COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

GUÍA DOCENTE
2026 - 2027

FICHA TÉCNICA DE LA ASIGNATURA

Datos de la asignatura	
Nombre completo	Ética y Explicabilidad
Código	DCIA-MUIAA-514
Impartido en	Máster en Inteligencia Artificial Aplicada [Primer Curso]
Nivel	Posgrado Oficial Master
Cuatrimestre	Semestral
Créditos	3,0 ECTS
Carácter	Obligatoria
Departamento / Área	Departamento de Ciencias de la Computación e Inteligencia Artificial
Responsable	Javier Jurado González
Horario de tutorías	Concertar cita por email

Datos del profesorado	
Profesor	
Nombre	Javier Jurado González
Departamento / Área	Departamento de Organización Industrial
Correo electrónico	jjuradog@icai.comillas.edu
Profesor	
Nombre	Jaime Pizarroso Gonzalo
Departamento / Área	Departamento de Ciencias de la Computación e Inteligencia Artificial (DCIA)
Despacho	Santa Cruz de Marcenado 26
Correo electrónico	jpizarroso@comillas.edu
Teléfono	2732

DATOS ESPECÍFICOS DE LA ASIGNATURA

Contextualización de la asignatura

Aportación al perfil profesional de la titulación

La Inteligencia Artificial ha dejado de ser un campo experimental para convertirse en un componente estructural de múltiples sistemas que afectan a la vida humana, desde la medicina hasta la justicia, desde las finanzas hasta las infraestructuras críticas. En este contexto, los profesionales de la IA ya no pueden limitarse a ser desarrolladores de modelos eficientes, sino que deben asumir una responsabilidad ética y social en relación con las decisiones que sus sistemas automatizados pueden llegar a tomar o influir.

Esta asignatura proporciona al alumno un marco integral para abordar dos dimensiones clave y complementarias de la IA contemporánea: su **dimensión ética y regulatoria** y su **explicabilidad**.

Por un lado, el bloque de **Ética y regulación** introduce al estudiante en los principios y corrientes fundamentales de la ética aplicada, así

como en los marcos legales emergentes que regulan el uso de algoritmos en contextos sensibles, tanto a nivel europeo como global. Le permite reconocer los dilemas morales específicos de la IA —como la opacidad algorítmica, la discriminación automatizada, la autonomía de los agentes artificiales o la asignación de responsabilidad— y evaluarlos críticamente desde una perspectiva plural y argumentada.

Por otro lado, el bloque de fundamentos técnicos de la **IA Explicable (XAI)** dota al alumno de herramientas para diseñar, seleccionar e implementar modelos que no solo funcionen, sino que también puedan ser entendidos, auditados y confiados por usuarios humanos. Se abordarán diferentes niveles y tipos de explicabilidad (global, local, intrínseca, post-hoc...), así como sus principales técnicas, limitaciones y desafíos, permitiendo al estudiante valorar el equilibrio entre transparencia, precisión y robustez según el caso de uso.

Con todo ello, el alumno no solo adquirirá competencias técnicas avanzadas en el desarrollo y evaluación de sistemas explicables, sino que también incorporará una perspectiva crítica y fundamentada sobre el papel que estas tecnologías juegan —y deben jugar— en una sociedad democrática y plural. Esta doble perspectiva contribuirá de forma decisiva a su perfil profesional, preparándole para liderar proyectos de IA en contextos reales que exigen tanto excelencia técnica como legitimidad ética y confianza social.

En definitiva, esta asignatura contribuye a la formación del perfil profesional del titulado al proporcionarle competencias fundamentales en el análisis crítico, ético y social de la Inteligencia Artificial. El estudiante adquiere herramientas para:

- **Evaluar el impacto social, económico y legal de los sistemas de IA**, integrando perspectivas de responsabilidad, sostenibilidad y justicia.
- **Aplicar marcos éticos en el diseño, desarrollo y uso de algoritmos de IA**, garantizando la toma de decisiones alineada con valores humanos y principios normativos.
- **Desarrollar técnicas de explicabilidad y transparencia en modelos de aprendizaje automático**, facilitando su interpretación, confianza y adopción en entornos profesionales.
- **Contribuir a la gobernanza y regulación de la IA**, aportando una visión crítica y constructiva en la definición de políticas y buenas prácticas.

