Manual Imprescindible

Guía de IA Responsable

Enfoque práctico para el cumplimiento y ética de la IA

Adrián González Sánchez
Juan Manuel Belloto Gómez
Alfonso Peralta Gutiérrez

Prólogos de Rodrigo Miranda Beltrán, Juan Bernabé-Moreno y Esther Checa Gutiérrez
Epílogos de Mario Hernández Ramos y Alberto Gago Fernández

ANAYA MULTIMEDIA

Guía de IA responsable

Enfoque práctico para el cumplimiento y ética de la IA

Manual Imprescindible

Guía de IA responsable

Enfoque práctico para el cumplimiento y ética de la IA

Adrián González Sánchez,
Juan Manuel Belloto Gómez,
Alfonso Peralta Gutiérrez

Primera edición: octubre de 2026

Realización de cubierta: Celia Antón Santos

Imagen de cubierta: © 2004-2026 Shutterstock, Inc.

Reservados todos los derechos. El contenido de esta obra está protegido por la Ley, que establece penas de prisión y/o multas, además de las correspondientes indemnizaciones por daños y perjuicios, para quienes reprodujeren, plagiaren, distribuyeren o comunicaren públicamente, en todo o en parte, una obra literaria, artística o científica, o su transformación, interpretación o ejecución artística fijada en cualquier tipo de soporte o comunicada a través de cualquier medio, sin la preceptiva autorización.

© 2026 Adrián González Sánchez, Juan Manuel Belloto Gómez, Alfonso Peralta Gutiérrez

© 2026 Rodrigo Miranda Beltrán (prólogo)

© 2026 Juan Bernabé-Moreno (prólogo)

© 2026 Esther Checa Gutiérrez (prólogo)

© 2026 Mario Hernández Ramos (epílogo)

© 2026 Alberto Gago Fernández (epílogo)

© EDICIONES ANAYA MULTIMEDIA (GRUPO ANAYA, S.A.U.), 2026

Valentín Beato, 21. 28037 Madrid

www.anayamultimedia.es

Depósito legal: M-10661-2026

ISBN: 978-84-415-5281-4

Impreso en España - Printed in Spain

A mi familia por su apoyo constante y a mis coautores
por compartir esta nueva aventura.

—Adrián

A mis padres, por sus enseñanzas y apoyo durante toda su vida.
A mi madre, por su afán de superación tras la partida de mi padre
y por su presencia durante muchos años más.
A Arancha, por su amor, y por ser una increíble compañera de viaje. También
por su comprensión, aunque me haya pedido que este sea mi último libro :).
A Hugo, por ser un ejemplo de madurez.

—Juanma

A mi mujer y a mi hijo.
Por su apoyo, ayuda y porque el tiempo que le dedico a muchos proyectos
se lo acabo robando a ellos y al sueño.

—Alfonso

gracias

Gracias una vez más al equipo de ANAYA por la confianza. Eugenio Tuya, Claudia Valdés-Miranda, Celia Antón y todas las personas que nos han ayudado a convertir nuestras ideas y textos en este libro. Gracias también a nuestros autores de prólogos y epílogos por el apoyo, y a todos los lectores por confiar en nosotros como parte de su proceso de aprendizaje y adopción de enfoques de IA responsable.

—ADRIÁN

Gracias a todos mis compañeros de viaje desde 2018 (empresas, entidades públicas y profesionales independientes), cuando decidí dedicarme a la IA dando un salto sin red, porque gracias a ellos la aventura está superando mis expectativas.

Especialmente a los Peaky Blinders de la IA: Adrián González Sánchez, Paul Van Branteghem y Henner Truchsess, con quienes he tenido la suerte de compartir muchos proyectos increíbles en estos años.

A Enrique Marco Martín y Cristina Sierra Romero de Campus AENOR por apostar por nosotros en su nuevo camino en el ámbito de la inteligencia artificial, y a Jose Ángel Valderrama Antón de AENOR por su apoyo. A ti, lector/a, por considerar este libro para tu aprendizaje.

—JUANMA

De nuevo a mi mujer y a mi hijo por aguantarme. A mis padres, por darme todo lo que han podido y enseñarme de la mejor manera. Espero que mi hijo pueda estar tan orgulloso como yo de ellos. A Juanma y Adrián por confiar en mí en este proyecto y a Anaya por hacerlo realidad. A los autores de los prólogos y epílogos por dedicar un poco de tiempo de su tan apretada agenda de manera desinteresada para añadir valor a este proyecto. Los hemos elegido porque para nosotros son algunos de los mayores referentes españoles en la materia. Gracias a todos.

—ALFONSO

Adrián González Sánchez

SOBRE EL AUTOR

Es un profesional de la inteligencia artificial con experiencia en diferentes empresas del sector, entre ellas Microsoft AI, donde ocupa el puesto de gestor de producto de IA y búsqueda web con las API de Bing.

Actualmente reside en España, ha desarrollado su carrera principalmente en Francia, Argentina y Canadá. Además, es director académico de dos másteres en IE University y profesor en universidades internacionales como la escuela ejecutiva de HEC Montréal y Concordia University. También ejerce como consultor de inteligencia artificial y producto en Toptal.

Con experiencia profesional como arquitecto de IA, desarrollador de software, consultor de telecomunicaciones, gestor de proyecto y producto, y especialista en la nube, decidió emprender un camino paralelo en los ámbitos académicos y de la divulgación de conocimiento.

Cuenta en su haber con varias publicaciones técnicas, incluidos dos libros publicados por ANAYA Multimedia (*Manual Imprescindible: Inteligencia artificial. Guía definitiva desde los orígenes a la IA generativa* y el presente, *Guía de IA responsable*), tres títulos de Marcombo Ediciones, ocho libros e informes de O'Reilly Media, y una nueva serie de libros con Packt Publishing.

