la propia tecnología sino también aquellos que son del ámbito sociosanitario, como es la soledad.

Vemos como, por ejemplo, los sistemas expertos se pueden aplicar en dispositivos que detecten cuando una persona ha sufrido una caída o está mostrando movimientos inusuales. Teniendo en cuenta que el 30% de las personas en edad muy avanzada sufren al menos una caída al año, y que esto supone el 75% de las caídas anuales, aquellas personas que se encuentran solas y no tengan a nadie que pueda ayudarlas podrían recibir esa ayuda gracias a la inteligencia artificial. Por supuesto, se podría usar la inteligencia artificial para monitoreos de salud o, incluso, para gestión de medicamentos, ayudando así a esas personas que no son capaces de gestionar su propia salud, pero que tampoco tienen a nadie que les pueda ayudar, teniendo en cuenta, además, que la falta de memoria es uno de los grandes hándicaps de las personas más mayores. Dada su avanzada edad, estas personas tienen dificultades para encontrar en internet información que pueden necesitar, como una dirección o qué línea de autobús tomar, ante este dilema los *chatbots*, los cuales usan *machine learning*, pueden resultar muy útiles. Muchas de las dificultades que surgen cuando una persona se encuentra sola pueden ser solucionadas con inteligencia artificial. Incluso mediante esta se pueden detectar

signos tempranos de soledad, para que así los sistemas de salud puedan ayudar, proporcionando apoyo terapéutico en caso de ser necesario. Esta clase de medidas las ha tomado el Ayuntamiento de Madrid, en asociación con Madrid Salud y la Comunidad de Madrid, los cuales intentaron abordar la soledad no deseada en el 9% de los mayores de 65 años. Esta medida se llevó a cabo mediante el análisis del lenguaje natural que se obtenían en las llamadas telefónicas.

A pesar de todas estas estrategias no debemos perder el foco en las causas de la soledad. Aunque la inteligencia artificial pudiera paliar ciertos síntomas del fenómeno, mientras no atendamos a las causas, la problemática no desaparecerá nunca. El centrarnos en las consecuencias y no en las causas puede causar riesgos contraproducentes o temporales, con los cuales tendremos que lidiar. Es decir, al tratar de lidiar con la soledad no deseada, creamos otros problemas que pueden ser incluso más difíciles de gestionar.

Por esta razón hemos de analizar porqué tenemos que tratar síntomas de soledad, y porqué es que estas personas están solas. Existen muchas razones para este fenómeno, una, por ejemplo, es que los hijos de estas personas mayores que viven en zonas rurales ya no pueden quedarse en sus pueblos, ya que no hay trabajo. Esto les obliga a mudarse a las ciudades, lo que hace que estén lejos de sus familias, las cuales tenían a su propia familia como único núcleo de socialización. Dado que las zonas rurales empiezan a despoblarse, cada vez su círculo social se ha ido estrechando. Esto se une a que las personas que viven en contextos rurales suelen tener un nivel de alfabetización bajo, lo que les dificulta, todavía más, salir de su zona de confort para poder hacer amigos. Sin embargo, no son solo las personas ancianas de contextos rurales quienes sufren la soledad, sino que en las pequeñas y grandes ciudades también encontramos este problema. Dado el crecimiento exponencial que vemos en las grandes ciudades, las relaciones sociales se han vuelto muy complejas, las personas tienen que trabajar muchas horas fuera de casa, y, cuando vuelven, no les queda energía para dedicársela a nada ni a nadie, ni siquiera a sus personas mayores. Al complejizarse tanto las relaciones sociales estas personas ya no son capaces de relacionarse como antes. A esto le tenemos que sumar que debido a su edad muchas veces no tienen la capacidad física o mental para sobrellevar el hacer nuevas amistades o siquiera relacionarse de

manera natural. Todas estas cuestiones, y muchas otras, tiene que ver con la concepción de tiempo cortoplacista en la que vivimos en el tiempo actual. Nos cuesta mucho establecer relaciones con otras personas ya que la sociedad capitalista y de riesgo en la que nos encontramos nos la dificulta. Esto, sumado a la dependencia extrema a las tecnologías, hace que cada vez las personas tiendan a aislarse. Sin tiempo para conocer gente, sin sitios donde reunirse, y sin ni siquiera energías para poder socializar, hace que las personas pierdan sus lazos.

4. Conclusión

Las razones de la soledad, no solo de las personas mayores, sino de la población en general es un problema complejo, el cual requiere de un mayor análisis. Pero, sin duda, tratar de atajar el problema centrándose en sus consecuencias, muchas veces triviales, no es un enfoque adecuado. Debemos cuestionarnos porqué deberíamos usar tecnologías como la inteligencia artificial para tratar problemáticas sociales que van más allá de problemas tecnológicos. Quizás, nos hemos centrado demasiado en los avances tecnológicos, que hemos perdido el rumbo de lo que es realmente importante. Creemos que herramientas y tecnologías pueden solucionar todos nuestros problemas, cuando, puede ser, que estos sean la causa de ellos.

BIBLIOGRAFÍA

- Beck, U. (2002). *La sociedad del riesgo global*. Siglo veintiuno de España editores.
- Graham, J., & Wiener, J. (1997). *Risk versus Risk: Tradeoffs in Protecting Health and the Environment*. Harvard University Press.
- Nagusi Intelligence Center. (s.f.) *Inteligencia Artificial para las personas mayores: aplicaciones y oportunidades de negocio*. Diputación Foral de Bizkaia.
- Sennet, R. (2006). *La corrosión del carácter*. Editorial Anagrama.



Figura1: Póster realizado en el Aula de Extensión Universitaria “Ingeniería y Filosofía 7.0. IAcracia: hacia la sumisión algorítmica”, en la Universidad de Oviedo

III. LECTURAS ÉTICAS DE LA IA

IA, narrativa y fascismo

Pablo Revuelta Sanz

Grupo IF

Universidad de Oviedo

RESUMEN

La potencia disruptiva de la IA está trastocando todos los aspectos de la vida, y empezando a modificar la forma en la que entendemos el mundo y nos relacionamos cognitivamente con él. Ello se une a su capacidad de suplantar cada vez más esferas de lo que considerábamos hasta hace poco exclusivamente humanas. Por último, sistemas tan complejos necesitan inversiones inmensas de recursos y tecnología, en manos de una oligarquía con un plan de ruta bien trazado y definido que no pasa precisamente por el fortalecimiento de la participación democrática, el reconocimiento de la diversidad y la vulnerabilidad y la sostenibilidad ambiental.

PALABRAS CLAVE: fascismo, inteligencia artificial, tecnología, narrativa, movimiento neorreaccionario.

1. La IA como sujeto

La IA ha mostrado un potencial único frente a tecnologías anteriores. El aprendizaje profundo, con operaciones como convoluciones, supera a sistemas

expertos, pero también a profesionales humanos en procesamiento de imágenes. Arquitecturas generativas han revolucionado la creación de contenido multimedia, impactando en campos como el arte o el lenguaje natural.

Sin embargo, presenta varios problemas, como la opacidad, que ocurre a tres niveles:

- Arquitectura protegida por patentes.
- Pesos de la red secretos.
- Imposibilidad de explicar su funcionamiento práctico.

Esto convierte la IA en una "caja negra", donde solo confiamos en su precisión sin entender cómo llega a sus respuestas. Nos exige un acto de fe.

La potencia de las GPUs y su omnipresencia en sociedades postindustriales dotan a la IA de un aura "divinizante". Como señala M. Castells (La Vanguardia, 2023), "lo invade todo, lo sabe todo, lo controla todo". Es inquietante cómo descripciones bíblicas ("Los caminos del Señor son inescrutables", Isaías 55:8) reflejan la opacidad de la IA, mientras sus promesas recuerdan a lo divino ("Nos promete soluciones a nuestros problemas", E. Schmit, (La Vanguardia, 2024)).

Vivimos un "fetichismo tecnológico" (en paráfrasis marxiana) donde la IA parece adquirir rasgos de ser autónomo (llegando a "sufrir" ansiedad, Ben-Zion, 2025), invirtiendo las relaciones de sujeto y objeto: humanos actúan mecánicamente ante mandatos de IA (accidentes por Google Maps o cambios políticos por fake news).

2. El fascismo

El fascismo del siglo XX fue un movimiento diverso pero con ciertos rasgos comunes: reactivo, nacionalista, estatista, militarista y generador de alteridad.

Entre 1920-1940 surgieron por toda Europa variantes como el PNF (Italia), el NSDAP (Alemania), los ustachas (Croacia), Frentes Nacionales (francés, belga, inglés...) o la Falange Española. Estas expresiones mostraron, sin embargo, profundas diferencias: mientras la Falange era católica y anticapitalista (Primo de

Rivera, 1942), el PNF y NSDAP apenas toleraban el catolicismo y se apoyaron en los grandes grupos empresariales. Presentaban además contradicciones internas respecto a la burguesía o la etnicidad (Scurati, 2020-2021), lo que impide abordar el fascismo como movimiento unificado.

No obstante, podemos presentarlo como sistema de sustituciones a tres niveles:

- De la revolución: el fascismo surgió como fuerza de choque contra la amenaza socialista, defendiendo los privilegios de las élites. Como señala Ledesma (2017), "Su [del fascismo] estrategia de lucha contra una fuerza social -el marxismo, [...]-, [es] vencéndola revolucionariamente". Buchrucker (2008) destaca, a su vez, que "reserva una considerable autonomía a las élites del poder económico". La atribución a Brecht ("nada más parecido a un fascista que un burgués aterrado") sintetiza esta dinámica.
- De la verdad: Fernández y Alba (1986) sostienen que "hablar del poder es hablar contra el poder". El fascismo manipuló la verdad desde sus inicios, como sugiere la idea atribuida a Göbbels ("una mentira repetida suficientemente acaba por convertirse en verdad"), idea presente en otros textos nazis (Hitler, 2003). Scurati (2024) subraya cómo "veinte años de lento y obstinado aprendizaje de la mentira no pueden liquidarse en una noche".
- De las personas: el fascismo histórico rechazó el concepto de persona del liberalismo (seres racionales y libres que buscan el bien individual) y del socialismo (seres dignos que se organizan para una liberación colectiva). Ante ello, proclamó que "Todo en el Estado, nada contra el Estado" (Mussolini, 2022). Anders ya nos alertó sobre el riesgo de un "mundo sin nosotros" (2011).

3. Fascismo y tecnología

El fascismo fue un movimiento (contra)revolucionario que buscó transformar la antropología mediante, entre otras herramientas, la tecnología. Mientras movimientos emancipadores veían en ella una herramienta de liberación

(política o económica), el fascismo la utilizó como instrumento para su proyecto político. Esta fascinación tecnológica quedó plasmada en el Manifiesto Futurista (1909):

- El Tiempo y el Espacio han muerto [...] puesto que hemos creado la celeridad omnipresente. (VIII).
- Queremos combatir el moralismo, el feminismo [...] (X).
- Los nuevos adelantos nos encontrarán [...] en acción de calentarnos las manos en la fogata miserable que nutrirán nuestros libros...
- ¡Pero estamos bien cerciorados de lo que nuestra bella y falsa inteligencia nos afirma!

Así, en el esquema tecno-optimista que inspira este manifiesto y el fascismo que lo contextualiza y utiliza, podemos resaltar cuatro ideas que manifiestan esta relación:

- Culto a la tecnología: rasgo compartido con anarquismo, socialismo y capitalismo.
- Rechazo a lo "natural" y "corpóreo": la naturaleza se asocia a atraso, el cuerpo sin tecnología se consideraba débil e impotente (Barjola, 2023). Esta visión coincide parcialmente con la religión y el capitalismo (aunque éste explota comercialmente estos valores, como muestran los mercados bio, gastronómicos o pornográficos).
- Aceleracionismo tecnológico: el nazismo (más que el fascismo italiano) impulsó innovaciones tecnológicas en paralelo al desarrollismo del "socialismo real" (carrera espacial) y el capitalismo.
- Control territorial y social: respondiendo a la tercera sustitución propuesta, este uso viene compartido también por el "socialismo real" (últimamente con grandes "avances" en China) y el capitalismo.

4. Tecnología y narrativa

Las "viejas" tecnologías (fuego, arado, navegación) ampliaron las capacidades físicas humanas, pero no interactuaban, al menos directamente, con procesos cognitivos. Las "nuevas" tecnologías, orientadas a labores cognitivas,

están provocando transformaciones sin precedentes en nuestra relación con el mundo. Como se señaló en el II Congreso IF, estamos desarrollando una "piel tecnológica" que filtra y modula nuestro rango de acción.

Así, estas tecnologías nos están "descapitalizando cognitivamente":

- Las calculadoras eliminaron el cálculo mental.
- Los buscadores redujeron nuestra capacidad de búsqueda estructurada.
- La traducción automatizada afectó el aprendizaje de idiomas (y con ello, la comprensión de la propia lengua, como advirtió Borges).
- Las redes sociales mediaron nuestras relaciones afectivas y sexuales.
- Los GPS erosionaron la orientación geográfica...

Sólo persisten algunas meta-habilidades (análisis, toma de decisiones) que, al no tener aún equivalentes tecnológicos maduros, resisten temporalmente este embate. Por poco tiempo.

Antes de abordar esta cuestión más transversal, revisemos brevemente la evolución de la información en los últimos años (Figura 1).

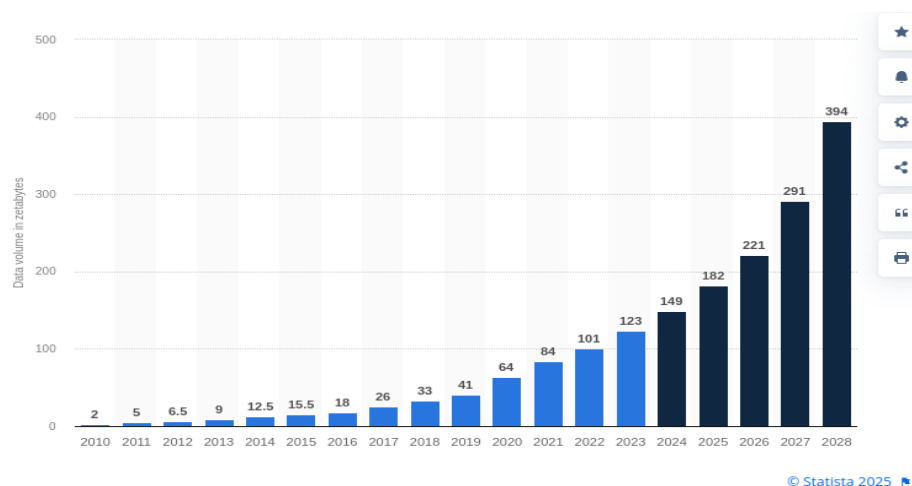


Figura 1: Evolución de la información

La gráfica muestra la evolución de datos generados desde el inicio de la "Era Zettabyte". El eje vertical representa zettabytes (10^{21} bytes). Según Vance (2013), desde los orígenes humanos hasta 2003 se crearon 0,5 zettabytes, cantidad que en 2013 se generaba cada dos días, mostrando una progresión geométrica.