Competencias - Objetivos

Competencias

Conocimientos o contenidos	
CO7	Conoce los sectores principales de aplicación de la IA, las oportunidades y los nuevos retos que se presentan, desde el punto de vista de negocio, ético, legal y de sostenibilidad (ODS)
Habilidades	
HA9	Evaluar modelos de Inteligencia Artificial aplicando las técnicas de interpretabilidad y explicabilidad adecuadas al caso de negocio y valorando sus implicaciones éticas y legales.
Competencias	
CP3	Colaborar eficazmente en equipos multidisciplinares y comunicar resultados técnicos y conclusiones con rigor, claridad y adecuación a públicos diversos, expertos y no expertos.
CP4	Desarrollar modelos de Inteligencia Artificial asegurando el cumplimiento de normativas legales y principios éticos en su aplicación en diversos sectores.
CP6	Evaluar explicaciones de modelos de IA, detectar sesgos o inconsistencias y proponer estrategias de mejora que aumenten su transparencia y fiabilidad.

BLOQUES TEMÁTICOS Y CONTENIDOS

Contenidos – Bloques Temáticos

Descripción general

1. Aspectos éticos y regulatorios. Principios éticos. Marco regulatorio nacional e internacional.
2. Fundamentos de explicabilidad. Introducción a la interpretabilidad y explicabilidad de modelos.
3. Técnicas en modelos clásicos. Modelos interpretables por diseño. Importancia de las características. Explicaciones en modelos combinados.
4. Introducción a la explicabilidad en sistemas de IA. Conceptos generales. Casos de uso en modelos avanzados

Descripción detallada

Bloque de Ética y regulación

1. **Fundamentos para una ética de la IA:** Tecnología como dimensión antropológica, el mito de la neutralidad tecnológica, la dimensión ética del ingeniero, historia de la ética de la IA, principios generales y marco regulatorio nacional e internacional.
2. **Sesgos y discriminación en la IA:** El problema de las variables proxy, taxonomía del sesgo algorítmico, justicia algorítmica, impacto en el Sur Global…
3. **Responsabilidad, Gobernanza y Explicabilidad:** Debate sobre la IA como sujeto moral (agencia/paciencia), el problema del alineamiento, XAI, abdicación moral, dignidad, transparencia, rendición de cuentas, AI Act…
4. **Privacidad, Protección de Datos y Propiedad Intelectual:** Los datos como bien público/privado, límites morales al capitalismo de la vigilancia, propiedad intelectual y entrenamiento masivo, normativa, RGPD…
5. **Futuro, Geopolítica e Impacto Social:** Oportunidades, riesgos y gestión de expectativas a futuro para: salud y longevidad, ciencia y tecnología, sostenibilidad ambiental, geopolítica, derecho al trabajo, democracia, riesgos cognitivos y emocionales, escatológicos/existenciales, desconocidos,…

Bloque de Explicabilidad

1. **Fundamentos de XAI:** Introducción a explicabilidad y Taxonomía de XAI: modelos transparentes y opacos, explicabilidad intrínseca y post-hoc, explicaciones locales y globales, agnósticas y específicas del modelo.
2. **Explicabilidad en datos tabulares:** Modelos interpretables: regresión, árboles y GAM/EBM. Métodos post-hoc para datos tabulares: importancia por permutación, PDP/ICE, LIME, valores de Shapley, SHAP y explicaciones contrafactuales.
3. **Explicabilidad en Visión por Computador:** Interpretación de representaciones internas en CNN, mapas de activación y ejemplos máximamente activadores. Métodos basados en perturbaciones y oclusión. Métodos basados en gradientes: saliency e Integrated Gradients. Métodos de localización basados en activaciones: CAM y Grad-CAM. Limitaciones de los mapas de atribución.
4. **XAI para modelos secuenciales y LLMs: de la atribución a la interpretabilidad mecanicista:** Particularidades de datos secuenciales: dependencia, orden y contexto. Atribución a variables, instantes y tokens mediante perturbaciones, Integrated Gradients y métodos tipo TimeSHAP. Representaciones internas y conceptos. Introducción a la interpretabilidad mecanicista: features, neuronas, cabezas de atención, circuitos, ablaciones, activation patching y análisis causal de componentes internos.