Colabora con LinkedIn Learning, DeepLearning.AI y The Linux Foundation entre otras organizaciones, en el desarrollo de cursos en línea. También forma parte del grupo de expertos del Sandbox de IA del Gobierno de España y es auditor líder certificado del estándar ISO/IEC 42001.

anayamultimedia.es/autor/adrian-gonzalez-sanchez
www.linkedin.com/in/adriangonzalezsanchez
x.com/adriangs86

Juan Manuel Belloto Gómez

SOBRE EL AUTOR

Es consultor y asesor independiente sobre estrategia e innovación tecnológica aplicada a negocio.

Toda su carrera profesional ha estado ligada a la transformación digital y a la innovación desde que, a finales de los 90, fue cofundador de la primera agencia de viajes *online* española. Su trayectoria desde desarrollador y arquitecto de software hasta desempeñar funciones ejecutivas y directivas le permite abordar desde la definición de objetivos de negocio y la estrategia para conseguirlos, hasta dirigir la implantación de la solución tecnológica necesaria para lograrlos.

Su relación con la IA comenzó en 2015, con la implantación de motores de recomendación en canales digitales y el uso de tecnologías de voz. En 2018 decidió focalizar su futuro profesional en la inteligencia artificial, asesorando y colaborando desde entonces con *startups*, pymes, grandes compañías multinacionales (entre ellas varias Big Four) y la Secretaría de Estado de Digitalización e Inteligencia Artificial del Gobierno de España (SEDIA), abarcando con todas ellas dimensiones estratégicas, de gobierno, de gestión de riesgos, legales, tecnológicas y de normas ISO. Además, es vocal de la red de expertos sobre inteligencia artificial de la UNE (Asociación Española de Normalización).

En el ámbito académico y divulgativo ha colaborado con universidades y en la actualidad es formador en varias instituciones entre las que destacan Campus AENOR y Bespoke School of Finance. Es coautor de cinco libros sobre inteligencia artificial publicados por Tirant lo Blanch y ANAYA Multimedia (*Manual Imprescindible: Inteligencia artificial. Guía definitiva desde los orígenes a la IA generativa* y el presente libro).

www.linkedin.com/in/juanmabg

Alfonso Peralta Gutiérrez

SOBRE EL AUTOR

Magistrado presidente de la Sección Civil y de Instrucción del Tribunal de Instancia de Guadix, Plaza n.º 2. Antes ha sido juez en Roquetas de Mar n.º 1, Vera n.º 4 y Guadix n.º 1. Primer español y primer magistrado en formar parte del Comité Consultivo de Inteligencia Artificial (AIAB) de la Comisión Europea para la Eficacia de la Justicia (CEPEJ) del Consejo de Europa.

Experto del Comité Europeo de Problemas de Criminalidad (CDPC) 2024-2027 y de su Grupo de Trabajo de Expertos en Inteligencia Artificial y Derecho Penal (CDPC-AICL) en la redacción de un instrumento del Consejo de Europa sobre IA y derecho penal. Único consultor español de la UNESCO en materia de inteligencia artificial y Estado de derecho. Experto en EL PACCTO 2.0 (Programa de la Comisión Europea de Asociación UE-América Latina y el Caribe en materia de justicia y seguridad) sobre IA y crimen. Miembro designado por el CGPJ del Comité de Digitalización de Cooperación Internacional de la Dirección General de Justicia y Consumidores de la Comisión Europea.

Autor del libro *La transformación de la justicia a través de la inteligencia artificial: análisis comparado mundial y 40 propuestas rupturistas*, editado por Aranzadi La Ley. Codirector desde 2020 del Plan de Formación Continua del CGPJ en distintos cursos sobre derecho e inteligencia artificial y cocoordinador del libro homónimo publicado en colaboración con la Universidad de Granada. Codirector del «Programa Superior en Derecho Digital: Protección de Datos, IA y Transformación Digital» de ediae y el Centro Mediterráneo de la Universidad de Granada. Doctorando en Derecho en la Universidad de Granada sobre la «Tecnocensura de los algoritmos a la libertad de expresión». Ha impartido ponencias en múltiples foros, es profesor de distintos másteres y ha escrito más de 50 publicaciones sobre IA y derecho, ética de la IA y marco normativo de la IA, entre otras materias.

Medalla al mérito policial de la ciudad de Guadix —por las operaciones realizadas en pro de la seguridad ciudadana y la convivencia— y miembro de proyectos de investigación de la Fundación COTEC, de la Universidad Complutense de Madrid, de la Universidad de Murcia, de la Universidad de Jaén y de la Universidad de Newcastle de Australia. Participa con artículos en los diarios *ABC*, *Confilegal* e *IDEAL* de Granada, así como en el programa *Código Cristal* de RNE 5.
www.linkedin.com/in/a-peralta/

Índice de contenidos

Prólogo de Rodrigo Miranda Beltrán 14
El nuevo contrato social entre humanos y máquinas 15

Prólogo de Juan Bernabé-Moreno 18
Comprender la revolución de la inteligencia artificial (responsable). 19

Prólogo de Esther Checa Gutiérrez 22
El mito de la neutralidad y el «hackeo» de la optimización opaca de los algoritmos 23

