Presentación

Este libro presenta de forma ampliada siete de las ponencias presentadas en la jornada que, con el título "El futuro de la inteligencia artificial", se organizó por los cuatro editores del libro en Funcas el 15 de octubre de 2025, y cuya grabación puede seguirse en la web de Funcas y en su canal de YouTube. Los trabajos incluidos en este libro abordan, desde distintos puntos de vista, el desarrollo futuro de la inteligencia artificial (IA), pero el título del libro se ha elegido para enfatizar un aspecto que consideramos fundamental en las contribuciones de los autores: la necesidad de un desarrollo responsable y socialmente beneficioso de la IA.

La IA se está convirtiendo cada vez más en una tecnología con profundas repercusiones socioeconómicas, que condiciona decisiones, comportamientos y oportunidades a escala social. Sin embargo, el debate público y técnico sigue aún centrado en el rendimiento (precisión, velocidad o eficiencia), y no tanto en las consecuencias, los valores y las formas de gobernanza que acompañan su despliegue. Este volumen parte de la idea de que el desarrollo futuro de la IA no depende únicamente de modelos más potentes, sino de nuestra capacidad colectiva para diseñar, evaluar y utilizar sistemas inteligentes de manera responsable, comprensible y socialmente legítima. Más allá de marcos éticos generales o normativos, quizás algo abstractos, los capítulos que componen este libro exploran cómo integrar la responsabilidad en el corazón mismo de la inteligencia artificial: en sus métricas, en sus arquitecturas de aprendizaje, en la formación de la ciudadanía, en la terminología que empleamos, en la comprensión de sus ciclos de evolución y, finalmente, en la gobernanza y la incorporación explícita de valores sociales en sistemas autónomos. Desde perspectivas complementarias, técnicas, epistemológicas, educativas y éticas, los autores ofrecen herramientas conceptuales y metodológicas para pensar una IA que no solo funcione bien, sino que también importe lo que hace, cómo lo hace y para quién lo hace.

En el capítulo I, "De la precisión a la responsabilidad: nuevas métricas para evaluar la IA", **Amparo Alonso Betanzos** ilustra con cuatro ejemplos la importancia de tener en cuenta las consecuencias sociales y económicas de la aplicación de la IA, ampliando el marco de evaluación para medir estas consecuencias e incluirlas junto a las medidas clásicas de eficacia de un sistema de decisión automática, como pre-

cisión, sensibilidad o eficiencia. Los ejemplos abordan cuatro temas que cubren un recorrido muy amplio y de gran actualidad. El primero es la identificación de usuarios tóxicos en redes sociales, un problema de importancia creciente que afecta desde el desarrollo de los adolescentes hasta la información difundida en redes sociales en las elecciones. El segundo, la mejora de los cada vez más usados sistemas de recomendación, buscando mayor explicabilidad y sostenibilidad. El tercero aborda las consecuencias de la priorización genética, que pretende identificar los genes con mayor probabilidad de asociación con procesos biológicos o patologías de interés. El cuarto ejemplo analiza el análisis de datos diádicos, que son aquellos que provienen de un par de entidades relacionadas, como dos usuarios de redes que interactúan entre sí. En estos datos, las predicciones sobre pares de entidades pueden tener consecuencias importantes en ámbitos sensibles que hay que tener en cuenta en los criterios de evaluación del sistema.