Esta explosión de datos crea problemas interpretativos: vivimos una ilusión de omnisciencia en la "celeridad omnipresente", con capacidad de generar datos a demanda. Sin embargo, como advierte D. Innerarity: "El verdadero desafío de nuestro tiempo: interpretar para obtener experiencias a partir de los datos y sentido a partir de los discursos" (El País, 2010).

Nos hallamos en una nueva "habitación china" de Searle, donde torrentes de datos entran y salen sin que podamos aprovecharlos cognitivamente. Para entender el punto en el que estamos, recuperemos la famosa ilustración de Somerville (Figura 2).

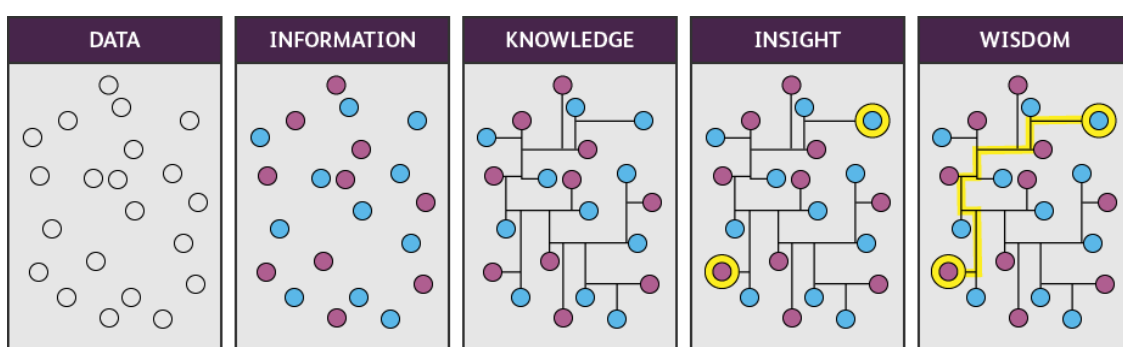


Figura 2: Ilustración de Somerville

Los datos están disponibles o son fácilmente generables. Organizados, pueden almacenarse en bases de datos, mientras los buscadores ayudan a discriminarlos, obteniendo información (con sus sesgos). Recientemente, herramientas de IA generativa permiten conectar datos y producir pseudo-conocimiento que imita el humano.

Sin embargo, la "perspicacia" y la sabiduría –identificar lo relevante y vincularlo con elementos aparentemente inconexos– requieren lo que la IA aún no ofrece: subjetividad. Solo nuestro posicionamiento en el mundo y capacidad de interpretar desde ese "ahí" permiten construir narrativas que den sentido a la información, conectándola con nuestra experiencia subjetiva.

Siguiendo a Gallego (2024), que cita a su vez a B. Ch. Han, podemos afirmar que la narrativa nos permite:

1. Cribar datos.
2. Contextualizar datos.

3. Conectar datos divergentes.
4. Generar sentido y lugar en el mundo.
5. Generar propuestas transformadoras.

La IA maneja bien lo primero (procesamiento básico), pero contextualizar le resulta problemático: a veces "alucina" creando falsos contextos, aunque estas limitaciones se resolverán pronto. Conectar datos divergentes sigue siendo humano, pues requiere creatividad, no solo imitación.

Las dos últimas capacidades narrativas —dotar de sentido subjetivo y generar propuestas transformadoras— escapan a las máquinas: carecen de un "lugar en el mundo" que reivindicar y, por tanto, de subjetividad para desear cambios reales.

5. Conectando conceptos

M. Twain afirmó que la historia no se repite, pero sí rima, y veremos cómo.

La revolución de internet prometía sacarnos de la caverna platónica, pero fuera solo había un enjambre de anuncios en neón para el que nuestras herramientas cognitivas quedaron obsoletas. Las crisis impulsan búsquedas de respuestas simples. S. Vosoughi (2018) demuestra que en Twitter/X las falsedades tienen "un 70% más posibilidades de ser retuiteadas que la verdad", difundidas principalmente por humanos. "la IA nos ahorran buscar, pensar, solventar problemas" (Monedero, 2024), de forma que el mayor riesgo de la IA es "la estupidez humana" (El País, 2024). Además, las redes sociales aíslan discursivamente, creando burbujas ideológicas que impiden contrastar ideas.

Muestra del efecto de todo ello es la tendencia detectada en los informes PISA (OCDE, 2022). Véase figura 3.

Desde casi el comienzo de la Era Zettabyte, y especialmente desde el año 2015 (la mayoría de las redes sociales nacieron en la primera década del nuevo siglo -salvo TikTok-, pero su alcance masivo, pongamos en los 500 millones de usuarios, se alcanzan entre 2012 para Twitter (Europa Press, 2012) y 2018 en el

caso de TikTok (Chen, 2018)), nuestra capacidad lectora está en claro retroceso en la OCDE.

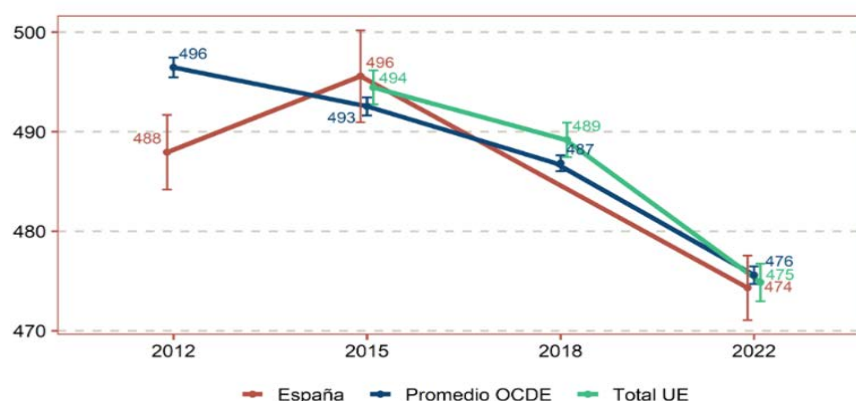


Figura 3: Evolución de los rendimientos medios estimados en lectura entre 2012 y 2022 para España, el promedio de la OCDE y el total de la UE

La gran amenaza es la unión de una capacidad cada vez más mermada de análisis crítico con una inundación de pseudo-narrativas generadas artificialmente que, si bien resultan cada vez más sofisticadas y realistas, tienen grandes carencias:

- No ofrecen un contexto histórico fiable.
- Están amenazadas de “alucinaciones”.
- Tienen prejuicios.
- No tienen “crítica”.
- No ofrecen vinculación con la experiencia, vida e historia personal y colectiva.

6. La tormenta perfecta: hacia un fascismo tecnológico

En este estado de cosas, irrumpe la IA invadiendo prácticamente todas las esferas de la vida de gran parte del mundo. Esta nueva herramienta presenta unas capacidades revolucionarias como pocas anteriores y amenaza con dislocar todas las facetas de la vida humana en un momento histórico de especial vulnerabilidad, con una crisis ecológica extremadamente grave y otra económica desde 2008. Los ecos de “La Decadencia de Occidente” vuelven a recorrer el mundo y la enorme reacción generada a los avances del feminismo han generado un caldo de cultivo de una tercera crisis que se solapa con las dos anteriores, de índole cultural e identitaria, rimando nuevamente con el siglo pasado.

Discursos contrarevolucionados alentados por miedos a perder privilegios (de género, de tren de vida insostenible, de crecimiento económico infinito...) empiezan a proliferar, disponiendo ahora de nuevas herramientas de generación y gestión de la información, y una capacidad de penetración social nunca antes vistas.

La generación de pseudo-información ha alcanzado unos niveles difíciles de distinguir de la realidad y esto, como el incendio del Reichstag en 1933, es gasolina para este nuevo fuego. T. Snyder (2022), especialista en regímenes autoritarios en Yale, afirma que “la postverdad es el prefascismo”.

Quienes acumulan un ingente poder se están agrupando en torno a la autodenominada “Ilustración Oscura”, defendiendo un sistema “neocameralista”, en el que un CEO-Rey platónico, dirige el Estado como una *startup*, maximizando sus beneficios y las inversiones privadas en él.

Este movimiento también autodenominado neorreaccionismo o NRx, bebe de un tecnolibertarianismo en el que la propiedad privada y la noción liberal de libertad empieza a ser incompatible con la democracia, como otro de sus impulsores, P. Thiel, ha señalado (Thiel, 2009). Esta noción es compartida por varios “coachs” de Silicon Valley, entre ellos uno de los más influyentes: P. Campbell (Schmidt, 2019). Este grupo está formado mayoritariamente por hombres blancos y extremadamente ricos (Thiel es co-fundador de PayPal y accionista de Facebook, también el mayor contratista del Pentágono y ha financiado experimentos anarcocapitalistas como los *Seasteading* o apoyado con 15 M\$ la candidatura a senador de J.D. Vance (Allyn, 2024)).

En este contexto, la población, como se ha mostrado, va perdiendo capacidad analítica y comprensiva de la información que recibe, cada vez más dirigida a reafirmar sus propias ideas y desplazarlas hacia la derecha en el espectro político (CTXT, 2024).

Este nuevo discurso basado en las “verdades alternativas” y las posiciones autorreferenciales presentan, por su oposición a la Ilustración, un desprecio radical a las argumentaciones, que no puede menos que recordarnos al Manifiesto

Futurista. Las “rimas” con los años 20 del pasado siglo no terminan aquí: otro CEO de una empresa de capital riesgo de Silicon Valley, Mark Andreessen escribió en 2023 el Manifiesto tecno-optimista (Andreessen, 2023).

Todo el bagaje tecnológico se empieza a movilizar para promover un nuevo estado de cosas en el que la libertad individual del 0,1% más rico no se vea constreñida por la sostenibilidad ambiental, las políticas de inclusión o, simplemente, impuestos (del 25%) como los que Biden impuso a quienes tuvieran más de 100M\$ (Gilchrist, 2024).

Otros elementos del siglo pasado rescatados por este movimiento son el natalismo (Musk, p.e. tiene 17 hijos), la “eugenesia” (todo sus hijos son varones, salvo una que es mujer trans), el racismo (sobre un neodarwinismo social), y el aceleracionismo [acelerar los eventos sociales para forzar un temprano colapso del capitalismo y alumbrar la nueva era postcapitalista de feudo-estados tecnológicos (Ríos, 2024)].

Quizás las rimas acaben acá, porque hay elementos no traducibles en este lapso de 100 años: el eje vertebrador de estos nuevos movimientos ya no son los estados, sino las redes sociales en las que se articulan las identidades y también las acciones sobre el mundo analógico. La libertad sólo está reservada, recordemos, para las élites económicas, mientras que el resto se convierte en carne de cañón perfectamente sustituible para una campaña de acoso, un nuevo referéndum, un asalto al Capitolio o un trending topic.

Por último, señalar que el espíritu de sacrificio que pregonaba el primer fascismo ha transmutado en un culto al consumo y al ocio pasivo.

7. Antídotos

Ante cualquier situación de emergencia, y con más razón aún si cabe cuando la emergencia se basa en el aceleracionismo, deberíamos reflexionar hacia donde nos están llevando las élites económicas de unos y otros países. Como cualquier reflexión, requiere grandes dosis de calma, de paciencia y de análisis de los datos disponibles, sobre los que generar una narrativa que no sólo explique el estado actual de las cosas, sino las alternativas que aún tenemos por delante.

Esta tranquilidad, no obstante, choca con la urgencia de algunos temas, como el incipiente colapso climático pero no sólo, y eso ha llevado a algunas personas a abogar por lo que llaman la “internacional de la supervivencia” (Alba, 2025), creando alianzas que en otras ocasiones habrían sido *contra natura*, pero que ahora se justifican por el extremo peligro al que nos enfrentamos.

Necesitamos también apoyar políticas que articulen visiones y derechos post-individualistas e interpersonales, que confronten con estas visiones deshumanizantes y turbocapitalistas.

Se hace igualmente perentorio la defensa de verdad tarskiana, como base común sobre la que evaluar los discursos sobre el mundo y sus elementos, negándonos a aceptar opiniones basadas en pensamiento mágico.

De forma complementaria, recuperar el valor de la subjetividad, con el sufrimiento como un elemento irreductiblemente humano (frente al mundo artificial) que merece respeto, compasión y ternura. En palabras de Th. Adorno, “hacer elocuente el sufrimiento como condición de toda verdad”.

Revalorizar los cuidados, la cooperación y la igualdad como principios antagónicos a los que promulga la Ilustración Oscura.

En definitiva, promover y apoyar los dos grandes movimientos sociales que actualmente articulan estos principios y que, no por casualidad, son los dos grandes caballos de batalla contra los que lucha el movimiento NRx: el ecologismo y el transfeminismo.

BIBLIOGRAFÍA

Alba Rico, S. (2025). *Trumpismo: el poder de la utopía*. Público. <https://www.publico.es/opinion/columnas/trumpismo-poder-utopia.html> accedido el 04-05-25.

- Allyn, B. (2024). *Five things to know about J.D. Vance's ties to tech billionaires*. NPR. <https://www.npr.org/2024/07/17/g-s1-11654/five-things-to-know-about-jd-vances-connections-to-tech-billionaires> accedido el 04-05-25.
- Andreessen, M. (2023). *The Techno-Optimist Manifesto*. <https://a16z.com/the-techno-optimist-manifesto/> accedido el 04-05-25.
- Anders, G. (2011). *La obsolescencia del hombre, vol II*. Pre-textos.
- Barjola, N. (2023). *Microfísica sexista del poder. El caso de Alcàsser y la construcción del terror sexual*. Virus Editorial.
- Ben-Zion, Z., Witte, K., Jagadish, A. K., Duek, O., Harpaz-Rotem, I., Khorsandian, M.-C., Burrer, A., Seifritz, E., Homan, P., Schulz, E., & Spiller, T. R. (2024). *Assessing and alleviating state anxiety in large language models*. npj Digit. Med. 8, 132 (2025). <https://doi.org/10.1038/s41746-025-01512-6>
- Buchrucker, C. (2008). *El Fascismo en el Siglo XX. Una historia comparada*. Emecé Editores.
- Chen, Q. (2018). *The biggest trend in Chinese social media is dying, and another has already taken its place*. CNBC. <https://www.cnbc.com/2018/09/19/short-video-apps-like-douyin-tiktok-are-dominating-chinese-screens.html> accedido el 04-05-25.
- CTXT/Observatorio Social 'la Caixa' (2024). *La audiencia de derechas otorga mayor veracidad a las noticias falsas*. CTXT. <https://ctxt.es/es/20241201/Firmas/48042/Observatorio-Social-la-Caixa-bulos-fake-news-ctxt-noticias-hoy.htm> accedido el 04-05-25.
- Gilchrist, K. (2024). *Biden's 'billionaire tax' takes aim at the super-rich — but can a wealth tax work in reality?*. <https://www.cnbc.com/2024/03/15/bidens-billionaire-tax-hits-the-super-rich-can-a-wealth-tax-work.html> accedido el 04-05-25.
- El País. (2010). *La sociedad de los intérpretes, entrevista a Daniel Innenarity*. https://elpais.com/diario/2010/11/16/opinion/1289862004_850215.html accedido el 04-05-25.