Introducción 26
Cuando la ética y la moral no bastan 27

1. La necesidad de la IA responsable en la era de la inteligencia artificial generativa 31

De la IA a la IA generativa y el impacto tecnológico en el valor de la ética 32
Vicios y virtudes de la IA (o cómo gestionar sus limitaciones) 34
La magnitud del impacto de la IA a nivel social 36
Confianza como ventaja competitiva y diferenciación en IA. 37
La transición de la experimentación en IA al despliegue a nivel empresarial 38
Casos y aprendizajes del pasado 39
COMPAS o cuando la libertad se convierte en una puntuación. 39
Amazon y el algoritmo que aprendió de un pasado desigual. 41
Los profesores de Florida o cuando medir no significa comprender. 42
Facebook y Cambridge Analytica o la privacidad como infraestructura democrática 43
El abogado neoyorquino que confió demasiado en ChatGPT. 44
La máquina expendedora de Anthropic o cuando un agente de IA descubre el mundo real 44
Microsoft Tay o el chatbot que aprendió lo peor de internet 45
Google Photos y el error que reveló sesgos en visión artificial 46
El algoritmo de las notas en Reino Unido o cuando el pasado de una escuela decide el futuro de un alumno 47
El escándalo neerlandés de las ayudas familiares o la automatización de la sospecha 48
Robodebt en Australia o el coste humano de una deuda calculada por promedios. 48
Reconocimiento facial y arrestos erróneos que convierten coincidencias en sospecha 49
Air Canada y el chatbot que prometió lo que la empresa no quería cumplir 50
Apple Card o cuando la percepción de injusticia también destruye confianza. 50
IBM Watson for Oncology o la dificultad de llevar la IA a la medicina. 51
Zillow Offers o cuando un modelo de predicción entra en el mercado real. 52
Clearview AI, la cara como dato y la privacidad como frontera. 53
Conclusiones a partir de patrones que se repiten con el tiempo. 53

2. Por qué es necesaria una IA responsable 57

Un enfoque pragmático . . . 58
Casos que ejemplifican la necesidad de una IA responsable . . . 58
La diferencia de la IA con otras tecnologías . . . 61
Las competencias necesarias en la IA . . . 62
Un factor diferencial: el volumen . . . 63
Iniciativas que ejemplifican la necesidad de una IA responsable . . . 64
Por qué es complejo conseguir una IA responsable . . . 65
Conclusión: la IA responsable no es opcional . . . 67

3. Dimensiones y herramientas para conseguir una IA responsable 69

Dimensión ética: para todos . . . 70
Dimensión regulatoria: solo para unos pocos . . . 73
Herramientas: para todos . . . 75
- Sistema de gestión . . . 76
- Gestión de riesgos . . . 76
- Controles . . . 79
- Medidas de diseño . . . 79
- Herramientas tecnológicas . . . 80

En resumen . . . 80

4. Dimensión ética 83

Principios de la Unión Europea . . . 84
- Supervisión humana . . . 87
- Solidez técnica y seguridad . . . 90
- Gestión de la privacidad y de los datos . . . 94
- Transparencia . . . 97
- Diversidad, no discriminación y equidad . . . 101
- Bienestar social y ambiental . . . 103
- Rendición de cuentas . . . 106
- Conclusiones . . . 109

Fundamentos técnicos detrás de la revolución de la IA generativa . . . 110
La aceleración de la adopción de la IA y el auge de la IA agéntica . . . 114
Riesgos generales y consideraciones clave . . . 114
Quién es quién en el tablero de la IA . . . 123
- Proveedores de infraestructura física . . . 124
- Hiperescaladores y proveedores de plataformas como servicio . . . 124
- Laboratorios frontera para el desarrollo de modelos . . . 125
- Comunidades y desarrolladores . . . 125
- Proveedores de datos y contenido . . . 126
- Mano de obra humana para la supervisión y mejora de la IA . . . 126
- Proveedores de servicios, integradores y consultoras . . . 127
- Entidades que adoptan las soluciones de IA . . . 127
- Usuarios finales de soluciones con IA . . . 128
- Reguladores y autoridades . . . 128
- Otros actores involucrados . . . 129
- Inversores en empresas que proveen y adoptan la IA . . . 129
- Actores que explotan las vulnerabilidades de la IA . . . 130

Enfoque ético aplicado a la IA generativa y agéntica . . . 130

5. Dimensión regulatoria 141

¿La regulación perjudica la innovación? Estrategia digital europea 142
Los problemas de competitividad europeos 142
¿La Europa del tapón de la botella? Estrategia digital para un despertar europeo 148
¿El problema es la regulación? 156
Soluciones al problema de competitividad europea 169
Contexto regulatorio global 183
China: Leyes de recomendación algorítmicas, «deep synthesis», servicios IAGen y servicios interactivos humanizados 185
Estados Unidos: Executive Order, Take it Down Act, normativas estatales y marco legislativo nacional 191
Convención del Consejo de Europa de IA 193
Ley italiana de IA. 195
España: Anteproyecto de Ley para el Buen Uso y la Gobernanza de la Inteligencia Artificial y modificación de la LO 1/82 197
Overview del RIA 207
Objeto del RIA 208
El porqué de su necesidad. 209
Estructura 209
Gobernanza y roles 213
Clasificación de los sistemas de IA 217
Detalles sobre los sistemas de alto riesgo. 230
Calendario de aplicación 237
Régimen sancionador 238
El RIA como facilitador de la IA responsable 239
Conclusiones 241
Relación del RIA con la IA ética 242
El origen: los derechos fundamentales de la Unión Europea 243
La presencia de los siete principios éticos en el RIA 246
Resumen 251
Aplicación del RIA a la IA generativa y agéntica 252
Aplicación del RIA a la IA generativa 252
Aplicación del RIA a la IA agéntica 258
Relación entre RIA y GDPR 262
Ómnibus Digital y Ómnibus de IA 278
Ómnibus Digital sobre datos, ciberseguridad y privacidad 279
Ómnibus de Inteligencia Artificial 286
Otras normativas que impactan en la IA 288
El impacto del DSA en los contratos de términos del servicio respecto a la moderación de contenidos online 288
La Directiva y propuesta de Reglamento CSA y Chatcontrol 304
Casos que han llegado a los tribunales 311
Algoritmos españoles y lo que llegará 311

