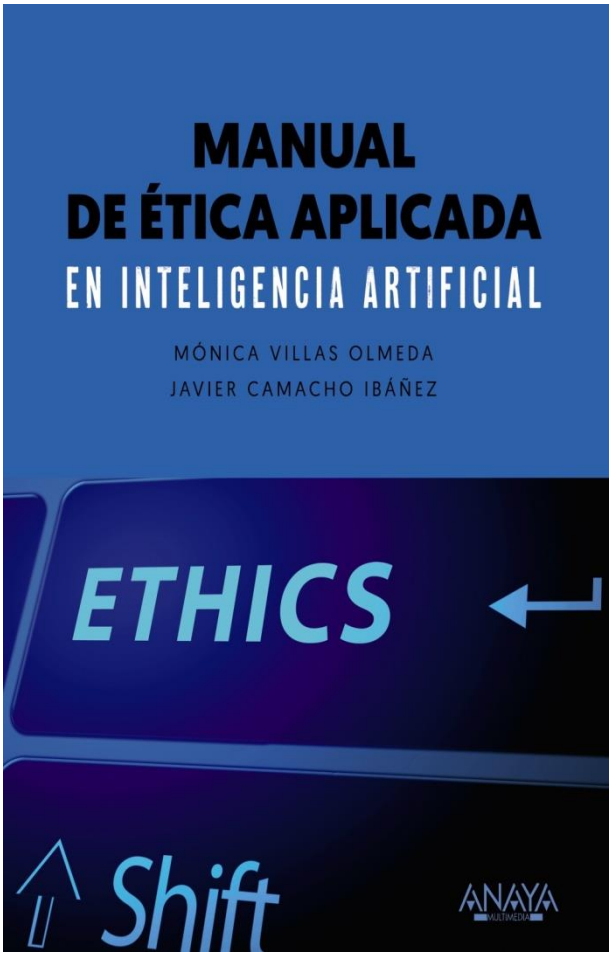
**Reseña del libro de Vilas Olmeda, Mónica y Camacho Ibáñez, Javier, *Manual de Ética Aplicada en Inteligencia Artificial.* Madrid: Anaya. 2022.**

**Por Fernando González Galán.**




Mónica Villas Olmeda es ingeniera Industrial egresada en 1994 del Instituto Católico de Artes e Industrias (ICAI) de la Universidad Pontificia Comillas, además posee un Master of Business Administration (MBA) por la Universidad Autónoma 2007, una Certificación de Cloud 2016 y otra Certificación en Design thinking 2017. Actualmente, además de trabajar en una tesis doctoral sobre Inteligencia Artificial en Ciberseguridad en la UNED, es consultora, docente y directora en programas de IA y tecnologías exponenciales en Deusto, UNIR, ESIC, *Immune Technology* y *Analyticae*. Cofundadora de ODISEIA (Observatorio de Impacto Social y Ético de la Inteligencia Artificial) y directora de formación, anteriormente desarrolló su carrera profesional en IBM. Javier Camacho

Ibáñez es profesor del Instituto Católico de Administración y Dirección de Empresas (ICADE) de la Universidad Pontificia Comillas. Además, posee una amplia formación tanto en administración empresarial como en telecomunicaciones. *Doctor en Economía y Empresa por la Universidad Pontificia Comillas (2016), Executive MBA por el IESE (Universidad de Navarra), Capability Development Program Coach, por la Universidad de Strathclyde (Glasgow), Máster en Investigación en Economía y Empresa por la Universidad Pontificia Comillas e Ingeniero de Telecomunicaciones por la Universidad Politécnica de Madrid. Mientras actualmente es miembro de la Cátedra de Ética Económica y Empresarial de la Universidad Pontificia Comillas, además* c*ompagina su labor como consultor y consejero de empresas con la labor docente e investigadora en temas de Ética Empresarial.*

