hogeschool Windesheim

Machine Learning Ethics

hogeschool Windesheim

With great power comes great responsibility

Leeruitkomst

De student herkent en benoemt diverse morele vragen bij een praktijkopdracht binnen een machine learning context en onderbouwt vervolgens een moreel oordeel met argumenten vanuit verschillende ethische invalshoeken.

Belangrijk
(anders dan andere semesters)

Ethiek is NIET een aparte module, maar een onderdeel van het semester (in totaal 30 EC)

Als je de leeruitkomst van ethiek niet behaalt, haal je dus het hele semester niet (30 EC)

Gebruik ChatGPT en soortgelijke hulpmiddelen

Voorwaarden voor gebruik:

- is dit helder vermeld (met prompt en antwoord)?

én

- kun je nog steeds zelf de leeruitkomst voldoende aantonen?

Opbouw workshops

Verdieping ten opzichte van ethiekworkshops bij Data Science

- Verdere inhoudelijke verdieping (o.a. begeleidingsethiek)
- Meer aandacht voor persoonlijke reflectie en
- Aanvaarden van morele verantwoordelijkheid

Morele verantwoordelijkheid

Een echte opdracht met een echt moreel oordeel

De morele goedkeuring van de opdracht en wijze van uitvoering en resulterend product ligt bij het projectteam van studenten,

dus niet (alleen) bij de opdrachtgever en/of bij de docenten…!

Overzicht workshops

Workshop 1	Workshop 2	Workshop 3	Workshop 4
• Herhaling basiskennis • Morele kwesties in Machine Learning	• Morele onderbouwing • Nastreven waarden m.b.v. Machine Learning	• Morele intuïtie en verantwoordelijkheid • Uitleg ‘function creep’	• Toepassing nuancering • Introductie begeleidingsethiek

Workshop 1

- Inleiding morele kwesties in Machine Learning
- Inhoud ethiek: herhaling basiskennis BET / DS, verdieping op tweezijdig argumenteren vanuit deontologie, utilisme en deugdethiek en reflectie op oordeel
- Morele uitgangspunten bij ML

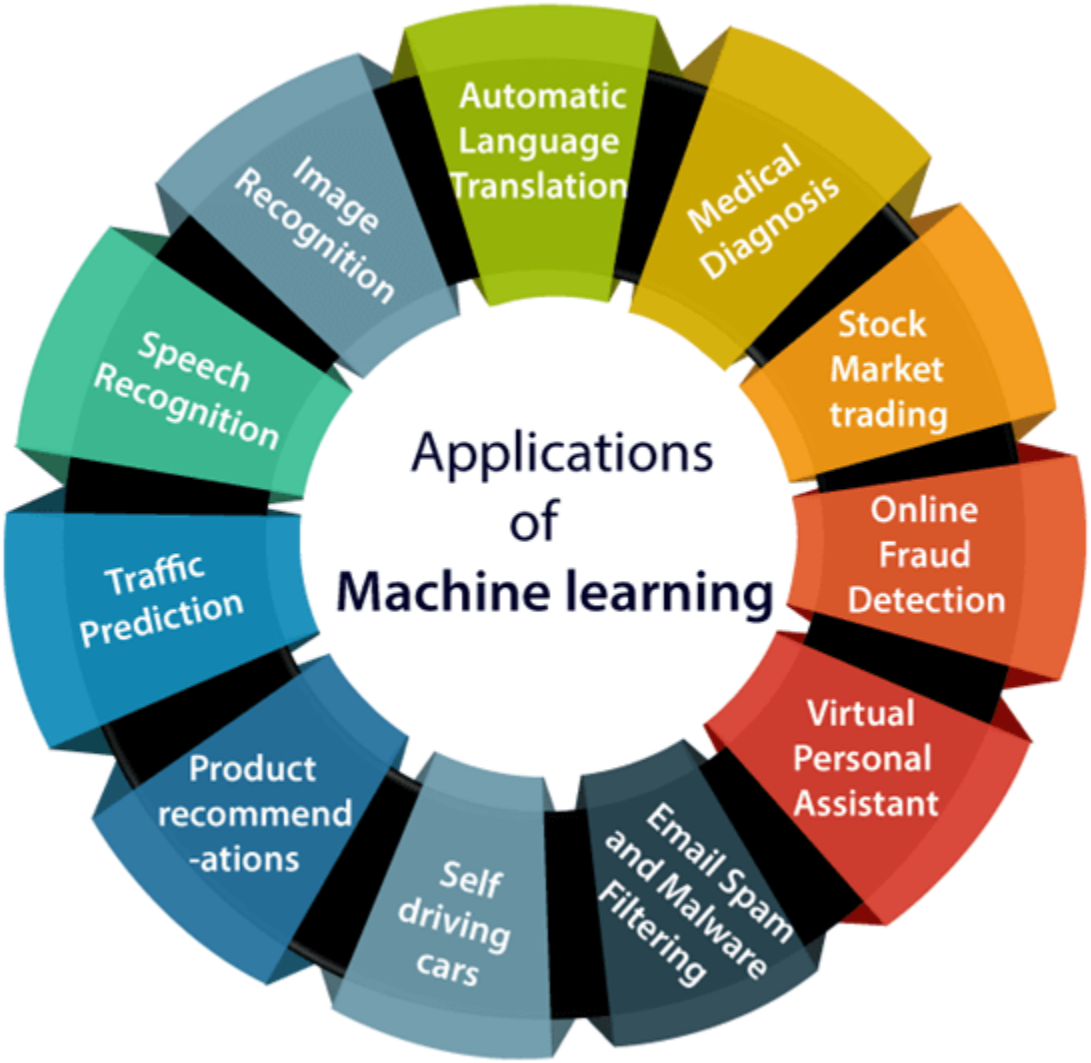
Denkvraag bij thema: Welke morele vragen kun je hierbij stellen?

