

Introducció a la ciència de dades.

PAC3: Detectar el dilema ètic i els riscos de no governar les dades

Presentació

Avaluació contínua (AC) dels continguts lectius corresponents:

- **Bloc 4: Ètica a l'àmbit de la ciència de dades.**
 - o Ètica i Big Data.
 - o Què he de fer?
 - o El bé, el que és just o el que és correcte? Teories.
 - o Conceptes fonamentals per a una ètica aplicada al Big Data.
 - o La presa de decisions ètiques.
- **Bloc 5: "Governar" les dades a les organitzacions.**
 - o La dada com a actiu de valor per a l'organització.
 - o Una altra perspectiva del cicle de vida de les dades.
 - o Govern de la dada.
 - o Components i maduresa del data governance.

Objectius i competències

- Entendre els conceptes utilitzats habitualment i relacionats amb el context de la ciència de dades.
- Entendre les fases del cicle de vida de la dada.
- Entendre la problemàtica ètica en un projecte de ciència de dades.
- Conèixer els elements bàsics de govern de dades.

Críteris d'avaluació generals de la PAC.

- Aportació d'alguna referència externa que complementi o sustenti els raonaments que s'hi exposen.
- No ometre cap de les preguntes de cada exercici.
- Respecteu l'extensió de paraules assenyalada a cada enunciat.
- Claredat en les respostes i els raonaments.
- Capacitat de síntesi.
- Originalitat.

Críteris de valoració dels exercicis

- Indicats a l'enunciat.

Format i data de lliurament

- Les respostes es lliuraran en **format PDF**.
- S'usarà el **Registre d'avaluació continuada** per pujar les respostes.
- El document **no ha d'incloure l'enunciat**, només les respostes.
- El document ha d'estar **estructurat** i el text en un color que en faciliti la lectura (negre o blau fosc).
- Cal assegurar **que el nom de l'estudiant consti al document/s** (per exemple: a la coberta i al peu de pàgina) i **que els documents s'han enviat/carregat correctament**.
- La data **màxima** per lliurar les respostes és el **09 de gener del 2022, a les 23:59h**. No es corregiran les PAC que no compleixin aquest requisit, excepte en casos de força major i degudament justificats.

Declaració de treball original (no plagi) de l'estudiant

Jo, *NomEstudiant*, declaro que per fer aquesta entrega... *(completeu la frase amb les vostres pròpies paraules)*

Pregunta 1 (30% puntuació)

Enunciat

Al llarg del mòdul 4 hem pogut veure la importància de les decisions ètiques dins d'una organització. La cultura d'una organització permet fer explícits els valors que els membres d'aquesta organització sostenen en comú. Això juntament amb la visió comuna que es tingui de la seva activitat influirà de manera ineludible en la presa de decisions sobre com es tracten conceptes tan rellevants com la privadesa, la identitat i la gestió de la propietat de les dades dels seus usuaris. En base a això, les organitzacions que tinguin en compte l'ètica com a factor fonamental en la construcció de valor estratègic formaran un rol decisiu en el lideratge d'estratègies ètiques. Considerant això:

1. Quins procediments poden dur a terme les organitzacions per facilitar una alineació estratègica on l'ètica sigui present? Indica un exemple pràctic en aquest sentit.
2. Indica breument per què un organisme o entitat que utilitzi la Ciència de dades com a presa de decisions ha d'adoptar aquests valors a la seva cultura organitzativa. Indica un exemple pràctic en aquest sentit.
3. Assenyala els punts de decisió ètica i termes més importants utilitzats per explorar les relacions entre valors i accions a les organitzacions. Indiqueu exemples pràctics d'aquests termes de decisió ètica.

Criteris d'avaluació

- Les preguntes 1 i 2 es valoraran **0,25 punts** per la resposta a la pregunta i **0,25 punts** per exemple aportat amb un màxim de **0,5 punts per l'apartat**.
- La pregunta 3 es valorarà **0.2 punts** com a màxim per cada exemple aportat i **0.4 punts** per l'explicació teòrica donant un màxim de **2 punts** per l'apartat. Haurà de tenir una extensió màxima de **500 paraules**. Es valorarà la capacitat d'identificar exemples pràctics per a cada terme.
- Totes les respostes han d'estar **desenvolupades i argumentades**.

Pregunta 2 (35% puntuació)

Enunciat

En els darrers anys, la utilització de dades privades per part de les companyies ha anat augmentant alhora que la preocupació dels usuaris. Aquesta gestió de les dades ens porta en primer lloc, a pensar en la propietat de les dades, i per tant apareixen preguntes com a qui pertanyen les dades, qui les genera o qui les emmagatzema. Aquest assumpte és de vital importància i en aquest dilema social l'ètica juga un paper fonamental a l'ara d'anticipar-se davant marcs que mai no estan regulats íntegrament i que per tant hem d'avaluar per saber cap a on posicionar-se per no perjudicar a cap de les parts.

A causa de la gran velocitat amb què el camp de la Intel·ligència Artificial està creixent ja que és gairebé impossible adequar la legislació vigent, apareixen casos com el següent:

Les companyies *GitHub* i *OpenAI* van llançar una nova eina basada en Intel·ligència Artificial anomenada *Copilot*, que permet auto completar fragments de codi de programació. Segons els seus creadors, analitza el codi ja escrit i genera codi nou, incloent-hi referències al codi existent. Tasques com importar tweets des de l'API de Twitter o dibuixar una gràfica, ja poden ser codificades amb aquesta eina.