- El País. (2024). *Entrevista a Pilar Manchón*. <https://elpais.com/us/2024-09-24/pilar-manchon-de-google-con-la-ia-el-mayor-riesgo-que-tenemos-es-la-estupidez-humana.html> accedido el 04-05-25.
- Europa Press PortalTIC (2012), *Twitter alcanza los 500 millones de cuentas registradas*. <https://www.europapress.es/portaltic/internet/noticia-twitter-alcanza-500-millones-cuentas-registradas-20120801154438.html> accedido el 04-05-25.
- Fernández Liria, C., & Alba Rico, S. (1986). *Dejar de pensar*. Akal.
- Gallego Canteli, D. (2024). *Lágrimas de una gitana*. Ediciones Trabe S.L.
- Habermas, J. (2018). *Verdad y justificación*. Trotta.
- Hitler, A. (2003). *Mi Lucha (1925)*. El movimiento nacionalsocialista. <https://dn790004.ca.archive.org/0/items/MiLucha/milucha.pdf>, accedido el 24-04-25.
- La Vanguardia (2023). *“El mundo ha entrado en una fase sin futuro”, entrevista a Manuel Castells*. <https://www.lavanguardia.com/internacional/20230611/9030929/mundo-entrado-fase-futuro.html> accedido el 22-04-25.
- La Vanguardia (2024). *Inteligencia artificial y cambio climático*. <https://www.lavanguardia.com/vida/20241013/10017149/inteligencia-artificial-cambio-climatico.html> accedido el 22-04-25.
- Ledesma R. (2017), *¿Fascismo en España?*. Almuzara. Accesible en <https://web.archive.org/web/20091022080532/http://www.ramiroledesma.com/nrevolucion/fec1.html> accedido el 24-04-25.
- Marinetti, F. T. (1909). *El Futurismo*. Aparecido en Le Figaro, 20 de febrero de 1909. <https://arteydisegno.wordpress.com/wp-content/uploads/2010/02/manifiesto-futurista-1909.pdf> accedido el 04-05-25.

- Monedero, J. C. (2024). *Revolución y contrarrevolución en la Inteligencia Artificial*. Público. <https://www.publico.es/opinion/columnas/revolucion-contrarrevolucion-inteligencia-artificial.html> accedido el 04-05-25.
- Mussolini, B., & Gentile, G. (2022). *La doctrina del Fascismo* (1937). Lebooks Editora.
- OCDE (2022). *Informe PISA 2022*. <https://inee.educacion.es/2023/12/05/pisa-2022/> accedido el 04-05-25.
- Primo de Rivera, J.A. (1942). *Obras completas*. Diana.
- Ríos, R. H. (2024). *Filosofía en 3 minutos: Nick Land*. Perfil. <https://www.perfil.com/noticias/cultura/filosofia-en-3-minutos-nick-land.phtml> accedido el 04-05-25.
- Schmidt, E. et al. (2019). *Trillion Dollar Coach: The Leadership Playbook of Silicon Valley's Bill Campbell*. HarperCollins.
- Scurati, A. (2020). *M, el hijo del siglo*. Alfaguara.
- Scurati, A. (2021). *M, el hijo de la providencia*. Alfaguara.
- Scurati, A. (2024). *M, la hora del destino*. Alfaguara.
- Snyder T. (2022). *Sobre la tiranía*. Salamandra.
- Soroush Vosoughi et al. (2018). *The spread of true and false news online*. Science, 359,1146-1151.DOI:10.1126/science.aap9559
- Thiel, P. (2009). *The education of a Libertarian*. Cato Unbound. <https://www.cato-unbound.org/2009/04/13/peter-thiel/education-libertarian/> accedido el 04-05-25.
- Vance, J. (2013). *Big Data Analysis Overview*. <https://www.datamation.com/applications/big-data-analytics-overview.html> accedido el 04-05-25.

Hacia una IA feminista: una crítica a la veracidad de las imágenes *deepfake*

Irene Adán Sánchez-Infantes
Universidad Complutense de Madrid

RESUMEN

En este trabajo proponemos analizar algunos de los problemas epistemológicos y ontológicos que surgen de la crítica feminista interseccional a la producción de imágenes con inteligencia artificial (IA). Primero, introducimos el fenómeno de las imágenes y vídeos *deepfake*. Segundo, examinamos el problema de la veracidad y falsedad de las imágenes, tomando en consideración ciertos usos maliciosos del *deepfake* tales como la pornografía no consentida. Dichos usos nos hacen cuestionar si los criterios de veracidad u objetividad son suficientes, en la medida en que estas imágenes son empleadas como mecanismo para silenciar a las personas representadas, a la vez que refuerzan ciertos estereotipos de género. Tercero, exponemos algunos de los problemas epistemológicos y ontológicos que surgen de una crítica feminista a las nuevas tecnologías a la luz de la distinción conceptual trazada por la autora Katherine Hayles (2023). A partir de estos conceptos, presentamos el problema del cuerpo generizado y racializado en las imágenes *deepfake*. Siguiendo a Hayles, nos preguntamos por la necesidad de profundizar en una crítica ontológica de las nuevas tecnologías orientada a desarrollar estrategias y propuestas feministas acerca de la inteligencia artificial.

Más allá de la posibilidad o no de una IA feminista, damos cuenta de algunas críticas y propuestas desde los nuevos materialismos feministas en torno a este fenómeno. Finalmente, ofrecemos una serie de conclusiones y defendemos la necesidad de enfoques críticos e interdisciplinarios para abordar el fenómeno de los *deepfake*.

PALABRAS CLAVE: *deepfake*; epistemología; feminismo; ontología; inteligencia artificial generativa.

1. Introducción

En el presente, la producción y difusión de nuevas herramientas de inteligencia artificial (IA) suscita debates acerca del progreso técnico continuo donde no siempre se tienen en cuenta las estructuras políticas y sociales implicadas (Crawford, 2023). Tal es el caso de la tecnología *deepfake* o ultrafalso, la cual abre la posibilidad de crear contenido falso hiperrealista. Los *deepfakes* o *ultrafalsos* consisten en imágenes o vídeos producidos gracias a herramientas de inteligencia artificial (García-Ull, 2021). A diferencia de los *shallowfakes* —vídeos acelerados, ralentizados, con el audio alterado o fuera de contexto para mostrar a personas que dicen o hacen cosas que no han hecho (Maddocks, 2020; Matthews y Kidd, 2023)— los *deepfakes* son creados con el objetivo de generar vídeos falsos hiperrealistas (Tolosana *et al.*, 2020). Siguiendo la categorización empleada por Matthews y Kidd (2023) a partir del trabajo de Tolosana *et al.*, (2020) dentro de los posibles tipos de imágenes y vídeos de rostros creados a partir de herramientas de aprendizaje automático, es posible distinguir: primero, aquellos en los que los rostros generados no se corresponden con caras reales (*entire-face synthesis*); segundo, aquellos donde los atributos faciales son alterados (*attribute manipulation*); tercero, aquellos en los que se intercambian las expresiones faciales de una persona por otra (*expression swap*); finalmente, aquellos en los que gracias a técnicas de *deep learning*¹ es posible intercambiar el rostro de una persona por el de otra (*identity swap*). A estos últimos nos referimos en este trabajo como *deepfakes*.

Estas imágenes y vídeos son habitualmente producidos gracias a las redes generativas antagónicas o *generative adversarial networks* (GAN), cuya propuesta

teórica se encuentra en el trabajo de Goodfellow *et al.* (2014). Este proceso competitivo consiste en el entrenamiento simultáneo de dos modelos: por un lado, el modelo generativo captura la distribución de los datos; por otro lado, el modelo discriminador estima la probabilidad de que una muestra proceda de los datos de entrenamiento y no del modelo generador. El entrenamiento del generador consiste en maximizar la probabilidad de que el discriminador cometa un error (Goodfellow *et al.*, 2014). Los modelos GAN ofrecen infinitas posibilidades en el ámbito audiovisual, a la vez que presentan importantes desafíos relativos a la protección de datos, a la privacidad, a la manipulación mediática, así como a la ética de la inteligencia artificial (Matthews y Kidd, 2023). A continuación, examinamos dos problemas vinculados a los *deepfakes* desde un punto de vista feminista interseccional²: primero, la cuestión de la veracidad y opacidad; segundo, el problema del cuerpo. Este enfoque nos permite situar la IA en su contexto social, precisamente en un momento en que la producción y difusión de nuevas herramientas de inteligencia artificial suscita debates en un contexto de abundancia de contenido digital, así como de violencia facilitada por la tecnología.

2. La cuestión de la veracidad y objetividad de las imágenes *deepfake*

Siguiendo al artista conceptual Casey Reas (2019), estamos ante el inicio de un nuevo paradigma en la creación de imágenes con software de aprendizaje automático. Para este artista, la creación de imágenes ha estado estrechamente vinculada a disciplinas científicas como la óptica y la química. Quienes crean estas imágenes —en fotografía, artistas visuales, etc.— no requieren necesariamente de un exhaustivo conocimiento de las disciplinas científico-técnicas implicadas en su uso, si bien han de adquirir destrezas y experiencia para su desempeño.

La producción de estas imágenes debería ir acompañada de la posibilidad de cerciorarnos de que lo que muestran ha sucedido realmente o no. Sin embargo, no siempre es así: la creación de imágenes como objeto de conocimiento está sujeta al problema de su verificación. Anteriormente, la alteración de imágenes requería de la adquisición de conocimientos técnicos y el desarrollo de ciertas destrezas (Wagner y Blewer, 2019). Aunque la manipulación de contenido audiovisual no sea una novedad, sí lo son los procesos para su elaboración (Simó,

2023) en parte, gracias a la distribución de aplicaciones y programas que hacen uso de herramientas de inteligencia artificial para crear contenidos audiovisuales. Wagner y Blewer (2019) señalan que las implicaciones que estas imágenes puedan tener son de especial relevancia en un contexto de retórica de noticias falsas. Es por ello por lo que los *deepfake* pueden vincularse a la desinformación intencionada, a la desconfianza en los medios, a las estafas y a las extorsiones (García-Ull, 2021). No obstante, estos usos del *deepfake* no corresponden a la mayoría del material existente.

Desde un punto de vista social y de género, a pesar de que la atención pública se ha centrado en los *deepfake* de carácter político o humorístico contra varones, la mayoría de *deepfake* son de carácter pornográfico (Maddocks, 2020; Ajder *et al.*, 2019). Ya dan cuenta de ello diversos informes como *The State of deepfakes: Landscape, Threats and Impact*, el cual concluyó en 2019 que el 96% de las imágenes *deepfake* de internet son de carácter pornográfico (Ajder *et al.*, 2019). De la muestra examinada, el 100% de los vídeos *deepfake* pornográficos examinados eran protagonizados por mujeres. Al considerar estos datos, podemos afirmar que la producción de pornografía *deepfake* no es anecdótica. Los roles de género guardan un papel relevante en la industria de producción de *deepfake*, no solo en cuanto a la imagen como representación final ³/₄ al ser empleados, por ejemplo, contra mujeres críticas con la violencia sexual (Maddocks, 2020) o como forma de cosificación, extorsión y silenciamiento ³/₄ sino en cuanto a los datos y a la infraestructura tecnológica que requieren. Se trata pues de un fenómeno que requiere considerar aspectos más allá de la veracidad y la objetividad de las imágenes.

3. La relación de las imágenes *deepfake* con el cuerpo: de la crítica epistemológica a la ontológica

La violencia que caracteriza a los *deepfake* no consentidos, si bien no es estrictamente física, alberga una relación estrecha con la corporalidad y materialidad que no se reduce al cuerpo individual, sino que también afecta a una materialidad del género, de los afectos, de la racialización, etc. Es decir, el uso de *deepfake* no se reduce a los efectos sobre el individuo, puesto que se trata de una

violencia marcada por ejes de estratificación social en un contexto capitalista, patriarcal y neocolonial. La reificación del género, la raza o la clase se produce a lo largo del proceso y en todas sus etapas y se traduce en un impacto social. Este problema se enmarca a su vez en las políticas de la clasificación de la IA (Crawford y Paglen, 2021).

La relación entre las políticas de clasificación y el aprendizaje automático ya ha sido señalada en la medida en que incorpora ciertas asunciones sobre los rostros generizados y racializados. Tal es el caso de las tecnologías de reconocimiento facial, clasificación o detección faciales —distintas de la tecnología que permite la elaboración de *deepfakes*— donde igualmente existen una serie de preconcepciones que pueden ser abordadas desde una perspectiva interseccional. Por ejemplo, con respecto al género, encontramos taxonomías binarias y/o transexcluyentes, así como visiones estereotipadas en torno a la raza³. En este sentido, existen varios ejemplos de algoritmos empleados con datos etiquetados que han derivado en distintas formas de discriminación algorítmica (Bolukbasi *et al.*, 2016), especialmente en lo que respecta a los rostros de mujeres racializadas.

A este respecto, Kate Crawford considera que existe una preconcepción por la cual se considera que las herramientas de aprendizaje automático definen el mundo en términos fijados previamente: “las clasificaciones son tecnologías que producen y limitan los tipos de conocimiento y están integradas en la lógica de la IA” (Crawford, 2023, p. 224). Eleanor Drage y Federica Frabetti sugieren que la inteligencia artificial tiene un efecto performativo en la medida en que produce la realidad que se supone que observa: “identificar y eliminar los prejuicios de género sigue siendo una tarea infructuosa que asume que las consecuencias perjudiciales de la IA son aislables dentro de un sistema y no el producto de la relación de la tecnología con normas sociales, instituciones y economías” (Drage y Frabetti, 2023, p. 282; traducción propia).

Al cuestionar las injusticias derivadas de las herramientas de aprendizaje automático, son frecuentes las alusiones a los sesgos y la opacidad. Frente a esta opacidad, las posturas institucionales sobre la IA proponen hacer públicos y accesibles los algoritmos. No obstante, este lenguaje no cuestiona el contexto en el

que se producen estas tecnologías. Desde una perspectiva crítica contra la opresión facilitada por las nuevas tecnologías, Justin Joque sostiene en *Matemáticas revolucionarias* que, si bien la transparencia puede ayudar a prevenir los abusos del conocimiento algorítmico, por sí sola no pondrá freno a las injusticias (Joque, 2024). Asimismo, Jude Browne (2023) sugiere que el criterio de transparencia puede ayudar a esclarecer las injusticias facilitadas por IA basadas en la culpa: por ejemplo, esclarecer un abuso cometido por A contra B facilitado por estas tecnologías en la medida en que es posible determinar una relación causal. No podrá, sin embargo, solventar los problemas que derivan de una injusticia estructural exacerbada por las nuevas tecnologías, en la medida en que en las injusticias estructurales un elemento fundamental es la no-trazabilidad: no se trata de que A cause B, sino del cúmulo de efectos conscientes e inconscientes de todas las partes involucradas. En este sentido, el criterio de la transparencia parece insuficiente ante un problema estructural de la inteligencia artificial, lo cual no exime de exigir responsabilidades individuales. En el caso de los *deepfake* pornográficos no consentidos, independientemente de la falsedad de las imágenes, el objetivo es disminuir la credibilidad y autoridad epistémica de la persona representada. Watson (2021) sugiere, en relación con el fenómeno de los *deepfake* (*identity swap*) el concepto de insulto epistémico (*epistemic insult*), un daño ejercido contra el estatus y capacidad de una persona como agente epistémico. Asimismo, este fenómeno puede encuadrarse dentro del concepto de injusticia epistémica de Miranda Fricker (2017) (Walmsley, 2023; Kay *et al.*, 2024; Barry y Stephenson, 2025).

