Universitat Autònoma de Barcelona

2023/2024

Ethics

Code: 106559
ECTS Credits: 6

Degree	Type	Year	Semester
2504392 Artificial Intelligence	FB	2	1

Contact

Name: Maria Pilar Dellunde Clave

Email: pilar.dellunde@uab.cat

Teaching groups languages

You can check it through this link. To consult the language you will need to enter the CODE of the subject.
Please note that this information is provisional until 30 November 2023.

Teachers

Miquel Domenech Argemi

Prerequisites

No prerequisites

Objectives and Contextualisation

Drawing from real-world case studies, this course is designed to instill in students awareness of the ethical and societal implications of artificial intelligence (AI). It provides comprehensive instruction on incorporating strategies and utilizing tools to minimize ethical risks while fostering the development of AI systems within the framework of responsible AI.

Competences

- Act with ethical responsibility and respect for fundamental rights and duties, diversity and democratic values.
- Act within the field of knowledge by evaluating sex/gender inequalities.
- Communicate effectively, both orally and in writing, adequately using the necessary communicative resources and adapting to the characteristics of the situation and the audience.
- Conceive, design, analyse and implement autonomous cyber-physical agents and systems capable of interacting with other agents and/or people in open environments, taking into account collective demands and needs.

- Develop critical thinking to analyse alternatives and proposals, both one's own and those of others, in a well-founded and argued manner.
- Identify, analyse and evaluate the ethical and social impact, the human and cultural context, and the legal implications of the development of artificial intelligence and data manipulation applications in different fields.
- Students must be capable of collecting and interpreting relevant data (usually within their area of study) in order to make statements that reflect social, scientific or ethical relevant issues.
- Work independently, with responsibility and initiative, planning and managing time and available resources, and adapting to unforeseen situations.

Learning Outcomes

1. Analyse AI application cases from an ethical, legal and social point of view.
2. Analyse sex/gender inequalities and gender bias in the field of knowledge.
3. Communicate effectively, both orally and in writing, adequately using the necessary communicative resources and adapting to the characteristics of the situation and the audience.
4. Critically analyse the principles, values and procedures that govern the practice of the profession.
5. Develop critical thinking to analyse alternatives and proposals, both one's own and those of others, in a well-founded and argued manner.
6. Evaluate how stereotypes and gender roles affect the professional exercise.
7. Evaluate the difficulties, prejudices and discriminations that can be found in actions or projects, in a short or long term, in relation to certain people or groups.
8. Explain the code of ethics, explicit or implicit, that pertains to the field of knowledge.
9. Identify the main sex- and gender-based inequalities and discrimination present in society today.
10. Identify the social, cultural and economic biases of certain algorithms.
11. Incorporate the principles of responsible research and innovation in AI-based developments.
12. Incorporate values appropriate to people's needs when designing AI-enabled devices.
13. Students must be capable of collecting and interpreting relevant data (usually within their area of study) in order to make statements that reflect social, scientific or ethical relevant issues.
14. Understand the social, ethical and legal implications of professional AI practice.
15. Work independently, with responsibility and initiative, planning and managing time and available resources, and adapting to unforeseen situations.

Content

1. Introduction: Why should AI professionals study ethics?
1.1. ACM Code of Ethics and Professional Conduct.
1.2. Ethical frameworks (consequentialism, theory of justice, virtue ethics...).
1.3. Ethical principles (equity, accountability, justice, privacy...).
1.4. Technological mediation.
1.5. Materialized morality.

2. Data Collection and Privacy

2.1. Basis of the importance of privacy
2.2. Technical approaches to data privacy (anonymization, encryption, differential privacy...).
2.3. Trade-off between privacy and other values (security, transparency...).
2.4. The use of data aggregation for predictive modeling.
2.5. Privacy beyond data (in context, by design,...).
2.6. Flilter "bubbles" and democracy.

3. Algorithm and decision-making and bias

3.1. Use of predictive algorithms, with focus on the criminal justice system.
3.2. Technical definitions of bias in algorithmic results.
3.3. Direct and indirect algorithmic discrimination.

3.4. Definition of fairness and fairness metrics.
3.5. Ethics guidelines for trustworthy AI:AI-Fairness Toolkits.
3.6. Trade-offs between predictive accuracy and competing values (fairness, transparency...).
3.7. Normative and ethical knowledge representation in AI.

4. Autonomous Systems and Explainability

4.1. The impact on liability, responsibility, and accountability in autonomous systems, focusing on the autonomous vehicles' case.
4.2. The importance of good explanations in AI systems.
4.3. Tools for explainability.

5. Responsible Research and Innovation (RRI) and AI

5.1. What is RRI?
5.2. RRI applied to AI.

6. Ethics and Robotics

6.1. Robots and society.
6.2. Ethical concerns in robotics.
6.3. Care robots/killer robots.

Methodology

The course's orientation is predominantly practical. Each class will typically commence with the presentation of a real-world case study, fostering a subsequent group discussion. Following that, concepts, methods, or AI systems related to the ethical concerns raised by the case will be introduced and explained. Finally, students will engage in individual or group practices to reinforce their learning of the lecture. In some classes, time will be kept for reviewing and correcting these practices. Few classes will consist of visits to AI research centers.

Annotation: Within the schedule set by the centre or degree programme, 15 minutes of one class will be reserved for students to evaluate their lecturers and their courses or modules through questionnaires.

Activities

Title	Hours	ECTS	Learning Outcomes
Type: Directed			
Case studies	50	2	4, 14, 5, 10, 9, 11, 6
Lesson attendance and active participation	30	1.2	14, 3, 5, 13, 1, 11, 12, 15
Practices and exercise	50	2	2, 5, 8, 9, 13, 1, 12, 15, 7

Assessment

The assessment can be carried out in two ways:

Continuous assessment. It will be ongoing and primarily focused on completing practical exercises during class. Students are required to complete a total of 10 practices, including both individual and group assignments. The course's final grade will be determined based on the performance in these practical

exercises. Students must present at least 7 practices to be evaluated in this continuous assessment way. Otherwise, the student will not have passed the continuous assessment and if they meet the conditions, they will have to present for recovery (see the Recovery section).

Single assessment. Students will have to submit practical exercises.

Recovery: The recovery test is a final exam. To participate in recovery, students must have previously been evaluated in a set of activities whose weight is equivalent to a minimum of 2/3 parts of the total qualification (continuous evaluation) or submit all the practical exercices tests (single assessment). The same recovery system will be applied for the continuous assessment.

On carrying out each evaluation activity, lecturers will inform students (on Moodle) of the procedures to be followed for reviewing all grades awarded, and the date on which such a review will take place.

Students will obtain a "No avaluable" course grade unless they have submitted more than 1/3 of the assessment items.

In the event of a student committing any irregularity that may lead to a significant variation in the grade awarded to an assessment activity, the student will be given a zero for this activity, regardless of any disciplinary process that may take place. In the event of several irregularities in assessment activities of thesame subject, the student will be given a zero as the final grade for this subject.

In the event that tests or exams cannot be taken onsite, they will be adapted to an online format made available through the UAB's virtual tools (original weighting will be maintained). Homework, activities, and class participation will be carried out through forums, wikis, and/or discussions on Teams, etc. Lecturers will ensure that students are able to access these virtual tools, or will offer them feasible alternatives.

Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Practices of the study cases.	100%	20	0.8	4, 2, 14, 3, 5, 8, 10, 9, 13, 1, 11, 12, 15, 6, 7

Bibliography

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

There will be no software.