En el capítulo II, "Aprendizaje federado para una sociedad de máquinas", **Senén Barro**, **José Miguel Burés** y **Roberto Iglesias** proponen una inmersión profunda en lo que califican como un cambio de paradigma en el aprendizaje automático: la transición de modelos centralizados a estructuras de aprendizaje distribuidas y colaborativas. El capítulo describe cómo el Aprendizaje Federado (FL) permite resolver la tensión histórica entre la necesidad de grandes volúmenes de datos para el entrenamiento de modelos y el derecho irrenunciable a la privacidad de los usuarios. Los autores detallan la mecánica operativa del FL, donde el entrenamiento se desplaza hacia el "borde" (también llamado *edge computing*), permitiendo que dispositivos como teléfonos móviles o sensores médicos procesen información localmente. De esta forma, en lugar de transmitir datos sensibles a una nube central, solo se envían las actualizaciones de los parámetros del modelo local, que luego se agregan para mejorar el modelo global. El capítulo analiza con precisión los desafíos que aún limitan su aplicabilidad universal, como la heterogeneidad de los datos, la deriva de concepto (cómo gestionar modelos en entornos donde los datos cambian con el tiempo, también llamado *concept drift*), invalidando patrones previos, o la eficiencia en la comunicación. Finalmente, Barro y sus colaboradores subrayan que este enfoque no es solo una solución técnica que además aumenta la seguridad y la privacidad de los datos, sino la base de una "sociedad de máquinas". En este ecosistema, el aprendizaje deja de ser un proceso aislado para convertirse en un esfuerzo colectivo y democrático, donde la personalización del aprendizaje permite que la IA se adapte a las particularidades de cada individuo sin sacrificar el beneficio del conocimiento compartido.

En el capítulo III, "Alfabetización en IA: hacia un uso responsable y controlado de la tecnología", **Francisco Bellas** llama nuestra atención sobre la necesidad de alfabetización en IA para controlar y hacer un uso responsable de esta tecnología. No se trata de generar especialistas en IA, sino de educar a la población en general para que comprenda sus fundamentos, conozca sus derechos y deberes según el marco legal

europeo y desarrolle un pensamiento crítico sobre sus aplicaciones. Solamente así tendremos ciudadanos capacitados para usarla de manera responsable.

En el capítulo IV, titulado "Agentic AI y Multiagentic: ¿Estamos reinventando la rueda? Una reinterpretación epistemológica", su autor, **Vicent Botti,** aborda con rigor crítico la proliferación reciente de términos como IA agentiva o sistemas multiagentivos, argumentando que no constituyen una ruptura conceptual, sino una reelaboración tecnológica de paradigmas bien establecidos en el campo de los agentes autónomos y los sistemas multiagente. A través de un análisis detallado que recorre los fundamentos filosóficos, las definiciones clásicas y las arquitecturas consolidadas, el capítulo subraya la importancia de emplear una terminología precisa y de aprovechar el vasto acervo de investigación previa para evitar redundancias y fragmentación en el desarrollo de sistemas inteligentes. Esta postura se alinea de manera natural con los principios de la ingeniería de valores en inteligencia artificial, al promover un diseño que optimice recursos conceptuales y técnicos, garantice la interoperabilidad, y facilite la creación de soluciones robustas, éticas y sostenibles. Lejos de menospreciar los avances impulsados por los grandes modelos de lenguaje, el texto aboga por una integración consciente y fundamentada, donde la innovación se construya sobre cimientos sólidos y no sobre modas terminológicas. Así, el capítulo es una llamada al rigor epistemológico y a la continuidad disciplinar, invitando a la comunidad a enfocar sus esfuerzos en la verdadera evolución técnica y ética de la inteligencia artificial.

En el capítulo V, "Panarquía artificial: expansión y ciclos adaptativos de la IA", **David Martinez Rego** propone una forma diferente y especialmente sugerente de entender la evolución de la inteligencia artificial. Frente a la narrativa habitual que habla de avances lineales o de ciclos de euforia y decepción, el capítulo invita a mirar la IA como un sistema complejo, vivo y en constante adaptación, influido tanto por los avances tecnológicos como por factores sociales, económicos y culturales. Apoyándose en el concepto de panarquía, el autor describe cómo la IA evoluciona a través de ciclos interconectados de crecimiento, consolidación, crisis y reorganización. Esta perspectiva permite explicar por qué algunas tecnologías florecen de forma explosiva mientras otras quedan relegadas, por qué ciertas ideas resurgen décadas después con renovada fuerza, y por qué la innovación no ocurre al mismo ritmo en todos los niveles, desde la investigación básica hasta la adopción industrial y social. El capítulo recorre momentos clave de la historia reciente de la IA (desde los primeros enfoques estadísticos hasta el auge del aprendizaje profundo y los grandes modelos de lenguaje, LLMs– *Large Language Models*), mostrando que cada avance se apoya en equilibrios frágiles entre datos, computación, teoría y expectativas sociales. Lejos de presentar una visión triunfalista, el análisis subraya también las limitaciones, tensiones y riesgos que acompañan a cada ola tecnológica.