6. Herramientas normativas. Gobernanza y gestión de riesgos 337
Sistema de gestión de la IA. ISO 42001 . . . 339
Introducción . . . 339
Cláusulas . . . 345
Controles . . . 361
Conclusiones . . . 364
Gestión de riesgos de la IA. ISO 23894 . . . 364
Qué es un proceso de gestión de riesgos . . . 364
Qué relación tiene con un sistema de gestión . . . 365
Qué objetivos tiene . . . 365
Qué beneficios proporciona . . . 365
Por qué es necesario . . . 366
Qué áreas están involucradas . . . 366
Metodologías y organizaciones . . . 367
Terminología . . . 368
Descripción general del proceso . . . 369
Objetivos a considerar en la gestión de riesgos . . . 371
Estándares para su implementación. ISO 31000 e ISO 23894 . . . 373
Relación entre la gestión de riesgos y el RIA . . . 374
Conclusiones . . . 374
Cómo integrar las normas . . . 375
Otras normas ISO de interés . . . 377
Alineamiento para la normalización de los reglamentos de IA . . . 383

7. Herramientas tecnológicas 387
Medidas de diseño para conseguir una IA responsable . . . 388
Herramientas para automatizar la IA responsable en empresa . . . 410
Integración de la IA responsable en las operaciones técnicas . . . 410
Flujos de trabajo para la gestión del riesgo de IA dentro de la empresa . . . 412
Model cards, datasheets y marcos de documentación . . . 414
Herramientas de detección de sesgos y pruebas de equidad . . . 416
Monitorización continua y detección de deriva del modelo . . . 417
Plataformas de automatización del gobierno de IA . . . 419
Gestión del riesgo de terceros y proveedores de IA . . . 421
Respuesta a incidentes y gestión de fallos de IA . . . 423
Formación y gestión del cambio . . . 425
Conclusión . . . 426

Epílogo de Mario Hernández Ramos 428
La IA responsable como imperativo humanista y democrático . . . 429

Epílogo de Alberto Gago Fernández 433
El valor económico de la IA responsable . . . 434
Innovación vs regulación . . . 434
AESIA – Agencia Española de Supervisión de Inteligencia Artificial . . . 435

Índice alfabético 437

Prólogo de

Rodrigo Miranda Beltrán

El nuevo contrato social entre humanos y máquinas

Cada gran revolución tecnológica ha redefinido las reglas de convivencia de su tiempo.

La máquina de vapor transformó el trabajo físico. La electricidad cambió la organización de las ciudades y las fábricas. Internet alteró la forma en que nos comunicamos, aprendemos y hacemos negocios.

La inteligencia artificial representa un salto diferente.

No se limita a amplificar nuestra fuerza, como hicieron las máquinas industriales. Tampoco únicamente amplía nuestra capacidad de comunicación, como hizo Internet. La inteligencia artificial comienza a intervenir en algo mucho más profundo: nuestra capacidad de pensar, decidir, crear y resolver problemas.

Por primera vez en la historia, millones de personas comparten parte de sus tareas cognitivas con sistemas capaces de generar contenido, analizar información, programar software, redactar informes, diseñar estrategias o tomar decisiones de apoyo en tiempo real.

Y eso nos sitúa ante una pregunta que trasciende la tecnología: ¿Cuál será el nuevo contrato social entre humanos y máquinas?

La respuesta no puede dejarse exclusivamente en manos de ingenieros, reguladores o grandes compañías tecnológicas. Nos afecta a todos:

- A las empresas que buscan mejorar su productividad.
- A los trabajadores que deben adaptarse a nuevas formas de empleo.
- A los gobiernos responsables de proteger la cohesión social.
- A las universidades y escuelas de negocio que forman a las futuras generaciones.

Y especialmente a aquellos colectivos que históricamente han sufrido con mayor intensidad los procesos de transformación económica.

Entre ellos, el talento sénior.

Durante décadas hemos asociado la innovación tecnológica a la juventud. Hemos construido un relato que, de forma explícita o implícita, identificaba experiencia con pasado y tecnología con futuro.

La realidad es bastante más compleja.

La inteligencia artificial generativa está demostrando que el valor diferencial de muchos profesionales ya no reside únicamente en la velocidad de ejecución o en el acceso a la

información. Cada vez cobra más importancia la capacidad de formular buenas preguntas, interpretar contextos, ejercer criterio, gestionar incertidumbre, comprender a las personas y tomar decisiones responsables.

Precisamente las capacidades que suelen fortalecerse con la experiencia.

Por eso creo que la inteligencia artificial puede convertirse en una oportunidad histórica para redefinir el papel del talento sénior en nuestras organizaciones.

Pero esa oportunidad no se materializará por sí sola.

Necesita liderazgo. Necesita inversión. Y, sobre todo, necesita diálogo social.

Durante demasiado tiempo hemos abordado las revoluciones tecnológicas desde posiciones enfrentadas. Empresas por un lado. Sindicatos por otro. Reguladores intentando seguir el ritmo de la innovación. Universidades formando profesionales para empleos que cambian más rápido que los planes de estudio.

La inteligencia artificial exige superar esa lógica.

No podemos permitirnos un debate basado únicamente en el miedo al reemplazo, ni tampoco un discurso ingenuo que ignore los riesgos reales asociados a esta tecnología.

Necesitamos una conversación madura entre empresas, representantes de los trabajadores, administraciones públicas, sistema educativo y sociedad civil.

Una conversación centrada en preguntas concretas:

- ¿Cómo redistribuiremos las ganancias de productividad?
- ¿Cómo acompañaremos a quienes necesiten recualificarse?
- ¿Cómo garantizaremos que nadie quede excluido por razones de edad?
- ¿Cómo mediremos el impacto real sobre el empleo?
- ¿Cómo aseguraremos la transparencia de los sistemas que influyen en las decisiones laborales?
- ¿Cómo preservaremos la dignidad humana en entornos cada vez más automatizados?

La respuesta a estas cuestiones determinará si la inteligencia artificial se convierte en un factor de progreso compartido o en una nueva fuente de desigualdad. Algunas de estas preguntas ya fueron planteadas por el papa León XIV en su encíclica *Magnifica Humanitas*. Adicionalmente, Europa posee una oportunidad singular para liderar este debate y, particularmente, España está en una posición muy relevante porque va a erigirse, en unos pocos años, en el primer país en esperanza de vida a nivel mundial.