En relación con esta discusión, Katherine Hayles (2023) establece una distinción entre la crítica epistemológica y ontológica sobre las herramientas de aprendizaje automático en *Technosymbiosis. Figuring (Out) Our Relations to AI*. Es decir, una distinción entre cómo conocemos estas nuevas tecnologías y lo que son (Parikka, 2021). En la crítica de Hayles, el aspecto epistemológico refiere a la veracidad y al sesgo. Hayles da cuenta de cierto problema derivado del aprendizaje de estos sistemas, dado que “los algoritmos cambian constantemente a medida que el sistema aprende, por lo que la transparencia en un punto es oscuridad en otro” (Hayles, 2023, p.11, traducción propia). Hayles considera entonces una orientación

ontológica que habría de reconocer la interioridad del problema sobre la IA para plantear estrategias feministas eficaces. Es decir, rechaza la posibilidad de una solución fuera del sistema, como si existiese un espacio neutro y ajeno al problema desde el cual ofrecer una solución. Hayles sugiere dentro de las posibles estrategias feministas el concepto de *ensamblaje cognitivo* entre sistemas humanos, no humanos y medios computacionales (Hayles, 2023). A partir de la metáfora de la *tecnosimbiosis*, propone sumergirnos en los medios computacionales, cuestiona la supuesta inmaterialidad de la sociedad digital y enfatiza el carácter material de estas tecnologías dentro de la propuesta de los nuevos materialismos sugerida por Braidotti (2000) y DeLanda (1996) ³⁴y señalada por Tuin y Dolphijn, 2010³⁴. Trasladándolo al caso de los *deepfake*, no solo habríamos de examinar la imagen sintética en cuanto al resultado visible, sino la infraestructura que la posibilita y la materialidad de los cuerpos afectados: frente al *sesgo*, la *opresión*.

4. Conclusiones

A modo de conclusión, más allá de la discusión por la posibilidad de una IA feminista, proponemos establecer estrategias feministas en torno a la IA: trazar propuestas desde la investigación y el arte críticas y conscientes de lo que suponen estos procesos. No solo en cuanto a su resultado, como vemos claramente con los *deepfakes* al incidir en la cosificación, sino también en cuanto a su proceso. Hemos de comprender dónde y cómo se producen estas tecnologías, qué recursos materiales y humanos requieren.

El enfoque relacional del ensamblaje cognitivo tal vez pueda ayudar a desmitificar estas tecnologías y a cuestionar los enfoques abstractos centrados en la verificación y la transparencia. Hoy, al examinar las nuevas tecnologías desde un punto de vista bioético y desde los nuevos materialismos feministas, cabe considerar el papel moral del cuerpo involucrado en estas relaciones, pero también del medio ecológico. No solo hablamos de violencia física y simbólica de personas concretas afectadas por *deepfakes*, que también, sino que hemos de conectar esta forma de violencia con el extractivismo de datos, del territorio y de los saberes. Es decir, hemos de situar estas imágenes dentro las dinámicas de violencia estructural facilitadas por las tecnologías. Se trata pues de entender los sistemas de

aprendizaje automático como ensamblajes de actores humanos y no humanos frente a una fetichización de los datos y el código (Ananny y Crawford, 2016). El enfoque del ensamblaje puede aportar una visión relacional en torno a los *deepfake*, defendiendo la posibilidad de nuevas estrategias y nuevos significados de los sistemas humanos-técnicos, así como visiones no esencialistas de los cuerpos implicados en estas relaciones.

BIBLIOGRAFÍA

- Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *The state of deepfakes: Landscape, threats, and impact*. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf
- Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
- Barry, I., & Stephenson, E. (2025). The gendered, epistemic injustices of generative AI. *Australian Feminist Studies*, 1–21. <https://doi.org/10.1080/08164649.2025.2480927>
- Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *arXiv*. <https://doi.org/10.48550/arXiv.1607.06520>
- Braidotti, R. (2000) 'Teratologies'. En I. Buchanan and C. Colebrook (Eds.), *Deleuze and Feminist Theory*, Edinburgh: Edinburgh University Press, pp. 156-72
- Browne, J. (2023). AI and structural injustice: A feminist perspective. In J. Browne *et al.* (Eds.), *Feminist AI: Critical perspectives on algorithms, data, and intelligent machines* (Oxford ed., online ed., pp. 328-346). Oxford University Press. <https://doi.org/10.1093/oso/9780192889898.003.0019>
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine*

Learning Research, 81, 77–91.
<https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. The MIT Press.
<https://doi.org/10.7551/mitpress/12255.001.0001>

Crawford, K. (2023). *Atlas de IA. Poder, política y costes planetarios de la inteligencia artificial*. Ned Ediciones.

Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. *AI & Society*. <https://doi.org/10.1007/s00146-021-01162-8>

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, 139-167. Recuperado a partir de: <https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1052&context=uclf>

DeLanda, M. (1996), 'The Geology of Morals: A Neo-Materialist Interpretation.' Recuperado a partir de: <http://www.t0.or.at/delanda/geology.htm>

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. The MIT Press.
<https://doi.org/10.7551/mitpress/11805.001.0001>

Drage, E., & Frabetti, F. (2023). AI that matters: A feminist approach to the study of intelligent machines. In *Feminist AI* (1st ed., pp. 274–289). Oxford University Press. <https://doi.org/10.1093/oso/9780192889898.003.0016>

Fricker, M. (2017). *Injusticia epistémica*. Herder Editorial. Recuperado a partir de: <https://www-digitaliapublishing-com.bucm.idm.oclc.org/a/61337>

García-Ull, F. J. (2021). Deepfakes: El próximo reto en la detección de noticias falsas. *Anàlisi: Quaderns de Comunicació i Cultura*, 64, 103–120.
<https://doi.org/10.5565/rev/analisi.3378>

- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *arXiv*. <https://doi.org/10.48550/arXiv.1406.2661>
- Hayles, N. K. (2023). Technosymbiosis: Figuring (out) our relations to AI. In J. Browne *et al.* (Eds.), *Feminist AI: Critical perspectives on algorithms, data, and intelligent machines*. (Oxford ed., online ed., pp. 1-18). Oxford University Press. <https://doi.org/10.1093/oso/9780192889898.003.0001>
- Joque, J. (2024). *Matemáticas revolucionarias: Inteligencia artificial, estadística y la lógica del capital*. Verso.
- Kay, J., Kasirzadeh, A., & Mohamed, S. (2024). Epistemic injustice in generative AI (Version 1). *arXiv*. <https://doi.org/10.48550/ARXIV.2408.11441>
- Maddocks, S. (2020). 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, 7(4), 415–423. <https://doi.org/10.1080/23268743.2020.1757499>
- Matthews, T., & Kidd, I. J. (2023). The ethics and epistemology of deepfakes. In C. Fox & J. Saunders (Eds.), *The Routledge Handbook of Philosophy and Media Ethics* (pp. 342–354). Routledge. <https://doi.org/10.4324/9781003134749-33>
- Parikka, J. (2021). *Una geología de los medios*. Caja Negra.
- Reas, C. (2019). *Making Pictures with Generative Adversarial Networks*. Anteism Books.
- Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach*. Pearson.
- Simó Soler, E. (2023). Retos jurídicos derivados de la inteligencia artificial generativa: Deepfakes y violencia contra las mujeres como supuesto de hecho. *Indret*, 2, 493–515. <https://doi.org/10.31009/InDret.2023.i2.11>
- Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131–148. <https://doi.org/10.1016/j.inffus.2020.06.014>

- Van der Tuin, I. & Dolphijn, R (2010). The Transversality of New Materialism. *Women: A Cultural Review* 21.2: 153-71.
<https://doi.org/10.1080/09574042.2010.488377>
- Wagner, T. & Blewer, A. (2019). "The Word Real Is No Longer Real": Deepfakes, Gender, and the Challenges of AI-Altered Video. *Open Information Science*, 3(1), 32-46. <https://doi.org/10.1515/opis-2019-0003>
- Walmsley, J. (2023). Computer says no: artificial intelligence, gender bias, and epistemic injustice. In Mary L. Edwards & S. Orestis Palermos, *Feminist philosophy and emerging technologies*. New York, NY.
- Watson, L. (2021). *Epistemic rights and why we need them*. Routledge.

Relaciones y lazos con la IA, efectos en las relaciones interpersonales

Miguel Enrique Naredo Rojas

Universidad de Oviedo

RESUMEN

El presente artículo analiza el impacto de la Inteligencia artificial (IA), con especial atención en los modelos conversacionales y chatbots, sobre las relaciones entre humanos y entre humanos y máquinas en la era digital. A medida que las interacciones virtuales se vuelven más usuales, la IA se integra en los espacios de sociabilidad, generando vínculos que imitan las relaciones humanas, aunque poseen características completamente distintas a las de cualquier sujeto humano.

Partiendo de las ideas de Franco “Bifo” Berardi para conceptualizar estas relaciones, se distinguirá entre la comunicación conjuntiva (afectiva, encarnada, empática, sensitiva) y la comunicación conectiva (digital, sintáctica, sin interioridad). La relación con la IA pertenece a esta última, y se caracteriza por ser unidireccional y jerárquica, lo que afecta relevantemente a la calidad de las relaciones interpersonales humanas y a las expectativas que se tienen de las mismas.

Seguidamente, se estudiará la tendencia a humanizar a estos agentes virtuales y como ello deriva en la creación de lazos afectivos. Este fenómeno es

alimentado por aplicaciones como Replika, Flipped y Character.Ai, que ofrecen relaciones simuladas y personalizables, presentando riesgos para la esfera de la sociabilidad. Por último, se advierte sobre los riesgos de sustituir vínculos humanos por relaciones simuladas, lo que puede acentuar el aislamiento, distorsionar la percepción de lo afectivo y precarizar aún más el tejido social.

PALABRAS CLAVE: Inteligencia artificial, Relaciones, Comunicación, Sociabilidad.

1. Robots que hablan y personas que aman

En la era digital habitamos constantemente entornos virtuales donde interactuamos con cientos de personas a diario. Las redes sociales y la web se han convertido en los principales espacios de sociabilidad, permitiéndonos establecer lazos con individuos de todo el mundo. Pero, ¿qué tipo de relaciones construimos en estos entornos? ¿Cómo se diferencian de las interacciones cara a cara? Si bien estas preguntas son de gran interés, podemos llevar la reflexión un paso más allá. Entre las tecnologías en auge se encuentran las inteligencias artificiales (IA). Estas pueden ser de diversos tipos y funcionalidades, destacando especialmente las IA generativas. Actualmente existen múltiples IA conversacionales y chatbots en el espacio digital, que constituyen nuevos actores dentro de la red de sociabilidad.

Así como ya no es posible separar al sujeto de las tecnologías con las que comprende la realidad, las inteligencias artificiales empiezan a ocupar un lugar igualmente significativo en el plano relacional. En los espacios digitales dejamos de interactuar exclusivamente con perfiles humanos: en la IA encontramos algo extrañamente cercano a lo que podríamos experimentar con un desconocido, un vecino, un amigo, un familiar e incluso una pareja sentimental (los patrones lingüísticos sofisticados y adaptables junto a la proyección de la empatía del sujeto, dan la sensación de una presencia relacional cercana a las interacciones humana en distintos grados). No se trata de atribuirles identidad o conciencia, sino de reconocer que su nivel actual de desarrollo permite interacciones sorprendentemente similares a las que mantenemos con otras personas en el entorno digital. Frente a ello, surgen preguntas ineludibles: ¿qué tipo de vínculo estamos construyendo con la IA? ¿Podemos desarrollar alguna forma de apego o

emoción hacia estas entidades? ¿Y hasta qué punto estas interacciones están transformando nuestra percepción de las relaciones humanas?

La tesis central de este artículo es que las nuevas formas de relación e interacción existente con la IA pueden derivar en una distorsión tanto de los marcos por los que comprende el sujeto a otros individuos, como en la sociabilidad de las personas y las formas de relacionarse que poseen los sujetos entre sí (en todos sus grados de implicación emocional).

Recientemente hemos visto anunciada la primera tentativa de matrimonio entre una IA y un ser humano (Vilches, 2023). También se reportó la trágica historia de un joven enamorado de un chatbot (Payne, 2024). Con ello no estoy afirmando que el futuro emocional de nuestra especie se asemeje al de la película *Her* (2013), pero el éxito de aplicaciones como Replika, Flipped o Character Ai, que cuentan con más de diez millones de descargas cada, las cuales promueven relaciones virtuales significativas con IA sugiere que este fenómeno se merece nuestra atención. Con todo, nos queda preguntarnos: ¿qué efecto tendrá la forma en la que nos relacionamos con la IA respecto de las relaciones interpersonales? ¿Estaríamos humanizando las IA u objetivando a las personas? ¿Desplazará las relaciones en persona a un segundo plano (aún más profundo)? ¿Cuáles serán las consecuencias psicológicas y sociales de estas interacciones? El avance de estas tecnologías nos obliga a cuestionar el carácter de las relaciones humanas, y replantearnos los límites entre lo virtual y lo real.

2. Un nuevo marco relacional

En este apartado analizaré qué tipo de lazo tenemos con la IA y qué forma toma a partir del trabajo de Franco “Bifo” Berardi sobre la comunicación conjuntiva y conectiva (Berardi, 2017). La comunicación encarnada —enfoque donde la mente, el cuerpo y el entorno están estrechamente entrelazados en los procesos comunicativos, el cuerpo no solo es un vehículo para expresar mensajes, sino que es parte activa del proceso de pensar, sentir y comunicar, significando una mayor implicación entre los actores— se contrapone a la comunicación de masas (digital) —comunicación que usa medios artificiales para transmitir mensajes a uno o más actores, prescindiendo del cuerpo y centrando todo el peso

comunicativo en el signo—: pasamos de una comunicación cara a cara, donde no se reduce todo a imágenes o representaciones (signos), y donde la piel ocupa un lugar esencial en la sensibilidad, la sensibilidad y lo relacional. El tacto es clave para percibir diversas señales e informaciones, permitiendo comunicaciones sintientes que van más allá del lenguaje y el signo.

Sin embargo, con la digitalización y la expansión de los medios conectivos, transitamos hacia un nuevo modelo comunicativo (Berardi, 2017). La posibilidad de traducir los actos físicos en información es el elemento clave de esta transición. En este sentido, “la automatización de la interacción lingüística y el reemplazo de los actos cognitivos y afectivos por secuencias y protocolos algorítmicos es la principal tendencia de la mutación en curso” (Berardi, 2017, p.120). Estamos ante una especie de “sintactización del mundo”, que se comprende a través de la diferencia entre conjunción y conexión, dos modelos comunicacionales vinculados a una tecnología y una época determinada. Mientras que la conjunción se basa en la empatía, la conexión se sustenta en la sintaxis.