Durante el transcurso de los bloques se realizarán competiciones en directo y un proyecto final basado en los contenidos.

METODOLOGÍA DOCENTE

Aspectos metodológicos generales de la asignatura

La asignatura combina un enfoque teórico-práctico e interdisciplinar, orientado tanto a la comprensión conceptual profunda como a la aplicación técnica y crítica de los contenidos. La metodología se estructura en torno a tres ejes complementarios:

- A través de clases magistrales participativas, análisis de casos reales, y debates guiados, el alumnado se introduce en los principales dilemas éticos y marcos legales en torno al desarrollo y uso de la inteligencia artificial. Se fomenta la capacidad de argumentación, la deliberación informada y la toma de conciencia sobre la responsabilidad profesional en contextos tecnológicamente complejos. Las sesiones incluyen ejercicios de evaluación de impacto ético-legal y discusión crítica a partir de situaciones actuales.
- El núcleo técnico se desarrolla mediante sesiones de explicación conceptual, talleres de análisis de modelos, y prácticas de programación sobre conjuntos de datos reales. Se proporciona al estudiante una *caja de herramientas* variada y se le entrena en su aplicación comparativa, interpretación crítica y limitaciones. Las clases alternan la exposición estructurada con laboratorios prácticos en los que se emplean modelos, se aplican técnicas explicativas y se evalúan sus propiedades (fidelidad, estabilidad, claridad, etc.).
- Aprendizaje basado en proyectos y práctica guiada: A lo largo de la asignatura, se desarrollan sesiones prácticas estructuradas que permiten aplicar los contenidos a contextos reales. Estas incluyen la exploración de sesgos en modelos, la generación de explicaciones para decisiones críticas y la evaluación de su calidad y robustez.

Metodología Presencial: Actividades

Clases magistrales expositivas y participativas	CO7, HA9, CP4, CP6
Actividades de evaluación continua del rendimiento	CO7, HA9, CP3, CP4, CP6
Sesiones prácticas o de laboratorio	CO7, HA9, CP3, CP4, CP6

Metodología No presencial: Actividades

Trabajos o Proyectos	CO7, HA9, CP3, CP4, CP6
Estudio Personal	CO7, HA9, CP4, CP6

RESUMEN HORAS DE TRABAJO DEL ALUMNO

Horas Presenciales		
AF1. Clases magistrales expositivas y participativas	**AF4. Actividades de evaluación continua del rendimiento**	**AF5. Sesiones prácticas o de laboratorio**
18	2	10
Horas no presenciales		
AF6. Trabajos o Proyectos	**AF10. Estudio personal**	
30	30	
		CRÉDITOS ECTS: 3 (90 horas)

EVALUACIÓN Y CRITERIOS DE CALIFICACIÓN

La asignatura se dividirá en dos bloques principales:

Bloque Ética y regulación (40%)	
Examen	70%
Evaluación continua	30%

Bloque Explicabilidad (60%)	
Examen	40%
Proyecto final	50%
Evaluación continua: participación en clase y sesiones de prácticas	10%

Calificaciones

Convocatoria Ordinaria

Para aprobar la asignatura, se deberá aprobar (>=5) cada uno de los bloques por separado. En caso contrario, la nota de la asignatura será la menor nota de los dos bloques, teniendo que recuperar solo los bloques suspensos en la convocatoria extraordinaria.

