UAB
Universitat Autònoma de Barcelona

Ètica

Codi: 106559
Crèdits: 6

2024/2025

Titulació	Tipus	Curs
2504392 Intel·ligència Artificial / Artificial Intelligence	FB	2

Professor/a de contacte

Nom: Vicente Costa Bueno

Correu electrònic: vicente.costa@uab.cat

Equip docent

Nuria Valles Peris

Idiomes dels grups

Podeu consultar aquesta informació al final del document.

Prerequisits

Aquesta assignatura no té prerequisits

Objectius

A partir casos d'estudi reals, aquest curs està dissenyat per introduir als estudiants les implicacions ètiques i socials de la intel·ligència artificial (IA). L'assignatura promourà la incorporació d'estratègies i l'ús d'eines per minimitzar els riscos ètics alhora desenvolupar sistemes d'IA en el marc d'una IA responsable.

Competències

- Actuar amb responsabilitat ètica i amb respecte pels drets i deures fonamentals, la diversitat i els valors democràtics.
- Actuar en làmbit de coneixement propi avaluant les desigualtats per raó de sexe o gènere.
- Comunicar-se de manera efectiva, tant oralment com per escrit, utilitzant adequadament els recursos comunicatius necessaris i adaptant-se a les característiques de la situació i de laudiència.
- Concebre, dissenyar, analitzar i implementar agents i sistemes ciberfísics autònoms capaços dinteractuar amb altres agents o persones en entorns oberts, tenint en compte les demandes i necessitats col·lectives.
- Desenvolupar pensament crític per analitzar de manera fonamentada i argumentada alternatives i propostes tant pròpies com alienes.

- Identificar, analitzar i avaluar limpacte ètic i social, el context humà i cultural i les implicacions legals del desenvolupament daplicacions dintel·ligència artificial i de manipulació de dades en diferents àmbits.
- Que els estudiants tinguin la capacitat de reunir i interpretar dades rellevants (normalment dins de la seva àrea d'estudi) per emetre judicis que incloguin una reflexió sobre temes destacats d'índole social, científica o ètica.
- Treballar de manera autònoma, amb responsabilitat i iniciativa, planificant i gestionant el temps i els recursos disponibles i adaptant-se a les situacions imprevistes.

Resultats d'aprenentatge

1. Analitzar críticament els principis, valors i procediments que regeixen lexercici de la professió.
2. Analitzar les desigualtats per raó de sexe o gènere i els biaixos de gènere en làmbit de coneixement propi.
3. Comprendre les implicacions socials, ètiques i legals de la pràctica professional en IA.
4. Comunicar-se de manera efectiva, tant oralment com per escrit, utilitzant adequadament els recursos comunicatius necessaris i adaptant-se a les característiques de la situació i de laudiència.
5. Desenvolupar pensament crític per analitzar de manera fonamentada i argumentada alternatives i propostes tant pròpies com alienes.
6. Explicar el codi deontològic, explícit o implícit, de làmbit de coneixement propi.
7. Identificar els biaixos socials, culturals i econòmics dels algoritmes.
8. Identificar les principals desigualtats i discriminacions per raó de sexe o gènere presents en la societat.
9. Que els estudiants tinguin la capacitat de reunir i interpretar dades rellevants (normalment dins de la seva àrea d'estudi) per emetre judicis que incloguin una reflexió sobre temes destacats d'índole social, científica o ètica.
10. Saber analitzar casos daplicació de la IA des dun punt de vista ètic, legal i social.
11. Saber treballar en equip en el disseny de projectes interdisciplinaris. Ser capaç de col·laborar amb no professionals i professionals daltres sectors.
12. Ser capaç dincorporar els principis de la recerca i innovació responsable en els desenvolupaments basats en la IA.
13. Ser capaç dincorporar valors adequats a les necessitats de les persones en el disseny de dispositius dotats dIA.
14. Treballar de manera autònoma, amb responsabilitat i iniciativa, planificant i gestionant el temps i els recursos disponibles i adaptant-se a les situacions imprevistes.
15. Valorar com els estereotips i els rols de gènere incideixen en lexercici professional.
16. Valorar les dificultats, els prejudicis i les discriminacions que poden incloure les accions o projectes, a curt o llarg termini, en relació amb determinades persones o col·lectius.

Continguts

Part I: Aspectes ètico-polític de la intel·ligència artificial

1. Introducció: per què són rellevants els aspectes polítics i socials de la intel·ligència artitifial
1.1. Teoria de la mediació tecnològica
1.2. Narrativa al voltant de la intel·ligència artificial i determinisme tecnològic
1.3 Innovació i recerca responsable (RRI)

2. Ètica i robòtica
2.1. Robots i societat.
2.2. Reptes ètics en robòtica.
2.3. Exemples aplicats de robòtica en l'àmbit quotidià

Part II: Aspectes ètics de la intel·ligència artificial

3. Introducció: Per què els professionals de la IA haurien d'estudiar ètica?
3.1. Codi Ètic i de Conducta Professional de l'ACM.
3.2. Marcs ètics (conseqüencialisme, teoria de la justícia, ètica de la virtut...).
3.3. Principis ètics (equitat, responsabilitat, justícia, privacitat...).

4. Recollida de dades i privadesa
4.1. La importància de la privadesa
4.2. Principals tècniques per a la privadesa de dades (anonimat, xifrat, privadesa diferencial...).
4.3. Privadesa més enllà de les dades (en el context, per disseny...).

5. Algorismes, presa de decisions i biaixos
5.1. Definicions tècniques de biaix en resultats algorísmics.
5.2. Discriminació algorísmica directa i indirecta.
5.3. Definició d'equitat i mètriques d'equitat.
5.4. Representació del coneixement normatiu i ètic en IA.
5.5. Directrius ètiques per a una IA fiable: AI-Fairness Toolkits.