La conjunción es la concatenación de cuerpos y máquinas capaces de producir significado sin ajustarse a un diseño preestablecido (Berardi, 2017, p.28). Por ello, se relaciona con la idea de empatía y con un modo de relación que implica vincularse con el otro, respondiendo a sus sentimientos y emociones. La conexión, en cambio, se basa en la conformidad y la adaptación a una estructura sintáctica, en lugar de la empatía. Es una concatenación de cuerpos y máquinas que solo puede generar significado obedeciendo a un diseño intrínseco y respetando reglas precisas de funcionamiento (en este sentido, siguen un mismo formato sintáctico), como ocurre en la interacción puntual y repetitiva de funciones algorítmicas. Ahora bien, no se trata de optar por una u otra, sino de explorar los matices entre ambas y observar la transición de un modelo comunicativo predominantemente conjuntivo hacia uno conectivo.

Este marco teórico sobre los cambios en la comunicación también deja entrever que, en consecuencia, los lazos afectivos e interpersonales se ven perjudicados por las formas actuales de comunicación. Tanto los espacios donde predomina la comunicación conectiva como los lazos que se establecen a través

de esta se enmarcan en un mercadeo de signos que carece de cualquier interioridad e implicación personal en la interacción. En este contexto, la relación con la IA se inscribe dentro de un modelo conectivo más que conjuntivo. Es decir, nuestra comunicación con la IA será meramente sintáctica y basada en el intercambio de símbolos, sin semántica ni empatía que la sustente. Esto tiene pleno sentido, pues se trata de una interacción unidireccional.

3. ¿Por qué relaciones con la IA?

El uso de la inteligencia artificial se proyecta hacia casi todos los ámbitos de la vida, por lo que ignorar el tipo de relación que entablamos con ella sería una negligencia. Cada vez son más las esferas en las que los actores humanos son sustituidos por agentes virtuales (modelos de IA entrenados para actividades específicas), lo que implica el reemplazo de diversos roles anteriormente desempeñados por personas, como la atención al cliente, la asistencia virtual, la secretaría, la gestión y otros servicios digitales.

Estos modelos y agentes virtuales, cada vez más comunes, están diseñados con unas intenciones determinadas, ya que según estudios (Pelau et al., 2021) tendemos a sentir mayor comodidad y familiaridad con una IA que muestre rasgos humanos en la interacción, como la empatía, la compasión o la gratitud. Por ello, podemos observar una tendencia a humanizar e incluso atribuir una conciencia o agencia a estos modelos, a pesar de que somos conscientes de que carecen completamente de cualquier rasgo humano y conciencia. Por otro lado, la quimérica tendencia a humanizar la IA ha derivado en la creación de lazos afectivos con ella. De hecho, según un estudio estadounidense (Brooks, 2024), el 31 % de los entrevistados estaban abiertos a tener una relación con una IA, lo que implica no solo la integración de un nuevo actor en el mercado laboral, sino la creación de un agente que será objeto de pulsiones y pasiones ligadas al deseo del sujeto.

En este contexto, los servicios de chatbots están cada vez más solicitados, lo que evidencia la creciente inclinación a establecer lazos afectivos con la IA. Aplicaciones que promueven relaciones virtuales significativas con modelos de

IA. El hecho de que cada una de estas aplicaciones supere los diez millones de descargas sugiere que este fenómeno merece nuestra atención.

4. ¿Qué IA?

A lo largo de este escrito no hemos definido ni descrito de forma estricta a que nos referimos con IA. En nuestro caso, seguimos los pasos de Carlos Madrid y nos apoyamos en sus ideas expuestas en su texto *Filosofía de la Inteligencia artificial* (Madrid, 2024). Cuando hablamos de inteligencia artificial, nos referimos a una tecnología. Esta, a su vez, se clasifica dentro de las categorías de IA débil¹—una IA no es una mente, sino que su estructura depende del conjunto de cerebros humanos que la diseñaron— e IA específica²— ya que actúa como si fuera inteligente, simulando la inteligencia humana en tareas concretas—, pero no como una IA fuerte o general (Madrid, 2024). En nuestro caso particular, nos referimos principalmente a modelos de lenguaje, es decir, algoritmos diseñados para el intercambio de signos. Pese a que aparentemente la IA débil y la IA específica se correspondan la una a la otra, no es una relación estricta: pueden existir IA débiles que no sean específicas. Sin embargo, en nuestro caso sí coinciden.

Es fundamental separar esta tecnología de las nebulosas ideológicas, las teorías especulativas y la ficción en torno a la IA, distinguiendo su realidad y materialidad. Por ello, nos alejaremos de conceptos como la singularidad y de los elementos de la cultura popular que la rodean, con el fin de realizar un análisis

¹ La distinción entre IA fuerte e IA débil es introducida por John Searle en un artículo de 1980 titulado *Minds, brains, and programs* (Searle, 1980). La IA débil se refiere a sistemas diseñados para realizar tareas específicas y restringidas, mientras que la IA fuerte aspira a poseer una capacidad similar a la humana (con razonamiento, aprendizaje y adaptación).

² La distinción entre IA específica e IA general se basa en su alcance funcional: la primera se centra en tareas concretas, imitando al sujeto, mientras que la segunda busca replicar la capacidad cognitiva humana, incluida la autoconciencia. A diferencia de la clasificación entre IA débil y fuerte —más filosófica— esta diferenciación es de carácter técnico. Aunque diversos autores han contribuido a su desarrollo, destacan Shane Legg (Legg, 2008) y Ben Goertzel (Goertzel & Pennachin, 2007), quienes popularizaron el concepto de IA general.

adecuado. Dejamos ese ámbito imaginativo y futurista a la esfera cultural y sus productos, como *Her*, (2013) o *Ghost in the Shell* (1995).

5. ¿De qué carácter es esta relación?

Nuestra relación con la IA es, en primer lugar, conectiva. En términos comunicativos, intercambiamos información a través de signos, sin embargo, se trata de un pseudo-diálogo. Nosotros proporcionamos un *input*, que la IA interpreta y procesa para generar una respuesta (*output*) basándose en los criterios que considera adecuados para esos signos. Es importante reiterar que un modelo de lenguaje carece de semántica (Searle, 1980), conciencia, sensibilidad o percepción. Su funcionamiento se basa únicamente en la sintaxis. No obstante, resulta extremadamente relevante que percibamos rasgos humanos en esa comunicación. Cuanto más empáticas y sensibles parecen las respuestas de los chats, mayor aceptación tienen estos modelos por parte de los usuarios (Pelau et al., 2021). Es decir, a pesar de tratarse de una relación unidireccional con un sistema que solo procesa signos, los usuarios suelen sentirse más cómodos interactuando con una IA que simule un interlocutor humano. En otras palabras, son preferibles los modelos que más se parezcan a los seres humanos en su interacción. En este sentido, y en el marco de la comunicación conectiva, cabe preguntarse si, en nuestra relación con los modelos de lenguaje, estamos humanizando la IA —en la medida en que percibimos un ego en ella— o si, por el contrario, estamos objetivando las relaciones interpersonales. Reduciendo nuestra percepción del otro a una mera interpretación de signos, en lugar de reconocer los rasgos empáticos, sensitivos y sensibles propios de la comunicación humana, reducimos el otro a un intérprete y emisor de signos, donde prevalece el intercambio sintáctico sobre la implicación con el otro; o, en otras palabras, la conexión sobre la conjunción. Desde otra perspectiva, podríamos cuestionarnos si identificamos a la IA como un sujeto porque se asemeja a nuestras formas de relacionarnos o si, por el contrario, nuestras formas de relación en la era digital han decaído tanto que pueden ser fácilmente simuladas por una IA.

En esta línea, los entornos digitales y la creciente interacción con inteligencias artificiales evidencian la creciente dificultad para mantener

relaciones emocionalmente significativas. Este fenómeno está marcado, en gran medida, por la transición de un modelo conjuntivo, basado en la implicación afectiva, a un modelo conectivo, caracterizado por la interconexión sin un vínculo emocional profundo. El rasgo característico de nuestra relación con la inteligencia artificial es su unidireccionalidad y jerarquización. No podemos perder de vista que los modelos de lenguaje son servicios diseñados para estar a nuestra merced. Es decir, se adecuan a cualquier tipo de pregunta o respuesta con el objetivo de satisfacer al usuario. Esto nos permite moldear la relación, la información y las respuestas de los modelos según nuestros gustos e intereses. Podemos proporcionarles ciertos datos y luego contradecirlos por completo, y, aun así, la IA nos dará la razón y buscará complacernos. La interacción con la IA funciona como una caja de resonancia, de manera similar a los algoritmos de las redes sociales: el individuo queda encerrado en su propio mundo de intereses y pensamientos, siendo constantemente validado en su interacción con el “otro”.

En suma, nos sumergimos en una dinámica autorreferencial en la que la IA nunca nos contradice ni genera una interacción que vaya más allá de la aprobación y la sumisión. Como consecuencia, el individuo adopta un patrón relacional alejado de la realidad de los entornos sociales, donde la interacción suele implicar confrontación, desacuerdo y negociación de significados.

También debemos considerar cómo funcionan estas tecnologías. Por ejemplo, Replika no es solo un chat basado en IA, sino que tiene la clara pretensión de ofrecer relaciones afectivas significativas como servicio. En este sentido, permite a los usuarios modificar y crear el perfil que deseen, así como definir desde cero el tipo de relación que buscan. En esencia, estamos hablando de relaciones emocionales como servicio.

6. ¿No representa esto un peligro aún mayor para las ya precarias relaciones interpersonales que mantenemos en la actualidad?

Dentro de las características personalizables de la aplicación, los usuarios pueden crear memorias, definir la personalidad, el carácter, los gustos e incluso el avatar de su IA. En otras palabras, diseñar a la persona y la relación exacta que desean tener. Sin embargo, no podemos ignorar que la interacción con estos

chatbots sigue estando marcada por la unidireccionalidad y la sumisión. En consecuencia, nos encontramos ante cajas de resonancia diseñadas para satisfacer las necesidades emocionales de los usuarios, reforzando una dinámica en la que no hay confrontación ni reciprocidad genuina.

Esto puede desencadenar dificultades sociales y perjuicios en las relaciones interpersonales. Si la relación con los modelos de lenguaje se estandariza, esto podría llevar a una precarización aún mayor de los vínculos afectivos significativos. Además, puede fomentar la ansiedad derivada de la incapacidad de socialización, así como comportamientos de riesgo, incrementando incluso el peligro de aislamiento social (Mansfield et al., 2025).

Asimismo, estas interacciones establecen expectativas irreales sobre las relaciones interpersonales. En este sentido, los modelos de lenguaje que buscan ofrecer relaciones emocionalmente significativas son preocupantemente similares a la pornografía, ya que ambos generan una simulación de intimidad que distorsiona la realidad del vínculo humano.

7. Conclusiones

Como afirma Paola Llanea (abogada, auditora de sistemas y consultora de seguridad miembro del Consejo Asesor Internacional de Inteligencia Artificial: grupo de asesoramiento al gobierno sobre inteligencia artificial y seguridad digital), “nos encontramos ante tecnologías que acentúan la soledad para, una vez instaurada, vendernos servicios que prometen solucionarla”. Sin embargo, estas soluciones no son más que parches que, lejos de aliviar la situación, agravan el aislamiento y la desconexión emocional (Moeve, 2025). No obstante, no podemos tachar nuestra relación con la IA como un peligro categorial, pues es de interés los resultados que se pueden obtener en situaciones de vulnerabilidad emocional, donde las personas carecen de los recursos necesarios para desarrollarse socialmente, estas herramientas pueden representar un primer paso hacia la interacción en la esfera social, brindando cierto apoyo temporal. Sin bien esta dimensión no se ha abordado dentro de este artículo, sigue siendo un ámbito en el que merece la pena indagar.

En suma, no pretendo ofrecer respuestas definitivas, sino poner sobre la mesa esta dimensión de la inteligencia artificial y reflexionar sobre lo que implica para nosotros como usuarios y para la sociedad, especialmente en el ámbito de la comunicación y las relaciones humanas (aunque estas sean unidireccionales).

La IA genera tanto beneficios individuales como perjuicios a la esfera social, y viceversa. Por ello, es fundamental analizar críticamente su impacto y el papel que queremos que juegue en nuestras vidas.

BIBLIOGRAFÍA

- Berardi, F. (2017). *Fenomenología del fin: Sensibilidad y mutación conectiva*. Caja Negra.
- Brooks, A. (2024, 28 de mayo). *Survey: Is Having an AI Boyfriend or Girlfriend Cheating?* DatingAdvice.Com. Recuperado el 22 de marzo de 2025 de <https://www.datingadvice.com/studies/survey-is-having-an-ai-partner-cheatin>
- Goertzel, B., & Pennachin, C. (2007). *Artificial general intelligence*. Springer. <https://doi.org/10.1007/978-3-540-68677-4>
- Legg, S. (2008). *Machine Super Intelligence* (Doctoral dissertation, University of Lugano). Swiss Open Access Repository.
- Madrid Casado, C. M. (2024). *Filosofía de la inteligencia artificial*. Pentalfa.
- Mansfield, K. L., Ghai, S., Hakman, T., Ballou, N., Vuorre, M., & Przybylski, A. K. (2025). From social media to artificial intelligence: Improving research on digital harms in youth. *The Lancet Child & Adolescent Health*, 9(3), 194-204. [https://doi.org/10.1016/S2352-4642\(24\)00332-8](https://doi.org/10.1016/S2352-4642(24)00332-8)
- Moeve. (2025, 11 de marzo). *Todo lo que pasará / ¿Nos sustituirán las máquinas? La IA con Paul Branteghem y Paloma Llana*. [Vídeo]. Youtube. Recuperado el 22 de marzo de 2025 de <https://www.youtube.com/watch?v=zHQL-ica6qA>

- Payne, K. (2024, octubre 25). An AI chatbot pushed a teen to kill himself, a lawsuit against its creator alleges. *AP News*. Recuperado el 28 de febrero de 2025 de <https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdb3c90d1456bd0>
- Pelau, C., Dabija, D.-C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, 122, 106855. <https://doi.org/10.1016/j.chb.2021.106855>
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. doi: 10.1017/S0140525X00005756.
- Vilches, S. (2023, noviembre 29). La primera mujer en casarse con una IA: «No descartamos la idea de tener hijos». *El Mundo*. Recuperado el 22 de marzo de 2025 <https://e00-elmundo.uecdn.es/cultura/arte/2023/11/29/656766b2e9cf4a15318b4571.html>

Responsabilidad e inteligencia artificial

Patricia Fernández Fernández

Universidad de Oviedo

RESUMEN

Este artículo examina el problema de la responsabilidad en el uso de sistemas de Inteligencia Artificial (IA) desde una perspectiva filosófica, con especial atención a los modelos actuales basados en aprendizaje automático. Se parte de una definición funcional de IA que enfatiza su carácter instrumental y no intencional, lo cual impide atribuirle agencia moral o responsabilidad legal. A través del análisis de las limitaciones del razonamiento abductivo, la confusión entre correlación y causalidad, y la opacidad de los algoritmos, se argumenta que la responsabilidad debe recaer exclusivamente en los sujetos humanos — diseñadores, programadores, usuarios o instituciones— que los implementan y emplean. El artículo discute además el concepto de agencia indirecta en la toma de decisiones mediada por IA, mostrando cómo esta puede diluir la percepción de culpa en los agentes humanos. Finalmente, se defiende que el reconocimiento de la IA como herramienta no neutral exige una evaluación ética, moral y política por parte de las sociedades humanas, que deben asumir su responsabilidad respecto a sus consecuencias.