Con un enfoque divulgativo y reflexivo, Panarquía artificial ofrece al lector herramientas conceptuales para interpretar el presente y pensar el futuro de la IA con mayor profundidad. Más que responder a la pregunta de "¿qué puede hacer la IA?", el capítulo ayuda a comprender cómo y por qué evoluciona, y qué implica esa evolución para la toma de decisiones, la gobernanza tecnológica y el desarrollo responsable de sistemas inteligentes.

En el capítulo VI, titulado "Autonomía basada en valores: hacia una gobernanza responsable de los agentes IA", los autores **Holger Billhardt**, **Alberto Fernández**, **Andrés Holgado** y **Sascha Ossowski** abordan uno de los desafíos más urgentes y complejos de la inteligencia artificial contemporánea: cómo garantizar que los sistemas autónomos y multiagente actúen de forma alineada con los valores éticos de la sociedad que los integra. A través de un análisis riguroso que combina perspectivas técnicas, legales y éticas, el capítulo explora mecanismos de gobernanza para sistemas abiertos —como los basados en subastas para la gestión de intersecciones de tráfico— y presenta modelos formales para dotar a los agentes de una conciencia de valores que les permita razonar sobre dilemas morales y tomar decisiones responsables. Esta aproximación se enmarca de manera natural en los principios de la ingeniería de valores en inteligencia artificial, al ofrecer metodologías concretas para traducir valores abstractos —como la equidad, la eficiencia o la seguridad— en diseños computacionales verificables. Lejos de limitarse a la especulación teórica, el texto avanza hacia propuestas prácticas de aprendizaje de sistemas de valores a partir de preferencias observadas, integrando así la reflexión ética en el ciclo mismo de desarrollo de los agentes autónomos. De este modo, el capítulo no solo alerta sobre los riesgos de una autonomía desvinculada de los valores sociales, sino que proporciona herramientas fundamentales para construir sistemas de IA que sean técnicamente robustos, legalmente conformes y socialmente legítimos, estableciendo un puente indispensable entre la teoría ética y la práctica ingenieril en el panorama actual de la inteligencia artificial.

Finalmente, en el capítulo VII, "Máquinas que saben lo que importa: ingeniería de la conciencia de los valores sociales", **Carles Sierra** aborda el importante problema de desarrollar métodos de IA que incluyan representaciones computacionales de los valores, de forma que automáticamente sean tenidos en cuenta en la decisión o recomendación creada por el sistema de IA. En el capítulo se analizan métodos para incrustar valores en el sistema teniendo en cuenta los dilemas éticos y el contexto social. Se analizan estrategias de negociación basadas en valores y se describe un proceso de decisión donde un agente analiza el contexto y la conducta de la contraparte, evalúa los conflictos entre valores y aplica un criterio híbrido de decisión que satisface los valores del agente, pero dando cierto peso a los valores de la contraparte.

Los editores queremos agradecer a los autores de los capítulos su excelente disposición para cumplir los plazos previstos y realizar el esfuerzo de prepararlos pensando en un público no necesariamente especialista en el tema. Agradecemos también al director general de Funcas, Carlos Ocaña, el apoyo a esta iniciativa y a todas las actividades relacionadas con el *big data*, la IA y sus aplicaciones. Agradecemos a Cecilia y Esperanza su siempre eficaz labor de organización de las jornadas y a Myriam González y a su equipo la atenta y eficaz edición de este libro.

Amparo Alonso Betanzos, **Daniel Peña**, **Pilar Poncela** y **Carles Sierra**

Enero 2026