Entre las motivaciones de publicar el presente *Manual* se pueden hallar, principalmente, exponer las implicaciones éticas de la IA, sobre todo a través de la *machine learning* (ML) cuyos algoritmos hacen posible precisamente el aprendizaje automático (AA). El algoritmo es una agrupación de pasos reglados bien definidos cuyo proceso permite solucionar un problema, realizar un cómputo, procesar datos o ejecutar diferentes tareas o actividades. Es interesante destacar que el aprendizaje automático se ha ganado un espacio en el ámbito de las ciencias de la computación constituyéndose incluso en una de las principales ramas de la inteligencia artificial gracias a que tiene por tarea lograr el aprendizaje de las computadoras a partir del desarrollo de una serie de técnicas. Sabemos que el conocimiento no viene dado genéticamente, ni está presente en ningún genotipo, sino que se adquiere con la experiencia en el trato humano, pero, además, si tenemos en cuenta el uso, especialmente, de datos, obtenemos también el aprendizaje en la computación automática la cual incorpora luz gracias curiosamente a la práctica.

La obra consta de seis capítulos acompañados de los correspondientes y muy útiles resúmenes al final de cada uno de los mismos. A ella se suma la introducción, las conclusiones, las referencias y un índice alfabético. La exposición es clara y sencilla de tal forma que cualquier persona interesada en el tema pueda conocer fácilmente su contenido. La idea del *Manual* es también diáfana, a saber, la comprensión y disponibilidad de respuestas para diferentes preguntas que inquietan al público especialmente dentro del incierto contexto actual en el que se desarrolla la Inteligencia

Artificial. Aunque el Manual está dirigido a estudiantes, profesores y técnicos también está abierto a lectores no técnicos.

Tras el prólogo, el primer apartado del libro, "1. Introducción", responde básicamente a la razón de publicar un manual sobre ética en IA dando unas pautas para su lectura.

El capítulo "2. Ética y Tecnología" entiende la ética no como freno que detiene la innovación sino como dispositivo que le concede seguridad y confianza. En este sentido la ética nace en una filosofía de la tecnología cuya noción de neutralidad se deja en manos del lector no sin antes brindarle diferentes perspectivas que puedan socorrerle en la faena de conocer la pericia para el logro de la imparcialidad de la tecnología. El instrumentalismo supone que la tecnología es neutral, el determinismo sospecha que la tecnología es parcial y responde a intereses previamente bien fijados, la mediación considera que la tecnología simplemente es mediadora entre las personas y el mundo. Los autores entienden la ética como un cuestionamiento continuado que debe hacerse cada persona respondiendo razonadamente a cómo tomar las mejores decisiones de tal forma que puedan así guiar el comportamiento humano de la forma posible más adecuada. Para este razonamiento proponen tres corrientes éticas principales la ética utilitaria que responde a las consecuencias, la ética del deber racional que responde a los principios y la ética de la virtud que responde a la excelencia. Finalmente, se introduce la concepción de la "ética por diseño" tras explorar aspectos éticos de la IA. La "ética por diseño" persigue evitar los males mayores que, vulnerando la ética, se observan como resultado de la aplicación de la IA, es decir, apremia incorporar valores, principios éticos, legales y sociales desde las etapas incipientes ubicadas en la misma concepción del diseño tecnológico hasta su culminación final.

El capítulo "3. Principios" trata de los principios éticos de la IA, los cuales, basados en los derechos humanos, son agrupados por los autores en "justicia", "no producción de daños", "generación de beneficios" y "respeto a la autonomía". Una aplicación efectiva de los principios éticos requiere para los autores conocer que la IA toma las decisiones de forma automática basándose en datos y, además, apreciar quiénes son los *"stakeholders"* (empleados, accionistas, clientes, proveedores, gobiernos, comunidades, etc.) pues éstos pueden tener diferentes intereses que afecten a los principios éticos. Llevar a la práctica estos principios requiere de marcos de trabajo o *"frameworks",* sin embargo, debido a la

falta de acuerdo global se están teniendo en cuenta tres bloques geográficos, a saber, Estados Unidos, Europa y China.