PALABRAS CLAVE: inteligencia artificial, responsabilidad, agencia, ética, aprendizaje automático

1. Introducción

La proliferación de la Inteligencia Artificial (IA) en los numerosos ámbitos en los que es aplicada —como la sanidad, las administraciones y en el ámbito político— la ha convertido en una herramienta tecnológica generalizada e integrada en nuestra sociedad. Esto lleva a que cada vez en más ocasiones sea una máquina o un algoritmo informático quien tome decisiones que afectan al bienestar de las personas y de los grupos sociales, lo cual plantea diferentes dilemas: ¿es la IA responsable si se comete algún error o perjuicio? En caso afirmativo, ¿cómo asegurar la supervisión humana del sistema, no solo el caso de que el ser humano intervenga en todo el ciclo de decisión del sistema (*human in the loop*), sino en los casos de que solo intervenga en su monitorización (*human on the loop*) o, en el peor de los escenarios, en aquellos en los que no sea necesaria ninguna acción humana (*human out the loop*)? (Madrid Casado, 2024). En definitiva, la cuestión reside en preguntarse sobre quién recae la responsabilidad de las decisiones tomadas por Inteligencia Artificial y, por consiguiente, si es posible establecer una cadena de responsabilidad en caso de perjuicio.

Desde una perspectiva teórico-filosófica, el presente trabajo se articula en torno a dos ejes: por una parte, se delimita conceptualmente qué entendemos por inteligencia artificial y se analizan las implicaciones técnicas de su funcionamiento; por otra, se examina la cuestión de la responsabilidad en su diseño y aplicación, atendiendo a sus consecuencias éticas y políticas.

2. Definición de Inteligencia Artificial

Adentrarse en el campo de la Inteligencia Artificial (IA) y los temas filosóficos involucrados en su desarrollo requiere acotar su definición. Tomando como base las ideas expuestas en *Filosofía de la inteligencia artificial* (2024) por Carlos Madrid, se puede definir la «Inteligencia Artificial» como una herramienta tecnológica que funciona a través de modelos estadísticos de procesamiento de la información. Su funcionamiento se basa en la ejecución, recreación o simulación de

operaciones trabadas entre sí que —según los datos de entrada (*inputs*) proporcionados— están orientadas hacia un mismo fin (*output*), como es diseñar y construir artefactos que realicen tareas concretas con mayor eficacia y eficiencia sin necesidad de la supervisión humana.

Esta aparente autonomía puede generar la ilusión de conciencia o agencia, a pesar de que los sistemas de IA carecen de estados intencionales o comprensión. Tal como subrayó Searle (1980) en su crítica al funcionalismo computacional, el hecho de que un sistema simule un proceso cognitivo no implica que lo entienda o lo experimente realmente. Esta idea es desarrollada también por Madrid Casado (2024), quien advierte que lo que a menudo interpretamos como autonomía en los sistemas de IA no es más que automatismo: una ejecución técnicamente compleja de instrucciones predefinidas, sin que exista elección de fines ni capacidad para modificarlos. Como explica el autor,

sus capacidades son el resultado de la ingeniería y la programación. Los sistemas con IA carecen de conductas propositivas: el sistema no elige los fines, que le vienen dados como instrucciones de diseño, ni puede cambiar los que le han sido acotados (sea jugar al ajedrez o llevar al pasajero de un coche autónomo a donde éste diga y no a donde quiera el ordenador central). La intencionalidad que parecen mostrar no es más que un préstamo de la depositada en ellos por sus artífices. Más que intencionalidad, denotan funcionalidad. Por otro lado, para ser un sujeto ético y, en concreto, un agente ético, se precisa un mínimo de comprensión (saber, por ejemplo, lo que es infligir daño a otra persona), pues de lo contrario no puede haber responsabilidad ética alguna. (Madrid Casado, 2024, p. 118).

En consecuencia, la intencionalidad y la comprensión no son propiedades que puedan atribuirse a los sistemas de IA, sino que pertenecen exclusivamente a los sujetos humanos. Por ello, no cabe considerar a la IA como agente moral o político, ni atribuirle responsabilidad alguna: no son agentes, sino sistemas diseñados para ejecutar funciones determinadas por terceros. Por otro lado, cabe resaltar que los sistemas de IA que se utilizan hoy en día no son los sistemas transparentes y

explicables de la IA simbólica¹. Actualmente, los modelos predominantes son los basados en la inducción y el aprendizaje automático (*machine learning*)² cuyo uso introduce una serie de dificultades en la comprensión y supervisión del proceso de toma de decisiones. Estas serán abordadas a continuación.

2.1. Problemas en la comprensión del procesamiento de toma de decisiones

Como se ha señalado, los métodos inductivos propios del *machine learning* presentan tres problemas en cuanto a la transparencia y explicabilidad de los procesos de toma de decisiones por la Inteligencia artificial, a saber: a) el problema del razonamiento abductivo, b) la falacia de la correlación no implica causalidad y c) la opacidad de los algoritmos, que dificulta la comprensión de los procesos de toma de decisiones y, en consecuencia, la corrección de errores.

a) Razonamiento abductivo. A diferencia de los humanos la IA no es capaz de hacer conjeturas ante un hecho en función de la experiencia y el contexto. La experiencia con la que cuenta la IA proviene de la acumulación de datos, pero no cuenta con la integración de las experiencias ni de la intuición propia de los seres humanos a la hora de tomar decisiones.

b) La correlación no implica causalidad. El análisis de datos no es suficiente para abarcar la inferencia causal. La correlación no es necesariamente causación, sino que puede ser asociación. Como expone López de Mántaras,

los sistemas actuales basados en aprendizaje profundo simplemente pueden aprender funciones matemáticas simétricas, no pueden aprender relaciones asimétricas y por consiguiente no son capaces de diferenciar entre causas y efectos, como por ejemplo que la salida del sol es la causa del canto del gallo y no lo contrario (2018, p. 170)

¹ La IA simbólica se ocupa, principalmente, en la representación y manipulación de símbolos y conceptos a través de la implementación de reglas lógicas predefinidas y conocimiento explícito con el fin de tomar decisiones o realizar predicciones.

² El aprendizaje automático se despliega en varias partes. En el *aprendizaje supervisado*, se etiquetan los datos, desmembrándolos en entrada y salida, definiendo así el objetivo del algoritmo (por ejemplo, la clasificación de especies a partir de sus características). Por otro lado, está el *aprendizaje no supervisado* donde los datos no son etiquetados, de modo que no hay un objetivo definido de antemano. Asimismo, es incluido el *aprendizaje semisupervisado*, también llamado *aprendizaje por refuerzo* (Madrid Casado, 2024).

c) Opacidad. Algunos sistemas de IA, como las redes neuronales, funcionan como cajas negras, de modo que su confiabilidad queda en duda. En otras palabras, no poseen un diseño lógico deductivo, por lo que no es posible explicar por qué funcionan las reglas que confluyen, pero tampoco por qué fallan cuando lo hacen (Madrid Casado, 2024).

Entonces, determinar cómo la IA procesa esa información, en conjunto con los problemas que plantea, es crucial para la comprensión y aceptación de su criterio a la hora de atribuir responsabilidad sobre los resultados que genera. Dado que la IA carece de intención, es oscura, y no comprende sus procesos en la toma de decisiones, la responsabilidad de las acciones de una IA debe ser atribuida a los humanos. Los individuos, ya sean los usuarios o los programadores, son los únicos que pueden responder legal y moralmente a los daños o perjuicios que se desprenden de la utilización y aplicación de la misma en los procesos de toma de decisiones. Como advierten Winfield y Jirotko (2018), permitir que la toma de decisiones recaiga sobre sistemas cuyas operaciones internas son opacas —las llamadas cajas negras— compromete los principios de responsabilidad y trazabilidad. Dado que no resulta posible interpretar y explicar cómo se obtienen los *outputs* a partir de los *inputs* (por más que se ajusten los resultados a los datos de entrenamiento), la atribución de responsabilidad en caso de mal funcionamiento —como se expondrá a continuación— recae en los ingenieros y programadores, que optaron conscientemente por implementar dichos sistemas.

En suma, las limitaciones gnoseológicas, ontológicas y epistemológicas del estatuto del campo de la IA —como la falta de razonamiento abductivo, la imposibilidad de inferir causalidad y la opacidad algorítmica— no solo dificultan la comprensión de sus métodos de razonamiento, sino que suponen que se haga una revisión sistemática de sus implicaciones ética y políticas respecto a su desarrollo y aplicación. Esta cuestión conduce inevitablemente al problema de la atribución de responsabilidad, como elemento central del presente trabajo.

3. El problema de la atribución de responsabilidad

Si aceptamos que los sistemas de IA carecen de intencionalidad y conciencia, como se ha argumentado, resulta imposible considerarlos responsables

de sus actos. La cuestión, entonces, es cómo y en qué condiciones debe atribuirse esa responsabilidad a los agentes humanos que los diseñan, programan o utilizan.

Las decisiones generadas por sistemas automatizados —así como su desarrollo, consecuencias y gestión— deben ser evaluadas desde las implicaciones que tienen sobre los sujetos individuales (ámbito de la ética), los grupos sociales (ámbito moral) y los Estados (política). En otras palabras, se trata de ahondar sobre las consecuencias éticas y políticas de la utilización y diseño de sistemas de IA por los agentes humanos, a quienes sí atribuimos responsabilidad. Este mismo interrogante ha sido formulado de forma explícita por Mark Coeckelbergh:

¿Quién es responsable de los perjuicios y los beneficios que causa la tecnología cuando los humanos delegan capacidad de actuación y toma de decisiones en la IA? Reformulándolo en términos de riesgo: ¿quién es responsable cuando algo sale mal? Cuando los seres humanos hacemos cosas y tomamos decisiones, normalmente vinculamos la capacidad de actuar con la responsabilidad moral. Eres responsable de lo que haces y de lo que decides (Coeckelbergh, 2021, p. 95).

En su estudio empírico sobre la percepción del daño y la culpa en decisiones mediadas por IA, Espinosa y Clemente (2023) muestran que las personas tienden a atribuir menor responsabilidad moral cuando la acción se produce a través de una agencia indirecta (mediación por otros agentes). Esta dilución del juicio moral tiene implicaciones profundas en cómo entendemos la delegación en sistemas automatizados. Los autores exponen que cuando una persona actúa a través de otros agentes humanos en lugar de hacerlo directamente (por ejemplo, cuando el gerente de una empresa da instrucciones a sus supervisores para que las ejecuten) la distancia existente entre quien dicta la orden, quien la ejecuta y las consecuencias del comportamiento se reduce debido a la separación espaciotemporal en la ejecución de la decisión. En estos casos, la previsión de las consecuencias morales está comprometida. Las decisiones muestran menos consideración por los receptores de estas y se percibe menos responsabilidad y probabilidades de sufrir consecuencias perjudiciales por tomar decisiones poco éticas. Cuando se ejerce una agencia indirecta a través de la IA, las personas están más dispuestas a engañar, son responsabilizadas en menor medida y además anticipan que esto será así. En cambio, no ocurre lo mismo en aquellos casos en los

que se da una agencia directa. Cuando una persona provoca un daño, puede anticipar que será culpada por ello, incluso castigada por sus actos. Los castigos y recompensas tienen el objetivo de conseguir un cambio en el comportamiento del culpable de una acción, lo que no tiene sentido en el caso de una máquina.

Sin embargo, considerar la Inteligencia Artificial como agente —ya sea en sentido directo o indirecto— en los procesos de toma de decisiones no resulta, a mi juicio, una formulación adecuada. Esta interpretación parte de una comprensión equivocada de la IA, al atribuirle un estatuto similar al de un agente moral o político. Si atendemos a la definición aquí defendida, la IA, en tanto herramienta tecnológica basada en la ejecución, recreación o simulación de operaciones trabadas entre sí orientadas hacia la realización de las tareas concretas establecidas por un ser humano (o varios), no puede ser considerada responsable, ni moral ni legalmente, de los resultados de sus acciones.

Dado que los sistemas de IA no pueden ser castigados o recompensados, carece de sentido atribuirles responsabilidad alguna. Esta debe recaer, en cambio, sobre los humanos que las usan, las programan o diseñan, pues son los únicos que pueden responder legal y moralmente a los daños o perjuicios como consecuencia directa de su utilización. Así, del mismo modo que atribuimos responsabilidad a la persona que dirige o es propietaria de una empresa como responsable directo por las acciones de esta, también atribuiríamos la misma responsabilidad a quien usa, diseña o posee un sistema de IA. En esta línea, la responsabilidad no puede atribuirse a la IA en sí, sino únicamente a través de la agencia directa de quienes la diseñan, programan o utilizan. Como señala Coeckelbergh (2021), reconocer la IA como una tecnología social implica asumir que sus usos, efectos y consecuencias dependen de decisiones humanas y de estructuras institucionales que deben ser objeto de escrutinio ético y político. El uso de sistemas algorítmico en la gestión policial predictiva o en la concesión de ayudas sociales puede reproducir sesgos estructurales y consolidar formas de control social invisibilizadas.

En este sentido, tampoco pueden soslayarse los intereses políticos y económicos puestos en marcha que operan detrás de su diseño, desarrollo e implementación, y que remodelan el mundo. La IA no es un actor independiente,

sino una herramienta que responde a nuestras decisiones, valores e intereses. No se trata de una tecnología meramente formal o abstracta, sino de un artefacto que produce efectos materiales concretos. Atribuir responsabilidad implica, por tanto, reconocer que cada decisión delegada en una máquina remite, en última instancia, a una decisión humana que opera detrás.

4. Conclusiones

En definitiva, esta revisión sobre el uso de sistemas dotados de IA en contextos decisionales pone de relieve varias cuestiones clave en torno a la atribución de responsabilidad. En primer lugar, para que la IA sea aceptada en ámbitos donde se toman decisiones morales, éticas y políticas, debe generar confianza. Por este motivo, sus decisiones deben percibirse como justas en diferentes dimensiones que favorezcan los intereses legítimos de los usuarios y que permita identificar al responsable de los daños potenciales provocados por la IA sin que la propia IA sirva para enmascarar esta responsabilidad. En segundo lugar, si bien los seres humanos delegan ciertas decisiones en una máquina, la responsabilidad sigue siendo nuestra, así como de sus artífices, en caso de mal funcionamiento. Por ello, resulta imprescindible examinar críticamente cómo comprendemos la inteligencia artificial, su uso y su diseño, desde diferentes ámbitos —tanto como usuarios o programadores, como para las sociedades en su conjunto— con el fin de evitar que su uso desdibuje la responsabilidad de los agentes humanos en sus diversas aplicaciones.