Actividad	**Porcentaje**
Bloque Ética y regulación	
Examen Final	70%
Evaluación continua	30%
Bloque Explicabilidad	
Examen Final	40%
Proyecto final	50%
Evaluación continua	10%

Para el bloque de Ética y Legal, se aplicarán los siguientes criterios:

- En caso de suspender el examen final (<5), la nota del bloque será la nota del examen final.

Para el bloque de Explicabilidad, se aplicarán los siguientes criterios:

- En caso de suspender el examen y/o el proyecto final (<5), la nota del bloque de explicabilidad será la menor nota del examen o el proyecto.
- En convocatoria extraordinaria se deberá(n) recuperar la(s) parte(s) suspensa(s) del bloque, manteniendo la nota de la parte aprobada si la hubiera.

En caso de suspender alguno de los bloques, la nota de la asignatura será la nota del bloque suspenso. En caso de aprobar o suspender ambos bloques, la nota de la asignatura será la media ponderada de ambos bloques:

Nota final = 40% Nota Ética y Legal + 60% Nota Explicabilidad

Convocatoria Extraordinaria

Para aprobar la asignatura, se deberá aprobar (>=5) cada uno de los bloques por separado. En caso contrario, la nota de la asignatura será la menor nota de los dos bloques.

Actividad	Porcentaje
Bloque Ética y Legal	
Examen Final	100%
Bloque Explicabilidad	
Examen Final	50%
Proyecto individual	50%

Para el bloque de Explicabilidad, se aplicarán los siguientes criterios:

- En caso de suspender el examen final y/o el proyecto final(<5), la nota del bloque de explicabilidad será la menor nota del examen o el proyecto.

En caso de suspender alguno de los bloques, la nota de la asignatura será la nota del bloque suspenso. En caso de aprobar o suspender ambos bloques, la nota de la asignatura será la media ponderada de ambos bloques:

Nota final = 40% Nota Ética y Legal + 60% Nota Explicabilidad

Uso de IA en la asignatura:

- El uso de IA para crear trabajos completos o partes relevantes, sin citar la fuente o la herramienta o sin estar permitido expresamente en la descripción del trabajo, será considerado plagio y regulado conforme al Reglamento General de la Universidad.
- Todos los exámenes de la asignatura se realizarán sin ningún tipo de asistencia por parte de modelos de IA. Los estudiantes deben demostrar sus conocimientos básicos y su capacidad para manejar el contenido de la asignatura a nivel personal.
- Para las prácticas semanales de la asignatura, así como para el proyecto final de la asignatura, se recomienda encarecidamente al alumnado no emplear modelos de IA de manera extensiva. El uso de los mismos debe limitarse, como máximo, a la planificación, desarrollo e investigación de las ideas. Las entregas finales de estos evaluables deben mostrar cómo se han desarrollado, implementado y refinado las ideas presentes en cada caso.

BIBLIOGRAFÍA Y RECURSOS

Bibliografía Básica

- Agencia Española de Protección de Datos. (2020). *Adecuación al RGPD de tratamientos que incorporan inteligencia artificial: Una introducción*. https://www.aepd.es/guias/adecuacion-rgpd-ia.pdf
- Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press. https://fairmlbook.org/
- Bilbao Alberdi, G., Fuertes Pérez, F. J., & Guibert Ucín, J. M. (2006). *Ética para ingenieros*. Desclée de Brouwer.
- Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. En K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 316–334). Cambridge University Press. https://doi.org/10.1017/CBO9781139046855.020
- Coeckelbergh, M. (2021). *Ética de la inteligencia artificial* (L. Álvarez Canga, Trad.). Cátedra. (Trabajo original publicado en 2020).
- Cortina Orts, A. (2024). *¿Ética o ideología de la inteligencia artificial? El eclipse de la razón comunicativa en una sociedad tecnologizada*. Paidós.