Mientras otras regiones compiten por desarrollar modelos más potentes o por captar mayores inversiones, Europa puede aportar algo igualmente valioso: una visión centrada en las personas y España, una visión centrada en el valor del talento sénior.

Una visión que entienda que la innovación tecnológica y la protección de los derechos no son objetivos incompatibles. Que la competitividad y la inclusión deben avanzar juntas. Y que la responsabilidad no es un freno para la innovación, sino una condición necesaria para que la innovación genere confianza.

Porque la confianza será, probablemente, el activo más importante de la economía de la inteligencia artificial. Las organizaciones que sepan construirla atraerán talento. Las instituciones que sepan protegerla fortalecerán su legitimidad. Y las sociedades que sepan cultivarla afrontarán el futuro con mayores garantías de prosperidad.

Este libro aborda precisamente ese desafío.

No desde el alarmismo ni desde la fascinación tecnológica, sino desde la convicción de que la inteligencia artificial es demasiado importante para dejarla únicamente en manos de la tecnología. El verdadero reto no consiste en construir máquinas más inteligentes. Consiste en construir una sociedad más preparada para convivir con ellas.

Y en esa tarea, empresas, sindicatos, administraciones públicas, universidades y ciudadanos compartimos una responsabilidad común.

Porque el futuro del trabajo no lo decidirán los algoritmos.

Lo decidiremos nosotros.

—Rodrigo Miranda Beltrán
Socio ISDI Digital Talent, Senior Advisor EY. Patrono Ashoka y Teaming.
Doctorando en IA Generativa y empleabilidad sénior.
Autor de los libros *#MiYoDigital*, *#ReinicIAndo*, *Reiniciarte*

EN UN TUIT:

«La pregunta no es si la IA sustituirá empleos, sino si seremos capaces de construir los mecanismos de diálogo social, formación y adaptación necesarios para que amplifique el valor de las personas, especialmente de aquellas cuya experiencia constituye uno de los activos más valiosos de nuestras organizaciones».

Prólogo de

Juan Bernabé-Moreno

Comprender la revolución de la inteligencia artificial (responsable)

Vivimos una revolución tecnológica distinta a todas las anteriores. La inteligencia artificial no se limita a acelerar procesos, automatizar tareas o mejorar herramientas existentes: empieza a ampliar capacidades que siempre hemos considerado profundamente humanas, como aprender, razonar, crear, conversar, interpretar información, formular hipótesis y tomar decisiones. Por primera vez, una tecnología no extiende solo nuestra fuerza física, nuestra memoria o nuestra capacidad de comunicación. Actúa sobre el conocimiento mismo.

Cada gran revolución tecnológica ha cambiado la historia. La máquina de vapor amplificó la fuerza mecánica; la electricidad reorganizó la vida moderna; internet conectó personas, información y mercados a escala global. La inteligencia artificial introduce una diferencia cualitativa: no solo nos permite hacer más rápido lo que ya hacíamos, sino que transforma la manera en que pensamos, aprendemos, trabajamos y resolvemos problemas.

Esta revolución no pertenece a un único sector ni a una única profesión. Atraviesa la ciencia, la medicina, la educación, la industria, las finanzas, la administración pública, la creatividad, la programación y la vida cotidiana. Allí donde hay información, incertidumbre o complejidad, la inteligencia artificial puede tener un papel.

Lo más llamativo es que esta tecnología nos resulta cercana de una forma inédita. Conversa, responde preguntas, resume documentos, escribe código, genera imágenes, propone ideas, traduce, explica y argumenta. Se expresa en el medio más natural para nosotros: el lenguaje. Esa cercanía la hace poderosa y accesible, pero también crea una ilusión peligrosa: la sensación de que comprende el mundo como lo comprendemos nosotros.

No es así. La inteligencia artificial puede producir resultados extraordinarios, pero carece de experiencia humana, conciencia, responsabilidad moral y comprensión vivida del significado. No aprende como aprende una persona. No recuerda como recordamos nosotros. Reconoce patrones y genera respuestas convincentes, pero también se equivoca, inventa, simplifica en exceso o reproduce los sesgos de los datos con los que fue entrenada.

Ahí está una de las diferencias esenciales entre la inteligencia artificial y la humana. Puede ampliar nuestra capacidad de analizar, crear y decidir, pero no sustituye el juicio. Puede ayudarnos a ver más lejos, pero no decide qué merece la pena mirar.

Puede acelerar el conocimiento, pero no responde por sí sola a las preguntas que importan: para qué, para quién, bajo qué límites y con qué consecuencias.

Por eso esta revolución trae oportunidades enormes y, a la vez, riesgos reales. Puede mejorar la productividad, acelerar la investigación científica, ayudar en diagnósticos médicos, optimizar recursos, personalizar la educación y democratizar el acceso al conocimiento. También puede amplificar desigualdades, concentrar poder, transformar empleos, generar dependencia tecnológica, producir desinformación, reforzar sesgos o introducir sistemas opacos en decisiones importantes. El riesgo no está solo en que la tecnología falle, sino en que la usemos sin entenderla.

Por eso necesitamos comprenderla desde sus fundamentos. No basta con aprender a usar herramientas. Hay que entender cómo funcionan los modelos, qué papel tienen los datos, cómo aprenden, cómo se evalúan, por qué fallan, qué límites tienen, qué supuestos esconden y qué problemas pueden o no resolver. En una época en la que las herramientas cambian cada pocos meses, los fundamentos son lo que permanece.