- <https://www.theverge.com/2021/6/29/22555777/github-openai-ai-tool-autocomplete-code>

Tot i això, recentment s'ha descobert que aquesta eina ha estat entrenada amb centenars de repositoris de codi obert. Això va obrir un debat sobre quines llicències han d'estar subjectes al programari resultant, i què comporta que l'eina generi codi creat prèviament per un altre desenvolupador. Això dona lloc a un debat de fins a quin punt el codi generat per una IA és simplement una "referència" i fins a quin punt és una feina original. D'altra banda, s'ha descobert que l'eina té certes limitacions al voltant de la generació de codi de manera esbiaixada produint resultats indesitjables.

- <https://www.xataka.com/robotica-e-ia/github-copilot-polemica-copyright-debate-ia-infringe-drets-autor-codigo-escrit-altres-programadors>
- <https://www.pwvconsultants.com/blog/questions-around-bias-legalities-in-githubs-copilot/>

En base a aquest cas particular, es demana que facis una reflexió crítica personal (entre 500 i 1000 paraules) que respongui a les qüestions següents:

1. Quins dilemes ètics planteja una eina com *Copilot*? Quina problemàtica comporta la utilització de codis d'*open-source*? Quin paper juga en aquest tipus de tecnologies el tractament del biaix i per què és important mitigar-lo?

2. Pel que fa a les lleis de propietat intel·lectual, cal considerar la generació automàtica per part d'aquesta eina com a autoria pròpia o incorre en una violació dels drets dels autors?
3. Si un desenvolupador no vol que el codi sigui utilitzat en l'entrenament d'aquesta eina, com s'hauria de gestionar aquesta relació? Creus que les administracions governamentals han de prendre partit en aquest tipus d'assumptes?

Criteris d'avaluació

- La pregunta 1 i 2 estan puntuades amb un **1 punt**.
- La pregunta 3 està puntuada amb **1.5 punts**.
- Poseu en pràctica els conceptes fonamentals d'una ètica aplicada al Big Data i als mitjans socials.
- Desenvolupa una reflexió sintetitzada, però ben argumentada.
- Utilitza referències bibliogràfiques per al desenvolupament dels arguments.

Pregunta 3 (35% puntuació)

Enunciat

La governança de dades és un camp molt ampli que afecta multitud d'àmbits. Un és la mobilitat i en particular els cotxes autònoms, un sector on la col·laboració entre ens públics i privats és crucial per al desenvolupament d'un marc legislatiu comú.

- <https://www.jipitec.eu/issues/jipitec-9-3-2018/4807>
- https://onlinepubs.trb.org/onlinepubs/AVSMForum/products/1-NCHRP_Data_Sharing_and_Governance_Final_10-28-20v2.pdf

Aquest repte fa del govern de la dada un exercici coral on els diferents actors han de prendre partida. Per això, multitud d'organitzacions estan treballant en el desenvolupament de plataformes i normes per a la compartició de dades de mobilitat, el gran impacte de les quals tingui en compte la perspectiva operacional i regulatòria aplicada a la privadesa dels usuaris.

Per exemple, a la *Royal Society of London* van plantejar al 2019 temes com, des de la perspectiva de la governança de dades, sobre qui ha de gestionar i emmagatzemar les dades en un accident de trànsit, per baixar les ràtios d'accidents i alhora preservar la privadesa dels conductors.

- <https://royalsociety.org/blog/2019/10/moving-from-data-governance-principles-to-practice-auto-insurance/>

D'altra banda, podem veure el cas de com el govern holandès ha aconseguit descodificar les dades de conducció dels automòbils Tesla, al·legant que les dades contenen una gran quantitat d'informació per a investigadors forenses i analistes d'accidents de trànsit i poden ajudar a clarificar-la investigació.

- <https://www.engadget.com/a-dutch-government-lab-has-decoded-teslas-driving-data-for-the-first-time-085709633.html>

En base a allò assenyalat anteriorment i recolzat en les referències i els mòduls teòrics d'aquesta part de l'assignatura, contesta:

1. Des d'una perspectiva de la governança de les dades, com aplicaries el “cicle de vida de les dades” al món dels cotxes autònoms? Quina mena de marc seria l'apropiat per a usuaris, agents reguladors i indústria?
2. De manera teòrica, suggereix una proposta de les fases de govern de dades vinculades al sector dels cotxes autònoms. Identifica'n els components en el cas d'ús que cal tractar.
3. Com es complementen la gestió de les dades i la governança? Creieu que cal la governança de dades en un cas com el dels cotxes autònoms? Quina proposta creieu que seria la més idònia?

Criteris d'avaluació

- La pregunta 1 i 2 es valoraran amb **1 punt** com a màxim cadascuna.
- La pregunta 3 es valorarà amb **1.5 punts** com a màxim.
- Entendre què és el *data governance* i la importància que tenen en el cas d'ús.
- Conèixer el procés de *data governance* i els components que els integren.
- Relacionar les bones pràctiques del *data governance* amb els problemes que es mitiguen.
- Cada resposta ha de tenir una extensió compresa entre 150 i 300 paraules