hogeschool Windesheim

Machines maken betere keuzes dan mensen

Waar of niet waar?

Voorbeelden toepassing Machine Learning



Applications of Machine Learning – Javatpoint, geraadpleegd op 19 jan. 2022.

‘Amazon stopt met sollicitatierobot vanwege vrouwendiscriminatie

Kunstmatige intelligentie werkt vaak op basis van patroonherkenning: het systeem bekijkt eerst wat voor mensen het bedrijf in het verleden aannam, vindt daar patronen in en past die vervolgens ook toe op toekomstige kandidaten.’

Bron: nu.nl, 10 oktober 2018.

Voorbeeld

‘De algoritmes die de Belastingdienst in de toeslagenaffaire gebruikte, schenden mensenrechten. Die conclusie trekt Amnesty International (…).’

‘Fiscus gebruikte in toeslagenaffaire algoritmes die mensenrechten schenden’, nu.nl, 25-10-21.

Voorbeeld

‘Facebook heeft zijn excuses aangeboden nadat een algoritme van het bedrijf zwarte mannen in een video als primaten had aangemerkt.’

‘Facebook zegt sorry nadat zwarte mannen waren aangemerkt als ‘primaten’’, nu.nl, 04-09-21.

“Politie stopt met algoritme dat voorspelt wie gewelddadig zou kunnen worden

(…) Het algoritme keek naar iemands geslacht, leeftijd en strafblad. Aan de hand daarvan beoordeelde de politie wie gewelddadig zou kunnen worden.
Mensen die uit de beoordeling naar voren kwamen, kregen te horen dat ze een "veiligheidsrisicosubject" waren. Dat betekende in de praktijk dat ze vaker gecontroleerd zouden worden en verplicht waren daaraan mee te werken. (…)”

Politie stopt met algoritme dat voorspelt wie gewelddadig zou kunnen worden | Tech | NU.nl, 25 aug. 2023.

“Microsoft stopt met omstreden gezichtsherkenning vanwege zorgen over misbruik

Microsoft stopt met de gezichtsherkenningstechnologie waarmee het bedrijf onder meer emoties van mensen zegt te kunnen raden. Volgens Microsoft riep de kunstmatige intelligentie vragen over privacybescherming op en bestond het risico op discriminatie en andere vormen van misbruik. (…)”

“Microsoft stopt met omstreden gezichtsherkenning vanwege zorgen over misbruik”, nu.nl, 22 jun 2022.