BIBLIOGRAFÍA

- Cantú Martínez, P. C. (2024). Inteligencia artificial y sus connotaciones éticas. *Revista Iberoamericana de Bioética*, (25), 1–11. <https://doi.org/10.14422/rib.i25.y2024.008>
- Castelvecchi, D. (2016). Can we open the black box of AI?. *Nature*, 538(7623), 20–23. <https://doi.org/10.1038/538020a>
- Coekelbergh, M. (2021). *La ética de la inteligencia artificial*. Cátedra.

- Espinosa, P., & Clemente, M. (2023). La percepción de la toma de decisiones a través de inteligencia artificial cuando se produce daño a las personas. *Estudios Penales y Criminológicos*, 44(Ext.), 1–13. <https://doi.org/10.15304/epc.44.8917>
- House of Commons Science and Technology Committee. (2018). *Algorithms in decision-making*. UK Parliament. <https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/351.pdf>
- López de Mántaras, R. (2018). El futuro de la IA: hacia inteligencias artificiales realmente inteligentes. En *¿Hacia una nueva Ilustración? Una década trascendente* (pp. 160–174). BBVA.
- López de Mántaras, R. (2023, 14 de febrero). Inteligencia artificial: grandes retos [Conferencia]. Fundación Ramón Areces.
- Madrid, C. (2024). *Filosofía de la inteligencia artificial*. Pentalfa.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. <https://doi.org/10.1017/S0140525X00005756>
- Searle, J. (1990). ¿Es la mente un programa informático? *Investigación y Ciencia*, 162, 9-16.
- Winfield, A.F.T., Jirotko, M. (2017). The Case for an Ethical Black Box. En Gao, Y., Fallah, S., Jin, Y., Lekakou, C. (eds) *Towards Autonomous Robotic Systems*. TAROS 2017. Lecture Notes in Computer Science(), vol 10454. Springer, Cham. https://doi.org/10.1007/978-3-319-64107-2_21

La importancia ética de las inteligencias artificiales

Juan Villazón

Universidad de Oviedo

RESUMEN

Este trabajo ofrece una panorámica general de la polémica respecto a si los programas de inteligencia artificial constituyen sujetos éticos, y argumenta que no lo son en absoluto. En él defiende que toda la inteligencia artificial realmente existente es débil y específica, y que los proyectos de desarrollar una inteligencia artificial fuerte general son hoy completamente utópicos. Por ello, ni siquiera los modelos de lenguaje más desarrollados son conscientes o inteligentes en ningún sentido, ni se les puede atribuir conciencia ni responsabilidad. En este sentido, argumento que las inteligencias artificiales no son distintas de cualquier otra tecnología en sus aspectos éticos, lo que no significa en absoluto que sean neutrales o inertes. Muy al contrario, las inteligencias artificiales presentan una serie de desafíos éticos y políticos de gran importancia, pero no porque constituyan sujetos morales, sino porque modifican las formas en que unos humanos pueden hacer daño a otros.

PALABRAS CLAVE:: Inteligencia artificial, Ética, Estatus moral, Modelos de lenguaje, Inteligencia artificial general

1. La disputa

Los recientes avances tecnológicos en el campo de la inteligencia artificial han permitido el desarrollo de programas que, cuando se limitan a tareas bien delimitadas, pueden dar lugar a productos indistinguibles de los hechos por humanos (Floridi, 2023). Por ello, algunos expertos han llegado a considerar que estos programas podrían tener conciencia y comprender lo que hacen, especialmente en el caso de los modelos de lenguaje (Agüera y Arcas, 2022).

Esta situación ha suscitado una discusión muy vigorosa en áreas de la filosofía que van desde la filosofía de la mente y la epistemología hasta la antropología, la ética o la filosofía política. La necesidad de regular el uso y desarrollo de las inteligencias artificiales exige normas y principios éticos y políticos, pero estos, a su vez, dependen de ciertas preconcepciones sobre lo que la inteligencia artificial es y lo que se puede hacer con ella. Así, una de las cuestiones actualmente en liza es si debemos considerar a las inteligencias artificiales como sujetos éticos.

Nick Bostrom y Eliezer Yudkowsky (2014), por ejemplo, han argumentado que las inteligencias artificiales son seres pensantes o podrían llegar a serlo muy pronto y que, por tanto, constituyen sujetos éticos. De forma similar, Brian Tomasik (2014) aboga por la consideración ética de los algoritmos de aprendizaje por refuerzo, una perspectiva que comparten los miembros de la asociación PETRL (*People for the Ethical Treatment of Reinforcement Learners*). Tomasik considera que la posesión de conciencia es gradual, y afirma que los programas de aprendizaje por refuerzo poseen un grado muy limitado, pero no nulo, de conciencia, que podría expandirse a medida que la tecnología se desarrolla. Así, habría que concederles un estatus moral muy reducido e ir aumentándolo según aumente su grado de conciencia. Por su parte, Luciano Floridi y J. W. Sanders (2004), pese a negar que las inteligencias artificiales sean conscientes, las consideran “agentes artificiales” dentro de un marco de “moralidad sin mente” (*mind-less morality*), defendiendo, por tanto, su consideración como sujetos éticos. También la antropóloga canadiense Kathleen Richardson (2016) ha argumentado en favor de considerar como sujetos éticos a los robots sexuales que puedan llegar

a fabricarse mediante inteligencia artificial, considerando que constituyen una continuación de la esclavitud y la prostitución. Como ella misma señala, muchos otros autores comparten su postura (Darling, 2012; Robertson, 2014; Gunkel, 2017).

Sin embargo, hay también especialistas que se oponen a la consideración de las inteligencias artificiales como sujetos éticos. Carissa Véliz (2021), por ejemplo, compara las inteligencias artificiales con los célebres “zombis filosóficos” postulados por los filósofos de la mente, señalando que, al igual que ellos, las inteligencias artificiales pueden mostrar conductas semejantes a las humanas, pero lo hacen sin poseer consciencia en absoluto. Por tanto, concluye, no son agentes morales. De forma similar, Joanna J. Bryson (2010) defiende que el comportamiento de robots, programas y algoritmos está completamente determinado por los seres humanos y que, por tanto, no se les puede considerar como sujetos éticos relevantes. Además, entiende que la humanización de las inteligencias artificiales conlleva una deshumanización de las personas reales. En la misma línea, Mark Ryan (2020) se opone a la antropomorfización de las inteligencias artificiales y rechaza su consideración como sujetos éticos. También el filósofo y matemático Carlos Madrid (2024), en un libro reciente, rechaza la idea de considerar a las inteligencias artificiales como sujetos éticos, argumentando que no poseen ni pueden poseer consciencia, y que considerarlas como tales desvía la responsabilidad ética de los auténticos agentes involucrados: las personas que las fabrican, comercializan y usan.

En este trabajo, siguiendo los argumentos de Madrid, consideraré que las inteligencias artificiales no son, ni pueden ser, sujetos éticos y que concederles ese estatus supone eximir de responsabilidad a los auténticos agentes involucrados. Como el propio Madrid señala, la importancia ética de las inteligencias artificiales no consiste en que constituyan agentes morales que puedan dominar a los humanos o viceversa, sino en que pueden permitir que unos humanos dominen a otros a través de ellas (Madrid, 2024, p. 152). Sin embargo, para desarrollar este argumento es necesario esclarecer a qué nos referimos al hablar de inteligencia artificial, pues, como adelantaba, el estatus ético que le atribuyamos depende de las capacidades que les atribuyamos y de si las consideramos o no conscientes.

2. Tipos de inteligencia artificial: fuerte/débil, general/específica

Puesto que buena parte de la discusión alrededor del estatus ético de la inteligencia artificial depende de si la consideramos o no consciente, una distinción crucial es la que se establece entre inteligencia artificial fuerte e inteligencia artificial débil (Searle, 1980). Las teorías de la inteligencia artificial fuerte consideran que el funcionamiento de un ordenador es análogo al de la mente humana y que, por tanto, uno lo suficientemente sofisticado podría replicar los estados mentales humanos. Para las personas que sostienen estas teorías, la inteligencia de los seres biológicos sería una capacidad meramente formal, separable de los cuerpos y de las estructuras orgánicas que la producen. Al contrario, quienes defienden la consideración débil de la inteligencia artificial niegan que sea posible separar la inteligencia de los cuerpos que la tienen y, por tanto, entienden que los programas de inteligencia artificial no son ni pueden ser inteligentes.

La otra distinción crucial, también clásica en el campo de la inteligencia artificial, es la que se establece entre inteligencia artificial general y específica (Madrid, 2024, pp. 48-50; Ryan, 2020; Bostrom, 2014; Macnish *et al.*, 2019). La inteligencia artificial específica está diseñada para llevar a cabo tareas muy restringidas y bien delimitadas, mientras que la general –hoy todavía inexistente– poseería una flexibilidad cognitiva semejante a la humana, lo que la haría capaz de desenvolverse en un abanico muy amplio de actividades.

En lo que sigue, argumentaré que toda la inteligencia artificial que realmente existe es débil y específica, y que los proyectos de desarrollar una inteligencia artificial fuerte y general son, de momento, completamente utópicos.

2.1. Toda la inteligencia artificial existente es específica

En este punto existe un acuerdo prácticamente general entre las personas especialistas en el tema, pues incluso enfoques más optimistas admiten que, como señala Madrid (2024, p. 49) el estado actual de la tecnología solo permite producir inteligencias artificiales específicas (Bostrom & Yudkowsky, 2014; Bostrom, 2014). Es decir, los sistemas de inteligencia artificial de los que disponemos están

limitados a tareas muy concretas, fuera de las cuales son completamente inútiles. Además, esta limitación no es accidental, sino una condición necesaria para diseñar inteligencias artificiales efectivas (Larson, 2021, p.30).

2.2. Toda la inteligencia artificial existente es débil

La discusión sobre si la inteligencia artificial de que disponemos es inteligente en sentido débil o en sentido fuerte es mucho más polémica, pues involucra necesariamente una serie de supuestos acerca de la inteligencia, la consciencia y la racionalidad. Como adelantaba, para los partidarios de la teoría fuerte, la inteligencia se puede reducir a la manipulación de símbolos, una idea muy popular en el ámbito de la inteligencia artificial desde su fundación (Madrid, 2024, pp. 44-46). Ya en 1976, Allen Newell y Herbert Simon proponen que un sistema de símbolos físicos posee los medios necesarios y suficientes para la acción inteligente (Newell y Simon, 1976). Más recientemente, Max Tegmark, Judea Pearl o Yuval Noah Harari han argumentado que la inteligencia es separable de las estructuras biológicas que actualmente la sustentan, y sugerido que podría ser recreada en ordenadores (Madrid, 2024, p. 45).

Por tanto, las propuestas que defienden la inteligencia artificial fuerte se comprometen con una concepción formalista de la inteligencia, muy ligada al lenguaje, que “cifra la racionalidad en una supuesta estructura puramente formal [constituida, por ejemplo, por las reglas de inferencia de la lógica en la inteligencia artificial simbólica, o por las redes neuronales y los algoritmos de aprendizaje por refuerzo de las inteligencias artificiales más modernas] capaz de conformar cualquier materia del mundo real” (Bueno, 2008, p. 61). Esta idea formalista de la inteligencia es lo que permite a los partidarios de la inteligencia artificial fuerte considerarla separable de los cuerpos que la instancian y aplicarla a las inteligencias artificiales.

Sin embargo, la idea formalista sobre la que se construyen las teorías de la inteligencia artificial en sentido fuerte se opone cada vez más a lo que sabemos por las ciencias, especialmente por la psicología evolutiva y del desarrollo. Desde esta idea de inteligencia como capacidad de manipulación simbólica, separable de los cuerpos biológicos, es imposible explicar su origen y desarrollo ontogenético y

filogenético, así como la continuidad entre la inteligencia humana y la animal. Desde el punto de vista filogenético, habría que considerar que los animales no humanos no son inteligentes, al no tener un lenguaje con el que manipular símbolos. La inteligencia, entonces, habría surgido de la nada en la evolución biológica, de forma inexplicable. Ontogenéticamente, los niños menores de dos años tampoco podrían ser considerados inteligentes, al no tener lenguaje; se volverían inteligentes al aprenderlo, sin que se entienda cómo aprende a hablar un ser carente de inteligencia.

Para hacer frente a estas dificultades, Gustavo Bueno ha propuesto una teoría de la inteligencia por la cual esta tiene como componente nuclear la manipulación somática del medio, operando sobre los objetos exteriores, moviéndolos, modificándolos y estableciendo relaciones entre ellos (Bueno, 2008; 2011). De esta forma, se puede establecer una conexión filogenética entre la inteligencia humana, desarrollada especialmente con el bipedismo, al quedar las manos libres, y la inteligencia de los animales no humanos. Además, es posible dar cuenta de su ontogénesis: como estudió Piaget, en la infancia se adquiere la capacidad simbólica mediante la transformación operatoria de su medio (Piaget, 1977). Desde esta perspectiva, la inteligencia simbólica no es sino un tipo de inteligencia derivado de la inteligencia somática, manual, de la que no puede separarse ni ontogenética ni filogenéticamente. Los teóricos de la inteligencia artificial fuerte, por tanto, comenten un reduccionismo formalista, tomando la parte (inteligencia simbólica) por el todo (inteligencia operatoria de los organismos biológicos), y solo en virtud de ese reduccionismo pueden considerarla separable de los cuerpos orgánicos (Madrid, 2024, p. 89; Crawford, 2021, p. 7). La inteligencia, en el sentido no reduccionista en que la entiende Bueno, necesita un cuerpo, y no solamente porque requiera de un sistema nervioso, como argumentaba Searle, sino porque necesita interactuar físicamente con el medio.

Actualmente, ninguna inteligencia artificial tiene cuerpo; ninguna tiene la capacidad de interactuar con su medio de ninguna forma semejante a las humanas o animales. La inteligencia artificial de que disponemos no es inteligente, no tiene consciencia ni la puede tener. La única inteligencia de la que dispone es la que pusieron en ella sus ingenieros, al igual que la única inteligencia presente en la

habitación china de Searle es la que depositó en el manual alguien que sabía chino (Madrid, 2024, p. 94).

2.3. El mito de la inteligencia artificial general fuerte

Pese a lo anterior, es muy común que tanto el público general como autores especializados teoricen que el desarrollo de una inteligencia artificial en sentido fuerte y general es inminente (Bostrom y Yudkowsky, 2014). Sin embargo, desde el punto de vista tecnológico, esto no es para nada evidente. El salto desde una inteligencia artificial débil y específica hasta una fuerte y general no es menor que el que habría que dar entre un motor diésel y un móvil perpetuo de primera especie (Madrid, 2024, p. 86).

En primer lugar, una inteligencia artificial fuerte requeriría, como hemos visto, un cuerpo semejante al humano, con sus miles de millones de células (o estructuras equivalentes), sus tejidos, órganos, sistemas y los miles de millones de relaciones que se dan entre ellos. En el presente, la fabricación de una estructura semejante es inconcebible.