Comprender desde la base no exige convertirse en especialista técnico. Exige desarrollar criterio: distinguir una promesa real de una exageración, saber cuándo una respuesta debe verificarse y reconocer que detrás de cada sistema hay datos, arquitectura, objetivos, decisiones de diseño, incentivos y limitaciones. Significa, sobre todo, no delegar nuestra comprensión del mundo en una tecnología que todavía necesita supervisión, contexto y responsabilidad humana.

Nunca ha sido tan importante seguir aprendiendo. La inteligencia artificial no será un episodio pasajero, sino una transformación continua. Las profesiones cambiarán, las organizaciones se rediseñarán y muchas tareas se reformularán. Pero aprender no será solo adaptarse a nuevas herramientas: será aprender a pensar con ellas, a cuestionarlas, a complementarlas y, cuando haga falta, a ponerles límites.

El futuro que se abre ante nosotros será intenso. Veremos sistemas más capaces, más integrados y más presentes en nuestra vida diaria. Veremos inteligencia artificial colaborando con científicos, médicos, docentes, ingenieros, artistas, emprendedores y responsables públicos. Veremos modelos que no solo generan contenido, sino que ayudan a razonar, experimentar, simular y descubrir. Veremos también tensiones, errores y debates difíciles. Necesitamos, por eso, una sociedad mejor preparada para entender esta tecnología y decidir cómo quiere usarla.

Y hay muchas razones para ser optimistas. La inteligencia artificial puede ayudarnos a abordar problemas que hasta ahora parecían demasiado complejos: acelerar el descubrimiento de nuevos materiales, mejorar tratamientos médicos, optimizar

sistemas energéticos, ampliar el acceso a una educación de calidad, reducir barreras creativas y permitir que más personas participen en la construcción del conocimiento. Puede convertirse en una herramienta extraordinaria para aumentar la capacidad humana, no para reemplazarla.

Pero el optimismo debe ser responsable. No se trata de aceptar la tecnología sin preguntas, ni de temerla sin entenderla, sino de conocerla a fondo para poder orientarla bien. El futuro de la inteligencia artificial no dependerá solo de la potencia de los modelos, sino de la calidad de las decisiones humanas que tomemos en torno a ellos: cómo los diseñamos, cómo los evaluamos, cómo los gobernamos, cómo los usamos y cómo distribuimos sus beneficios.

Este libro nace de esa necesidad. La inteligencia artificial no es solo una colección de herramientas nuevas: es una transformación de nuestra relación con el conocimiento, el trabajo, la creatividad y la toma de decisiones. Entender sus fundamentos es el primer paso para participar activamente en esa transformación: no como espectadores fascinados por la velocidad del cambio, ni como críticos paralizados por el miedo, sino como personas capaces de comprender, cuestionar y construir.

Quizá esa sea la verdadera invitación de esta época: aprender lo suficiente para no quedar al margen, preguntar lo suficiente para no perder el criterio y actuar con la ambición necesaria para que esta revolución tecnológica esté al servicio de una sociedad más capaz, más justa y más humana.

—Juan Bernabé-Moreno
Director de IBM Research Europa

Prólogo de

Esther Checa Gutiérrez

El mito de la neutralidad y el «hackeo» de la optimización opaca de los algoritmos

La tecnología siempre encarna decisiones, intereses, prioridades, sesgos y prejuicios de quienes la diseñan y la financian. La tecnología no es neutral. Reconocer esto es un primer paso para exigir transparencia, equidad y ética en cualquiera de los ámbitos donde la tecnología opere.

La narrativa a la que estamos acostumbrados implica que los sistemas de inteligencia artificial son árbitros imparciales, debido a que su forma de operar está constituida por modelos matemáticos y estadísticos. Sin embargo, este software se ejecuta sobre datos seleccionados por humanos, se procesa mediante categorías definidas por humanos y se optimiza para cumplir objetivos humanos. Cathy O'Neil[1] acuñó el término «weapons of math destruction» para describir a aquellos modelos que regulan aspectos fundamentales de la vida moderna.

O'Neil explica cómo estos sistemas se caracterizan por ser opacos —las personas no comprenden las reglas del juego—, son altamente escalables —su alcance afecta a millones de personas— y provocan daños —las poblaciones más vulnerables tienden a ser las más penalizadas—. Con carácter general, la complejidad técnica de esta tecnología actúa como una barrera que bloquea el debate democrático y el escrutinio público.

La práctica de entrenar a los grandes modelos del lenguaje (LLM) con datos históricos inevitablemente provoca la automatización de desigualdades, sesgos e injusticias estructurales bajo una falsa apariencia de neutralidad. La socióloga Ruha Benjamin, a través del concepto «The new Jim Code»[2], explica cómo las tecnologías digitales reproducen y profundizan de manera sistemática las jerarquías raciales y sociales. Asimismo, indica que los algoritmos codifican la desigualdad de tres formas: amplificando las jerarquías raciales, ignorando las divisiones sociales e intentando «solucionar» el sesgo racial.

Además, la optimización en el diseño de un algoritmo puede entrañar juicios de valor políticos e ideológicos. La optimización es una decisión de valores, ¿qué se busca: eficiencia, maximización de beneficios empresariales, equidad o eficiencia de procesamiento? Esta optimización marca qué objetivos técnicos se subordinan a los intereses humanos.

1 O'Neil, C. (2018). *Armas de destrucción matemática: cómo el big data aumenta la desigualdad y amenaza la democracia*. Capitán Swing.

2 Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press.

Estos juicios y sesgos —raza, género— son detectables, otros son estructurales e invisibles —el idioma de entrenamiento, el origen de los datos, los perfiles que configuran el equipo de trabajo, etc.—. La propia opacidad del ecosistema tecnológico hace que resulte complejo percibir una discriminación algorítmica, ya que esta opera en múltiples capas de abstracción. El sesgo visible se manifiesta de forma evidente en los sistemas de análisis facial, donde los atributos de raza y género se correlacionan de forma directa; sin embargo, los sesgos invisibles y más estructurales se encuentran arraigados en la lingüística computacional, un ejemplo del sesgo semántico, de género y cultural invisible fue documentado en el sistema word2vec de Google.

