Universitat Autònoma de Barcelona

2023/2024

Ética

Código: 106559
Créditos ECTS: 6

Titulación	Tipo	Curso	Semestre
2504392 Inteligencia Artificial	FB	2	1

Contacto

Nombre: Maria Pilar Dellunde Clave

Correo electrónico: pilar.dellunde@uab.cat

Idiomas de los grupos

Puede consutarlo a través de este enlace. Para consultar el idioma necesitará introducir el CÓDIGO de la asignatura. Tenga en cuenta que la información es provisional hasta el 30 de noviembre del 2023.

Equipo docente

Miquel Domenech Argemi

Prerrequisitos

Esta asignatura no tiene prerrequisitos

Objetivos y contextualización

A partir de casos de estudio reales, este curso está diseñado para introducir a los estudiantes las implicaciones éticas y sociales de la inteligencia artificial (IA). La asignatura promoverá la incorporación de estrategias y el uso de herramientas para minimizar los riesgos éticos y desarrollar sistemas de IA en el marco de una IA responsable.

Competencias

- Actuar con responsabilidad ética y con respeto por los derechos y deberes fundamentales, la diversidad y los valores democráticos.
- Actuar en el ámbito de conocimiento propio evaluando las desigualdades por razón de sexo/género.
- Comunicarse de manera efectiva, tanto de forma oral como escrita, utilizando adecuadamente los recursos comunicativos necesarios y adaptándose a las características de la situación y de la audiencia.
- Concebir, diseñar, analizar e implementar agentes y sistemas ciber-físicos autónomos capaces de interactuar con otros agentes y/o personas en entornos abiertos, teniendo en cuenta las demandas y necesidades colectivas.

- Desarrollar pensamiento crítico para analizar de forma fundamentada y argumentada alternativas y propuestas tanto propias como ajenas.
- Identificar, analizar y evaluar el impacto ético y social, el contexto humano y cultural, y las implicaciones legales del desarrollo de aplicaciones de inteligencia artificial y de manipulación de datos en diferentes ámbitos.
- Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética.
- Trabajar de forma autónoma, con responsabilidad e iniciativa, planificando y gestionando el tiempo y los recursos disponibles, adaptándose a las situaciones imprevistas.

Resultados de aprendizaje

1. Analizar críticamente los principios, valores y procedimientos que rigen el ejercicio de la profesión.
2. Analizar las desigualdades por razón de sexo/género y los sesgos de género en el ámbito de conocimiento propio.
3. Comprender las implicaciones sociales, éticas y legales de la práctica profesional en IA.
4. Comunicarse de manera efectiva, tanto de forma oral como escrita, utilizando adecuadamente los recursos comunicativos necesarios y adaptándose a las características de la situación y de la audiencia.
5. Desarrollar pensamiento crítico para analizar de forma fundamentada y argumentada alternativas y propuestas tanto propias como ajenas.
6. Explicar el código deontológico, explícito o implícito, del ámbito de conocimiento propio.
7. Identificar las principales desigualdades y discriminaciones por razón de sexo/géneros presentes en la sociedad.
8. Identificar los sesgos sociales, culturales y económicos de los algoritmos.
9. Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética.
10. Saber analizar casos de aplicación de la IA desde un punto de vista ético, legal y social.
11. Ser capaz de incorporar los principios de la investigación e innovación responsable en los desarrollos basados en la IA.
12. Ser capaz de incorporar valores adecuados a las necesidades de las personas en el diseño de dispositivos dotados de IA.
13. Trabajar de forma autónoma, con responsabilidad e iniciativa, planificando y gestionando el tiempo y los recursos disponibles, adaptándose a las situaciones imprevistas.
14. Valorar cómo los estereotipos y los roles de género inciden en el ejercicio profesional.
15. Valorar las dificultades, los prejuicios y las discriminaciones que pueden incluir las acciones o proyectos, a corto o largo plazo, en relación con determinadas personas o colectivos.

Contenido

1. Introducción: ¿Por qué es importante la ética para los profesionales de la IA?
   1.1. Código Ético y de Conducta Profesional de la ACM.
   1.2. Marcos éticos (consecuencialismo, teoría de la justicia, ética de la virtud...).
   1.3. Principios éticos (equidad, responsabilidad, justicia, privacidad...). 1.4. Mediación tecnológica.
   1.5. Moral materializada.

2. Recogida de datos y privacidad
2.1. La importancia de la privacidad
2.2. Principales técnicas para la privacidad de datos (anonimato, cifrado, privacidad diferencial...).
2.3. Conflicto entre privacidad y otros valores (seguridad, transparencia...). 2.4. El uso de la agregación de datos para el modelado predictivo.
2.5. Privacidad más allá de los datos (en el contexto, por diseño,...).
2.6. "Bubble filters" y democracia.

3. Algoritmos, toma de decisiones y sesgos
3.1. Uso de algoritmos predictivos, con especial atención al sistema de justicia penal.
3.2. Definiciones técnicas de sesgo en resultados algorítmicos.
3.3. Discriminación algorítmica directa e indirecta.
3.4. Definición de equidad y métricas de equidad.
3.5. Directrices éticas para un uso ético y confiable de la inteligencia artificial: AI-Fairness Toolkits.
3.6. Conflicto entre precisión predictiva y otros valores (equidad, transparencia...).
3.7. Representación del conocimiento normativo y ético en IA.