- Couldry, N., & Mejias, U. A. (2019). *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.
- Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
- Floridi, L. (2024). *Ética de la inteligencia artificial* (J. Anta Pulido, Trad.). Herder Editorial. (Trabajo original publicado en 2023).
- 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
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2
- Molnar, C. (2025). *Interpretable machine learning: A guide for making black box models explainable* (3.ª ed.). https://christophm.github.io/interpretable-ml-book/
- Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21. https://doi.org/10.1109/MIS.2006.80
- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
- Parlamento Europeo y Consejo de la Unión Europea. (2024). Reglamento (UE) 2024/1689, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial. *Diario Oficial de la Unión Europea, L, 2024/1689*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
- Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
- UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence
- Véliz, C. (2021). *Privacidad es poder: Datos, vigilancia y libertad en la era digital* (A. Santos, Trad.). Debate. (Trabajo original publicado en 2020).
- Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195374049.001.0001
- World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. https://www.who.int/publications/i/item/9789240029200

Bibliografía Complementaria

- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820. https://doi.org/10.15779/Z38RV0D15J
- Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human–robot interaction. *Engaging Science, Technology, and Society, 5*, 40–60. https://doi.org/10.17351/ests2019.260
- European Data Protection Board. (2024). *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models*. https://www.edpb.europa.eu/documents/opinion-of-the-board-art-64/opinion-282024-on-certain-data-protection-aspects-related-to_en
- Fulgu, R. A., & Capraro, V. (2024). Surprising gender biases in GPT. *Computers in Human Behavior Reports, 16*, Article 100533. https://doi.org/10.1016/j.chbr.2024.100533
- Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2
- Gordon, J.-S., & Nyholm, S. (2021). Ethics of artificial intelligence. *Internet Encyclopedia of Philosophy*. https://iep.utm.edu/ethics-of-artificial-intelligence/
- Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt.
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics, 4*(37), Article eaay7120. https://doi.org/10.1126/scirobotics.aay7120
- Harris, C. E., Jr., Pritchard, M. S., Rabins, M. J., James, R. W., & Englehardt, E. E. (2019). *Engineering ethics: Concepts and cases* (6th ed.).

Cengage Learning.

- High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
- Hortal Alonso, A. (2002). *Ética general de las profesiones*. Desclée de Brouwer.
- Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press.
- Johnson, D. G., & Wetmore, J. M. (Eds.). (2021). *Technology and society: Building our sociotechnical future* (2nd ed.). MIT Press.
- Jonas, H. (1984). *The imperative of responsibility: In search of an ethics for the technological age*. University of Chicago Press.
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. En C. H. Papadimitriou (Ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Vol. 67, Article 43, pp. 43:1–43:23). Schloss Dagstuhl–Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43
- Larson, E. J. (2022). *El mito de la inteligencia artificial: Por qué las máquinas no pueden pensar como nosotros lo hacemos* (M. J. Krmpotić, Trad.). Shackleton Books. (Trabajo original publicado en 2021).
- Maldonado, C. E. (2024). *Inteligencia artificial y ética*. Ediciones Desde Abajo.
- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342
- Parlamento Europeo y Consejo de la Unión Europea. (2016). Reglamento (UE) 2016/679, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos. *Diario Oficial de la Unión Europea, L 119*, 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj
- Perrigo, B. (2023, 18 de enero). Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. *Time*. https://time.com/6247678/openai-chatgpt-kenya-workers/
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x
- Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, Article 15. https://doi.org/10.3389/frobt.2018.00015
- Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. En *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598
- Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. En *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Article 17, pp. 1–9). Association for Computing Machinery. https://doi.org/10.1145/3465416.3483305
- Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
- van de Poel, I., Nihlén Fahlquist, J., Doorn, N., Zwart, S., & Royakkers, L. (2012). The problem of many hands: Climate change as an example. *Science and Engineering Ethics, 18*(1), 49–67. https://doi.org/10.1007/s11948-011-9276-0
- Villas Olmeda, M., & Camacho Ibáñez, J. (2022). *Manual de ética aplicada en inteligencia artificial*. Anaya Multimedia.
- Winner, L. (1980). Do artifacts have politics? *Daedalus, 109*(1), 121–136.

A lo largo de la asignatura se podrá facilitar información bibliográfica actualizada adicional.