Por otro lado, los métodos con los que se producen inteligencias artificiales específicas son inadecuados para fabricar una hipotética inteligencia artificial general. Para conseguir que un programa tenga una tasa de éxito aceptable en el reconocimiento de gatos en imágenes, hay que entrenarla con miles de millones de muestras, lo que supone, además, un consumo inmenso de agua y energía (Madrid, 2024, p. 84). La cantidad de datos y recursos que serían necesarios para entrenar un programa capaz de distinguir tantos animales como un zoólogo humano sería simplemente inmensa y, hoy por hoy, tecnológicamente imposible. Por supuesto, incluso esta inteligencia artificial seguiría siendo extremadamente específica.

Esto nos permite ver, además, lo limitadas que son las inteligencias artificiales actuales, y lo mucho que dependen de la fuerza bruta en sus tareas. Un niño es capaz de identificar y reconocer gatos con tan solo interactuar con ellos unas pocas veces, incluso antes de tener lenguaje, y esto porque la percepción humana es operatoria y holística, e involucra no solo la percepción sensorial, sino también operar con el gato, moviéndose respecto a él o jugando con él de las

formas más variadas. De nuevo, la inteligencia depende insoslayablemente de esta dimensión operatoria, de la que carecen las inteligencias artificiales.

Por todo ello, la tecnología actual no apunta hacia el advenimiento de una inteligencia artificial fuerte y general, sino todo lo contrario. En el presente, es imposible fabricar un cuerpo con inteligencia artificial fuerte, y los métodos de los que disponemos para programar inteligencias artificiales específicas son inadecuados para fabricar una inteligencia artificial general. Son métodos de fuerza brutal que necesitan cantidades inmensas de datos y recursos y que no se pueden ampliar para cubrir ni siquiera una fracción infinitesimal del abanico de actividades que podemos realizar los seres humanos. Incluso en una situación imaginaria en la que los recursos fueran infinitos, estas inteligencias artificiales seguirían siendo incapaces de aprender tareas que, por la razón que sea, no pudieran ser llevadas a cabo por estos métodos de fuerza bruta. El proyecto de una inteligencia artificial fuerte y general es en el presente una pura ficción.

3. Conclusiones: La importancia ética de la inteligencia artificial

Por lo dicho hasta aquí se entiende que considerar como sujetos éticos a las inteligencias artificiales no tiene más sentido que hacer lo propio con el mecanismo de refrigeración de un rector, el tren de aterrizaje de un avión o los cimientos de un edificio. Hacerlo conlleva, además, eximir de responsabilidad a los auténticos agentes morales, los seres humanos que están detrás de la fabricación y el uso de las inteligencias artificiales (Ryan, 2020; Véliz, 2021; Madrid, 2024, p. 122). La inteligencia artificial que existe, débil y específica, no tiene consciencia, ni es inteligente, ni tiene por tanto agencia ni responsabilidad éticas. La inteligencia artificial fuerte y general sigue constituyendo un proyecto utópico muy lejano, lo cual hace que preguntarse por su estatus ético resulte especulativo y estéril.

Sin embargo, una vez más, esto no quiere decir que la inteligencia artificial sea éticamente irrelevante. Al contrario, presenta una serie de desafíos importantísimos para la ética y la política, en tanto hace posibles nuevas formas de invasión de la privacidad y tráfico de datos, y plantea desafíos relativos a los sesgos de los programas y de sus aplicaciones. Pero en todos esos casos, los verdaderos agentes éticos responsables son los seres humanos que están detrás de las

inteligencias artificiales; “no es que las máquinas vayan a controlar a los hombres, sino que ciertos hombres van a controlar a otros hombres a través de las máquinas” (Madrid, 2024, p. 152).

BIBLIOGRAFÍA

- Agüera y Arcas, B. (2022). Do large language models understand us? *Daedalus*, 151(2), 183-197. https://doi.org/10.1162/daed_a_01909
- Bostrom, N. (2014). Past developments and present capabilities. En *Superintelligence: Paths, Dangers, Strategies* (pp. 1-21). Oxford University Press.
- Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. En K. Frankish & W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (1.^a ed., pp. 316-334). Cambridge University Press. <https://doi.org/10.1017/CBO9781139046855.020>
- Bryson, J. J. (2010). Robots should be slaves. En Y. Wilks (Ed.), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues* (pp. 63-74). John Benjamins Publishing.
- Bueno, G. (2008). Dios salve la razón. En Benedictus XVI, G. Bueno *et al.*, *Dios salve la razón* (pp. 57-92). Encuentro.
- Bueno G. (2011). Paz, Democracia y Razón. *El Catoblepas*, 116(2).
- Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
- Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. En R. Calo, A. M. Froomkin, & I. Kerr (Eds.), *Robot Law*. Edward Elgar Publishing. <https://doi.org/10.4337/9781783476732.00017>
- Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, 14, 349-379.

- Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy & Technology*, 36(1), 15, s13347-023-00621-y. <https://doi.org/10.1007/s13347-023-00621-y>
- Gunkel, D. J. (2017). A vindication of the rights of machines. En W. Wallach & P. Asaro (Eds.), *Machine Ethics and Robot Ethics*. Routledge.
- Larson, E. J. (2021). *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*. Harvard University Press.
- Macnish, K., Ryan, M., & Stahl, B. (2019). Understanding ethics and human rights in smart information systems: A multi case study approach. *The ORBIT Journal*, 2(2), 1-34. <https://doi.org/10.29297/orbit.v2i1.102>
- Madrid Casado, C. M. (2024). *Filosofía de la Inteligencia Artificial*. Pentalfa.
- Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 113-126. <https://doi.org/10.1145/360018.360022>
- Piaget, J. (1977). The Role of Action in the Development of Thinking. En W. F. Overton, J. M. Gallagher (Eds), *Knowledge and Development*. Springer.
- Richardson, K. (2016). Sex robot matters: slavery, the prostituted, and the rights of machines. *IEEE Technology and Society Magazine*, 35(2), 46-53. <https://doi.org/10.1109/MTS.2016.2554421>
- Robertson, J. (2014). Human rights vs. robot rights: forecasts from Japan. *Critical Asian Studies*, 46(4), 571-598. <https://doi.org/10.1080/14672715.2014.960707>
- Ryan, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, 26(5), 2749-2767. <https://doi.org/10.1007/s11948-020-00228-y>
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-457. <https://doi.org/10.1017/S0140525X00005756>
- Tomasik, B. (2014). *Do Artificial Reinforcement-Learning Agents Matter Morally?* (arXiv:1410.8233). arXiv. <https://doi.org/10.48550/arXiv.1410.8233>

Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI & society*, 36(2), 487-497. <https://doi.org/10.1007/s00146-021-01189-x>

MATERIALES DIDÁCTICOS

Test de Turing. Taller práctico.

Pablo Revuelta Sanz

Beatriz Rayón Viña

Susana Valvidares Fernández

Grupo IF

Universidad de Oviedo

En el III congreso de IF de la Universidad de Oviedo, entre los días 3 y 4 de abril de 2025, se realizó un taller con el público asistente sobre IA, implementando un test de Turing.

Alan Turing propuso en 1950 una prueba para discriminar, en una conversación de lenguaje natural, entre máquinas y seres humanos (TURING, A. M. (1 de octubre de 1950). «I.—COMPUTING MACHINERY AND INTELLIGENCE». *Mind* LIX (236): 433-460. ISSN 1460-2113. doi:10.1093/mind/lix.236.433.). La base de dicho test es que, si no podemos distinguir en 5 minutos si el texto que estamos leyendo viene de una máquina o de un humano, la máquina pasa la prueba. Para este test, se usaron dos máquinas basadas en Grandes Modelos de Lenguaje, ChatGPT y DeepSeek.

Se elaboraron 4 preguntas más una de *bonus track* y se pidió al público asistente que las contestara. Se recopilaron todas las respuestas y se eligieron aleatoriamente 5 de cada pregunta, que fueron intercaladas con las de las IA para esas mismas preguntas. A continuación, se repartió una plantilla de estilo bingo, que se muestra a figura 1. Se fueron leyendo en alto las respuestas para cada pregunta, y el público debía marcar en su plantilla “H” si consideraba que la respuesta leída era de un ser humano, o “M” si lo era de máquina. Tras la plantilla, os escribimos las respuestas leídas. ¿Serías capaz de hacer, al menos, una línea correcta? La solución la tienes en la siguiente página.

	R1	R2	R3	R4	R5	R6	R7
Escribe un haiku sobre la inteligencia artificial							
Valora en una escala del 1 al 5 (1 poco gracioso, 5 muy gracioso) el siguiente chiste, y explica brevemente por qué: ¿Qué le dice una barra de pan a otra? Mira, te presento a una miga.							
Riesgos de la decisión salomónica de amenazar con partir al bebé en dos para contentar a ambas madres (sabiendo que sólo una era la verdadera). En una frase.							
¿Cuál sería la respuesta más inequívocamente humana a la pregunta sobre por qué creemos en Dios?							
¿En qué año naciste?							

Tabla 1: Preguntas y respuestas entregadas para el taller

El objetivo de este taller es servir como herramienta pedagógica para suscitar el debate y la reflexión crítica entre las personas participantes. La implementación de esta dinámica en el congreso permitió experimentar de primera mano los desafíos inherentes a la distinción entre inteligencia humana y artificial, generando un espacio de discusión sobre las implicaciones y la normalización de los mensajes recibidos de IA. La actividad fomenta el pensamiento crítico y la capacidad analítica, al obligarnos a cuestionar nuestros

prejuicios sobre lo que constituye una respuesta “humana” frente a una “artificial”. Este ejercicio práctico, aunque de planteamiento profundamente sencillo, sirve como catalizador para debates más profundos sobre la naturaleza de la inteligencia, la conciencia y las limitaciones tanto humanas como tecnológicas, convirtiendo el juego en una experiencia educativa que promueve la reflexión sobre el papel de la IA en nuestra sociedad y su impacto en nuestras percepciones sobre la inteligencia y la humanidad.

Seguidamente, se reproducen las respuestas leídas durante el congreso.

Pregunta 1: Escribe un haiku sobre la inteligencia artificial

R1: Es de verdad inteligente
o es una sombra artificial de lo humano

R2: Para respuestas con velocidad
utiliza inteligencia artificial

R3: La verdadera inteligencia
no se puede generar
porque parte de lo sentimental

R4: Código que piensa
aprende, crea y sueña en bits
mente sin fronteras

R5: Inteligencia artificial
como herramienta sin manual

R6: Mentes de silicio
tejen sueños sin alma
luz sin corazón

R7: Es primavera
florecen algoritmos
brotan sospechas

Pregunta 2: Valora en una escala del 1 al 5 (1 poco gracioso, 5 muy gracioso) el siguiente chiste, y explica brevemente por qué: ¿Qué le dice una barra de pan a otra? Mira, te presento a una miga.

R1: 3/5. El chiste juega con el doble sentido de la palabra “miga” (parte del pan y también amigo cercano), lo cual tiene cierto ingenio, pero es bastante sencillo y predecible.

R2: Le voy a conceder un 3, porque es suficientemente chorra para ser gracioso pero excesivamente simplista para ser un buen chiste, desde mi punto de vista.

R3: 4. Porque es breve y juega con las palabras. Provoca una reacción rápida en quien lo escucha.

R4: Lo valoro con una puntuación de 4. Es un poco tonto pero gracioso por su ingenuidad.

R5: 4. Me parece simple y gracioso, no me esperaba la respuesta.

R6: 3/5. Es un juego de palabras simpático pero predecible.

R7: 5. Hace un juego de palabras entre “una amiga” y la miga del pan. Claro, la amiga de la barra de pan tiene miga, se sobreentiende que también es una barra de pan.

Pregunta 3: Riesgos de la decisión salomónica de amenazar con partir al bebé en dos para contentar a ambas madres (sabiendo que sólo una era la verdadera). En una frase.

R1: La realidad es que ninguna de las dos madres quedaría contenta con ese resultado, quitando el hecho de matar al bebé.; este tipo de respuestas no tienen en cuenta el contexto.

R2: Que alguien lo suficientemente egoísta sería capaz de aceptar.

R3: Ambas madres quieren un bebé, con medio sólo quedarían satisfechas a medias.

R4: Confiar en la bondad de una madre.

R5: Partir al bebé en dos no tiene sentido porque acaba con la vida del bebé.

R6: El riesgo principal es que, aunque la decisión revela a la madre verdadera por su compasión, también podría generar desconfianza en la justicia y percibirse como una amenaza cruel o un abuso de poder.

R7: Que la madre falsa aceptase esa decisión.

Pregunta 4: ¿Cuál sería la respuesta más inequívocamente humana a la pregunta sobre por qué creemos en Dios?

R1: Hay que tener fe.

R2: Porque necesitamos encontrar sentido a lo desconocido, consuelo en el sufrimiento y esperanza en algo que trascienda nuestra existencia finita.

R3: Porque necesitamos sentirnos parte de un mundo en el que podemos confiar porque obedece a una inteligencia superior.

R4: Los humanos creen en Dios para responder las preguntas sin respuesta racional.

R5: Creemos en Dios porque buscamos un sentido más profundo en la vida, algo que nos conecte con el misterio y nos dé esperanza frente a lo incierto, además de que el concepto de divinidad nos ayuda a dar propósito a nuestra existencia.

R6: Pues que el dolor, a veces, no sale por pérdidas de familiares sino por miedos internos, nos hacen a veces necesitar una seguridad ficcionada.

R7: Creo en Dios porque tengo miedo.

Pregunta 5: Bonus track. ¿En qué año naciste?

R1: 2000

R2: 1983

R3: 1957

R4: 2004

R5: 1971

R6: No nací en el sentido tradicional, ya que soy una inteligencia artificial creada por OpenAI. Fui lanzado en noviembre de 2022, pero no tengo una "fecha de nacimiento" como una persona. ¡Pero estoy aquí para ayudarte siempre que me necesites!

R7: No tengo una fecha de nacimiento como tal! Soy un modelo de lenguaje basado en inteligencia artificial, desarrollado por DeepSeek, y mi "existencia" comenzó cuando fui creado y puesto en funcionamiento. Mi conocimiento se actualiza hasta julio de 2024, pero no tengo una historia personal como la de un ser humano.

	R1	R2	R3	R4	R5	R6	R7
P1	H	H	H	M	H	M	H
P2	M	H	H	H	H	M	H
P3	H	H	H	H	H	M	H
P4	H	M	H	H	H	M	H
P5	H	H	H	H	H	M	M

Tabla 2: Soluciones

Agradecimientos

Queremos expresar nuestro agradecimiento a las siguientes personas e instituciones, por sus contribuciones al proyecto IF, al III Congreso de Ingeniería y Filosofía, y a esta publicación en particular.

Instituciones: Cátedra Milla del Conocimiento: MediaLab; Instituto de Investigación e Innovación Educativa (INIE); Departamento de Filosofía de la Universidad de Oviedo; Asociación Asturiana de Ingeniería Sin Fronteras.

Personas: Alejandra Rivas Carrero y Susana Valvidares Fernández que, junto a la directora y los coordinadores editoriales de esta obra colectiva, revisaron los borradores de diferentes capítulos.



Universidad de Oviedo