Cuando los equipos de trabajo definen las variables, los umbrales estadísticos y las taxonomías que definen un sistema de inteligencia artificial, están actuando como «legisladores silenciosos», puesto que estas directrices probablemente afecten a la vida de millones de personas sin debate democrático. Esto nos lleva a la aparición del concepto «algocracia»[3] y la necesidad de construir sistemas algorítmicos que empoderen y prioricen el impacto social real de la tecnología por encima de su rendimiento predictivo o su rentabilidad.

Los investigadores Shakir Mohamed, Marie-Therese Png y William Isaac defienden que la inteligencia artificial no es neutral y hablan del fenómeno «colonialidad algorítmica»[4] a través del cual describen las dinámicas de poder opresivas, extractivas y discriminatorias, propias de la era colonial y que continúan reproduciéndose en el desarrollo y uso actual de la IA. A este fenómeno, los autores proponen un «giro decolonial» fundamentado en tres ejes: reconfigurar las suposiciones del diseño de la IA a través de los cuales se reconozcan los desequilibrios de poder y se recojan perspectivas múltiples en lugar del etnocentrismo occidental; visibilizar a las comunidades históricamente marginadas y que estas asuman un rol de instrucción y guía de la tecnología; y fomentar el desarrollo de la IA basándose en la solidaridad, la equidad y la soberanía tecnológica.

En *Atlas de IA,* Kate Crawford, explica también cómo la inteligencia artificial opera como una industria de extracción planetaria, centra su tesis en tres aspectos: los modelos de IA consumen ingentes recursos naturales para construir y mantener sus centros de datos, acelerando el colapso ecológico; la ingesta indiscriminada —y en la mayoría de los casos no consentida— de información de internet no tiene ningún retorno social, y el funcionamiento de los algoritmos modernos depende de un ejército de trabajadores precarizados que realizan tareas de etiquetado de datos,

[3] Aneesh, A. (2006). *Virtual Migration: The Programming of Globalization.* Duke University Press.

[4] Mohamed, S., Png, M.T. y Isaac, W. (2020). *Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence*. arXiv. Disponible en: arxiv.org/abs/2007.04068.

transcripción de audio y filtrado de contenido, y cuyas condiciones laborales en algunos casos carecen de protección social básica.

Pasar de declaraciones éticas abstractas a la implementación de mecanismos regulatorios de obligado cumplimiento en los sistemas automatizados de toma de decisiones es el reto de gobernanza de nuestra sociedad.

Frente a estos desafíos, este libro nace con el propósito de ir más allá de la teoría y ofrecer un camino práctico hacia una inteligencia artificial responsable. A lo largo de sus páginas, el lector podrá encontrar un análisis detallado de las dimensiones éticas y los marcos regulatorios actuales, con especial atención al nuevo Reglamento de Inteligencia Artificial de la Unión Europea (RIA) y su relación con la privacidad de los datos. Además, esta obra no solo analiza los riesgos, sino que aporta herramientas normativas internacionales y soluciones tecnológicas concretas. Con este enfoque, cualquier persona lectora podrá entender cómo auditar y gestionar la tecnología, así como comprender el valor de desarrollar una IA al servicio de las personas y de la sociedad.

—Esther Checa Gutiérrez
Directora Global de Innovación en t2ó.

Introducción

Cuando la ética y la moral no bastan

Pese a trabajar en la industria de la inteligencia artificial desde dentro y hacerlo a nivel técnico, considero la ética de la IA una temática social aún más interesante que los fundamentos tecnológicos de base, y sobre todo una curiosa fuente de inspiración para conversaciones diversas, ya sean de laboratorio, de bar o de autobús.

A lo largo de los años me he dado cuenta de que este concepto, ligado al desarrollo y adopción de la inteligencia artificial, es uno de esos temas que llaman la atención de la gente. Todo el mundo, independientemente de su nivel de conocimiento técnico, ve en la IA un reflejo de lo que la sociedad debería ser, y entra en debates sociales, morales y sobre el futuro del trabajo con una facilidad pasmosa. Y esto tiene consideraciones maravillosas porque depende de los valores de cada persona, empresa, geografía o incluso religión. No hay una idea única de lo que es ético o moral, por mucho que nos alineemos en torno a principios universales que defienden el bienestar humano o el rol de las personas en un mundo robotizado.

Recuerdo algunas de mis primeras clases en 2019 como profesor de Concordia University en Montreal. Era un entorno de educación continua en el que enseñaba a estudiantes con perfiles muy diversos. La mayor parte de ellos carecía de experiencia y de conocimientos previos en temas de datos, y menos aún en inteligencia artificial. Lo curioso del asunto era la manera de interactuar con ellos en función de la temática. En el resto de las sesiones en las que veíamos temas de infraestructura técnica, arquitectura y de manipulación de datos, los alumnos asistían y participaban en mayor o menor medida según sus conocimientos y capacidades. Pero cuando llegaba la octava sesión dedicada a los temas éticos y sociales, por lo general ya entrada la primavera, los debates se encendían cual proceso de polinización.

Recuerdo sobre todo la pasión con la que dichos estudiantes entraban en los detalles de casos pasados —que veremos en profundidad a lo largo del libro— como el sesgado COMPAS estadounidense para el análisis de la reinserción de presos, el uso indebido de los datos de Facebook junto a la empresa de «marketing político» otrora llamada Cambridge Analytica, o los sistemas de revisión de candidaturas de empleo con algoritmos que discriminaban a candidatas mujeres en una conocida empresa tecnológica. Como profesor, esa mezcla de opinión, pasión y razonamiento para analizar lo que era o no correcto suponía el mejor caldo de cultivo posible para conversaciones dinámicas y ricas en debates.