6. Explicabilitat
6.1. L'impacte sobre la responsabilitat i la rendició de comptes en els sistemes autònoms, centrant-nos en el cas dels vehicles autònoms.
6.2. La importància de les bones explicacions en els sistemes d'IA.
6.3. Eines per a avaluar l'explicabilitat.

Activitats formatives i Metodologia

Títol	Hores	ECTS	Resultats d'aprenentatge
Tipus: Dirigides			
Asistència i participació activa a classe	30	1,2	10, 4, 5, 12, 13, 9, 3, 14
Casos d'estudis	50	2	1, 5, 15, 8, 7, 12, 3
Pràctiques i exercicis	50	2	10, 2, 5, 16, 6, 8, 13, 9, 14

L'orientació del curs és predominantment pràctica. Cada classe començarà generalment amb la presentació d'un cas d'estudi real, que donarà lloc a una discussió grupal. A continuació, s'introduiran i explicaran els conceptes, els mètodes o els sistemes d'IA relacionats amb els reptes ètics plantejats pel cas d'estudi. Finalment, l'alumnat farà pràctiques individuals o grupals per reforçar el seu aprenentatge del contingut de la classe. En algunes sessions es reserva temps per repassar i corregir aquestes pràctiques. Algunes classes consistiran en visites a centres de recerca d'IA.

Nota: es reservaran 15 minuts d'una classe, dins del calendari establert pel centre/titulació, per a la complementació per part de l'alumnat de les enquestes d'avaluació de l'actuació del professorat i d'avaluació de l'assignatura/mòdul.

Avaluació

Activitats d'avaluació continuada

Títol	Pes	Hores	ECTS	Resultats d'aprenentatge
Pràctica avaluativa 1	34%	7	0,28	10, 2, 4, 1, 5, 15, 16, 6, 8, 7, 12, 13, 9, 3, 14
Pràctica avaluativa 2	33%	7	0,28	10, 4, 1, 5, 16, 6, 8, 7, 9, 3, 14
Pràctica avaluativa 3	33%	6	0,24	4, 5, 12, 13, 9, 11, 14

L'avaluació es pot dur a terme de les dues maneres descrites a continuació.

A v a l u a c i ó c o n t i n u a d a .
Els estudiants han de realitzar tres tasques d'avaluació: una corresponent a la Part I i dues relacionades amb la Part II. La tasca corresponent a la Part I serà un examen a fer a classe (presumiblement a l'octubre). La segona tasca serà una prova d'avaluació a fer a classe, que inclourà preguntes curtes sobre la Part II i una anàlisi d'un cas real de l'aplicació d'un sistema d'IA. La tercera tasca consistirà en l'ús de kits d'eines d'IA per avaluar i discutir diferents mètriques relacionades amb l'ètica d'un sistema d'IA.
Per ser elegible per a l'avaluació continuada, l'estudiant ha d'haver completat les tres tasques d'avaluació. A més, per aprovar l'assignatura, cal obtenir almenys una nota de 5 en les tres tasques d'avaluació. En cas contrari, l'estudiant haurà de fer la recuperació (vegeu la secció de Recuperació).
La nota final de l'assignatura en aquesta modalitat es determinarà com la mitjana de les tres tasques d'avaluació.

A v a l u a c i ó ú n i c a .
L'estudiant farà un examen final al gener. L'examen contindrà tres parts i, per aprovar l'assignatura, l'estudiant haurà d'obtenir almenys una nota de 5 (sobre 10) en cadascuna de les parts. La nota final de l'assignatura en aquesta modalitat serà la mitjana de les notes obtingudes en cadascuna de les parts de l'examen.

R e c u p e r a c i ó .
Per ser elegible per a la recuperació, els estudiants han d'haver completat les tres tasques d'avaluació (avaluació continuada) o haver fet l'examen al gener (avaluació única). Només es farà un examen final de recuperació individual. Per aprovar l'assignatura en aquesta modalitat, la nota de l'examen final de recuperació ha de ser igual o superior a 5. La nota final serà la nota de l'examen final de recuperació.

Si l'estudiant està matriculat en l'avaluació continuada, només cal que recuperi les parts de l'examen corresponents a les seves tasques d'avaluació suspeses.

En dur a terme cada activitat d'avaluació, els professors informaran els estudiants (a Moodle) dels procediments que s'han de seguir per revisar totes les notes atorgades i la data en què es farà aquesta revisió.

En cas que un estudiant cometi alguna irregularitat que pugui comportar una variació significativa en la nota atorgada a una activitat d'avaluació, l'estudiant rebrà un zero en aquesta activitat, independentment de qualsevol procés disciplinari que es pugui dur a terme. En cas de diverses irregularitats en les activitats d'avaluació de la mateixa assignatura, l'estudiant rebrà un zero com a nota final d'aquesta assignatura.

En el cas que les proves o exàmens no es puguin fer presencialment, s'adaptaran a un format en línia posat a disposició a través de les eines virtuals de la UAB (es mantindrà la ponderació original). Els deures, activitats i la participació a classe es realitzaran mitjançant fòrums, wikis i/o discussions a Teams, etc. Els professors vetllaran perquè els estudiants puguin accedir a aquestes eines virtuals o els oferiran alternatives factibles.

Bibliografia

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.

Programari

Per determinar (Part II).

Llista d'idiomes

Nom	Grup	Idioma	Semestre	Torn
(PAUL) Pràctiques d'aula	711	Anglès	primer quadrimestre	tarda
(TE) Teoria	71	Anglès	primer quadrimestre	tarda