El capítulo "4. Responsabilidad" la entiende como concepto que funda la disposición de sistemas de IA en los cuales puedan confiar los distintos usuarios, empresas e instituciones. Se distingue responsabilidad legal, responsabilidad moral, responsabilidad impactada por la tecnología y cómo afectan los sistemas de IA al sujeto, objeto y razón de la responsabilidad. Los autores tratan los principios de beneficencia, autonomía y justicia como fundamentales para la acción responsable sin olvidar los obstáculos que impiden la decisión cabal y el concepto de gobernanza.

El capítulo "5. Privacidad" la observa como derecho universal particular y personal de cada ser humano que no se puede transferir ni al cual se puede renunciar. El apartado "privacidad e identidad" relaciona identidad con privacidad porque aquella contiene datos personales que deben ser protegidos como privacidad de cada usuario. Los autores muestran como el "Reglamento General de Protección de Datos" de la Unión Europea persigue ese objetivo. A continuación, el capítulo presenta dos procedimientos para lograr la preservación de la privacidad en la gestión de los sistemas relacionados con la IA: la anonimización y la resolución del dilema privacidad transparencia a través de la privacidad por diseño, la integridad contextual, la transparencia estructurada, la privacidad diferencial, la encriptación homomórfica, la computación multiparte, el aprendizaje federado y los datos sintéticos. Todo ello permite desarrollar una ingeniería de la privacidad. La "anonimización" de datos se entiende como un proceso mediante el cual se elimina la información personal, privada o sensible, para evitar su asociación con identidad alguna. Ello posibilita que, sin menoscabar el derecho a la privacidad, sea factible manejar y acceder a la información.

El capítulo "6. Equidad" se inicia tratando la diferencia entre sesgo y discriminación. El sesgo se produce en el proceso de formación del algoritmo dando lugar a diferentes tipos en función de la etapa de evolución del cifrado algorítmico. Así los autores hablan de los sesgos relacionados con el conjunto de datos a utilizar, a saber, sesgo histórico, sesgo de representación o muestreo y sesgo de medida y de sesgos relacionados con el diseño del algoritmo o su despliegue, es decir, sesgo de aprendizaje, sesgo de evaluación, sesgo de agregación y sesgo de implementación. La discriminación es definida como el trato diferente que causa perjuicio por atribuciones de raza, sexo, ideas políticas, religión, etc.

Clasifican los distintos tipos de discriminación como directa, indirecta, sistémica, estadística, explicable y no explicable. La equidad aplicada tanto a *machine learning* como a su ciclo, la equidad grupal e individual junto a las métricas de equidad y las diferentes herramientas en el mercado completan este capítulo.

El capítulo "7. Explicabilidad" se contempla como el proceso que tiene como objetivo detallar decisiones tomadas por los algoritmos en relación con las acciones de los usuarios. Estos consumidores pueden ser desarrolladores de IA, expertos en una industria o área, reguladores o auditores y usuario final. Los modelos de explicabilidad se pueden clasificar en función de la complejidad del modelo, a saber, explicabilidad intrínseca y explicabilidad post – hoc; en función de la tipología de la explicación, explicabilidad global y explicabilidad local; y en función del modelo, explicabilidad ligada al modelo y explicabilidad agnóstica del modelo. El capítulo finaliza presentando las diferentes herramientas del mercado junto con algunos ejemplos y los retos futuros.

El libro resulta extraordinariamente entretenido por su asequible lectura, su aguda explicación y el detalle que dedica a un tema, la ética aplicada a la IA, por lo demás, nada fácil de abordar. Un manual imprescindible para toda aquella persona con inquietudes éticas en medio del inmenso dispositivo que representa el desafía de la IA.