Recuerdo también el ejemplo de un compañero de trabajo en una conocida empresa de IA canadiense. Trabajaba como científico de datos y estaba participando en un proyecto de precios dinámicos en estaciones de servicio remotas. Comentaba

frustrado cómo el uso de la IA contribuía en este caso a buscar datos de demanda y distancia entre gasolineras para así subir el precio en momentos en los que los conductores no tenían muchas opciones cerca, o simplemente vivían en zonas más remotas. Esta frustración lo llevó a abandonar la empresa al considerar que ese proyecto no se alineaba con sus principios individuales y su visión de una IA en beneficio de la sociedad.

Al mismo tiempo, hemos visto suficientes casos en los que la intuición de las personas ha ido más rápido que la voluntad proactiva de las empresas o incluso los movimientos regulatorios. Aunque parezca un tema reciente, los denominados *deepfakes* en los que se utiliza la IA para suplantar identidades —por lo general para fines nocivos e incluso ilegales—, no son una idea nueva y ya se veían venir en la segunda década de este siglo. Si bien los modelos de IA generativa modernos son mucho más avanzados que los de antes, ya había voces por aquel entonces que denunciaron el uso de estas técnicas para generar contenido pornográfico falso sin el permiso de las personas afectadas.

Y qué decir del papel de la IA en los entornos bélicos. Mientras que todos los mortales —profesionales de la IA o no— pensábamos que el uso de esta tecnología no era aplicable al contexto de la guerra, nos dimos cuenta casi de un día para otro de que no era así. Que la IA estaba bien presente en los conflictos internacionales y que no hay una idea clara de qué se puede o no hacer, y sobre todo cuáles son las reglas escritas o no escritas (más allá de la ley del más fuerte). Aún peor, que la IA podía ser usada según qué gobiernos y administraciones públicas para buscar patrones en los datos de ciudadanos en pro de sus intereses, que *a priori* podían parecer lícitos, pero que *a posteriori* resultaban más discutibles.

Todo esto pone de manifiesto dos verdades:

- Que la IA es un tema de conversación apasionante, que en general genera debates casi intuitivos que nos permiten, como humanos, anticipar riesgos que nunca nos planteamos con otros tipos de tecnologías. Que esas conversaciones son raramente algo técnico y se nutren de distintas personas con puntos de vista diferentes. Eso nos permite dilucidar «lo que queremos hacer».
- Que los enfoques regulatorios o ese «lo que debemos hacer» suelen llegar con cierto retraso, lo que nos fuerza a estar en una especie de «Salvaje Oeste» en el que seguimos apoyándonos en esa voluntad de hacer las cosas bien, mientras esperamos la llegada de las reglas oficiales y que los impactos provocados por quienes no quieren hacer las cosas bien se mitiguen con el tiempo.

Y ahí estamos. Si bien las regulaciones de protección de datos llevan años en el tablero internacional, los ejemplos recientes de reglamentos de inteligencia artificial que veremos a lo largo del libro se han cocinado a fuego lento. Muy lento. Tanto que ha dado tiempo a que los borradores de leyes salgan, se revisen, se discutan, se negocien y se publiquen. Y a que el contexto político y económico influya en los enfoques regulatorios y el continuo —y a mi parecer estéril— debate de regulación *versus* innovación.

Pero llegue tarde o no, saber qué se debe hacer mientras el mundo comienza a entender lo que quiere hacer es un paso fundamental y el *leitmotiv* de elementos como las nuevas regulaciones y los estándares de inteligencia artificial. Además, dichos elementos nos permiten evolucionar de lo ético y moral a algo que se pueda considerar realmente responsable, y que nos traiga al terreno de la rendición de cuentas (denominada *Accountable AI* en la lengua de Shakespeare).

Por otro lado, me gusta pensar que las piezas del puzle están siendo ordenadas y que los aspectos necesarios para crear una IA responsable son ya una realidad. Que el origami de la portada representa esa pieza responsable en el conjunto de las actividades de las empresas y la sociedad en general, y que cada vez más gente es consciente de su necesidad. Que el trabajo comienza ahora, y que además ser responsable y crear un impacto positivo tiene su recompensa con los clientes e incluso con los inversores.

Por ello, a nuestro ávido lector, le presentamos una obra que habla de la IA responsable desde lo tangible. Si bien incluye cuestiones éticas y casi filosóficas, no duda en ir al detalle de lo que esta responsabilidad representa: el alineamiento de lo ético, lo regulatorio, lo normativo y lo tecnológico. Y lo hacemos con la confianza que da trabajar al pie del cañón en estos temas, cada uno de nosotros —humildes coautores— con perfiles diferentes y complementarios, y con las ganas de compartir una visión única de cómo vemos la IA en el mundo moderno, obviamente sesgados por lo que creemos que debería ser.

Que lo disfrutéis y que este libro os permita contribuir a la noble misión de crear un mundo en el que la inteligencia artificial tenga un impacto positivo y los enfoques responsables sean la norma y no la excepción. Va por vosotros.

—Adrián González Sánchez

Guía de IA Responsable

Manual Imprescindible

Este libro es una obra didáctica que puede utilizarse como material de apoyo para el aprendizaje personal de quienes estén interesados en la IA ética y responsable, así como una guía de estudio para cursos formativos.

El texto está organizado en varios capítulos que ofrecen las claves para concretar, desde un enfoque práctico, los conceptos de la IA responsable, en ocasiones etéreos. Para ello, integra todas las dimensiones necesarias para conseguirla —la ética, la regulación y la normativa— y abarca desde la estrategia de la organización hasta la implementación de soluciones inteligentes. Todo ello se basa en estándares extraídos de la industria y de la amplia experiencia de los autores.

Se trata de una obra actualizada que aspira a convertirse en un referente de aprendizaje en el ámbito profesional y académico de la IA responsable, con alcance global y en lengua española.

ISBN 978-84-415-5290-6
2311399
9 788441 552906

ANAYA MULTIMEDIA

www.anayamultimedia.es