4. Sistemas autónomos y explicabilidad
4.1. El impacto sobre la responsabilidad y la rendición de cuentas en los sistemas autónomos, estudio del caso de los vehículos autónomos.
4.2. La importancia de las buenas explicaciones en los sistemas de IA.
4.3. Herramientas para la explicación.

5. RRI y IA
5.1. ¿Qué la RRI?
5.2. RRI aplicada a la IA.

6. Ética y robótica
6.1. Robots y sociedad.
6.2. Retos éticos en robótica.
6.3. Robots de cuidado/killer robots.

Metodología

La orientación del curso es predominantemente práctica. En general, cada clase comenzará con la presentación de un caso de estudio real, que dará lugar a una discusión grupal. A continuación, se introducirán y explicarán los conceptos, métodos o sistemas de IA relacionados con los retos éticos planteados por el caso de estudio. Por último, el alumnado realizará prácticas individuales o grupales para reforzar el aprendizaje del contenido de la clase. En algunas sesiones se reserva tiempo para repasar y corregir estas prácticas. Algunas clases consistirán en visitas a centros de investigación en IA.

Nota: se reservarán 15 minutos de una clase dentro del calendario establecido por el centro o por la titulación para que el alumnado rellene las encuestas de evaluación de la actuación del profesorado y de evaluación de la asignatura o módulo.

Actividades

Título	Horas	ECTS	Resultados de aprendizaje
Tipo: Dirigidas			
Asistencia a clase y participación activa	30	1,2	3, 4, 5, 9, 10, 11, 12, 13
Casos de estudios	50	2	1, 3, 5, 8, 7, 11, 14
Prácticas y ejercicios	50	2	2, 5, 6, 7, 9, 10, 12, 13, 15

Evaluación

La evaluación se puede llevar a cabo de dos formas.

Evaluación continua. Se centrará principalmente en la realización de ejercicios prácticos durante la clase. El alumnado deberá completar un total de 10 prácticas, incluyendo tareas individuales y grupales. La nota final de la asignatura se determinará en función del rendimiento en estos ejercicios prácticos. El alumnado debe presentar al menos 7 prácticas para ser evaluado mediante esta modalidad. En caso contrario, el alumnado no habrá superado la evaluación continua y, si cumple las condiciones pertinentes, deberá presentarse a la recuperación (véase el apartado Recuperación).

Evaluación única. El alumnado deberá presentar ejercicios prácticos.

Recuperación. La prueba de recuperación será un examen final. Para poder presentarse a la recuperación, el alumnado deberá haber sido previamente evaluado en un conjunto de actividades cuyo peso equivalga a un mínimo de 2/3 partes de la calificación total (evaluación continua) o presentar todas las pruebas prácticas (evaluación única).

Al realizar cada actividad de evaluación, el profesorado informará al alumnado (a Moodle) de los procedimientos a seguir para revisar todas las calificaciones concedidas y la fecha en que se realizará esta revisión.

Los estudiantes conseguirán una calificación de "No evaluable" a menos que hayan presentado más de 1/3 de los ítems de evaluación.

En caso de que un/a alumno/a cometa alguna irregularidad que pueda comportar una variación significativa de la nota otorgada a una actividad de evaluación, se le dará un cero por esta actividad, con independencia de cualquier expediente disciplinario que se pueda abrir. En caso de que haya varias irregularidades en las actividades de evaluación de una misma asignatura, el estudiante recibirá un cero como nota final de esta asignatura.

En caso de que no se puedan llevar a cabo pruebas o exámenes presenciales, se adaptarán a un formato online puesto a disposición a través de las herramientas virtuales de la UAB (se mantendrá la ponderación original). Los deberes, actividades y participación en clase se realizarán a través de foros, wikis o debates en equipos, etc. El profesorado se asegurará de que el alumnado pueda acceder a estas herramientas virtuales u ofrecerá alternativas factibles.

Actividades de evaluación continuada

Título	Peso	Horas	ECTS	Resultados de aprendizaje
Prácticas de los casos de estudio	100%	20	0,8	1, 2, 3, 4, 5, 6, 8, 7, 9, 10, 11, 12, 13, 14, 15

Bibliografía

Margaret A. Boden, *AI: Its nature and future*, Oxford University Press, 2016.

Mark Coeckelberg, *AI Ethics*, The MIT Press, 2020.

Crawford, K. (2021). The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI." Berkman Klein Center for Internet & Society, 2020.

Mehrabi N., Morstatter F., Saxena N-, Lerman K., Galstyan A. *A Survey on Bias and Fairness in Machine Learning.* Association for Computing Machinery Surveys, (2021), 54(6)

Sparrow, R. (2007) 'Killer robots', Journal of Applied Philosophy, 24(1), pp. 62-77.

Vallès-Peris N and Domènech M (2020) Roboticists' Imaginaries of Robots for Care: The Radical Imaginary as a Tool for an Ethical Discussion. Engineering Studies, 12 (3): 156-176.

Vallès-Peris, N., Domènech, M. (2021) Caring in the in-between: a proposal to introduce responsible AI and robotics to healthcare. AI & Society.

van de Poel, I. (2020) 'Embedding Values in Artificial Intelligence (AI) Systems', Minds and Machines, 30(3), pp. 385-409.

van Wynsberghe, A. (2013) 'Designing Robots for Care: Care Centered Value-Sensitive Design', Science and Engineering Ethics, 19(2), pp. 407-433.

Verbeek, P.-P. (2006) 'Materializing Morality: Design Ethics and Technological Mediation', Science, Technology & Human Values, 31(3), pp. 361-380.

Software

No habrá software.