“Microsoft stelt nieuwe AI-richtlijnen op en zet gezichtsherkenning in Azure stop

Microsoft gaat strakkere richtlijnen opstellen voor hoe het kunstmatige-intelligentiesystemen bouwt. In die Responsible AI Standard staan regels voor stem- en gezichtsherkenning. Met dat laatste stopt Microsoft gedeeltelijk, omdat er te veel risico's aan zitten.

Microsoft zegt dat het de meest recente versie van zijn Responsible AI Standard heeft vrijgegeven. (…)”

Microsoft stelt nieuwe AI-richtlijnen op en zet gezichtsherkenning in Azure stop - IT Pro - Nieuws – Tweakers, 22 juni 2022.

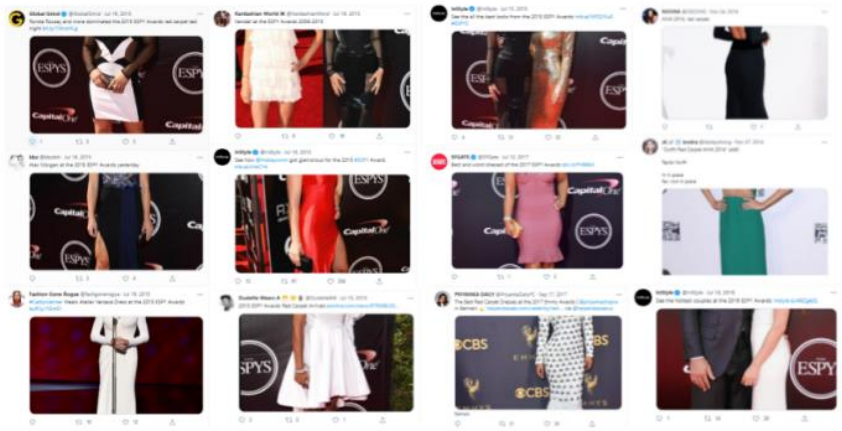


Figure 2: A collage of real-world user-uploaded images on

Als Twitter foto's van vrouwelijke beroemdheden verkleind weergeeft, dan bij vrouwen vaak ingezoomd op het lichaam.

Waarom?

De AI vond de logo's het belangrijkste aspect van de foto's

Devansh, Twitter, Racism, and the Male Gaze: Featuring AI from Google and Apple. Are they all prejudiced? [Twitter, Racism, and the Male Gaze | by Devansh | MLearning.ai | Feb, 2022 | Medium]

Creditcard limiet

Mannen kregen bij de Apple Card meer limiet op hun creditcard dan vrouwen, ook bij gelijke inkomsten, spaartegoeden, credit score, belastingaangifte, etc.

Aanvragers was niet naar hun gender gevraagd. De beslissing was dus niet gebaseerd op man/vrouw.

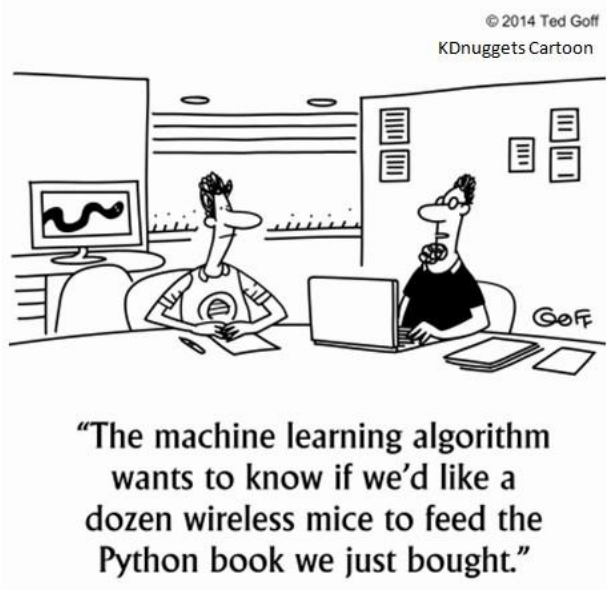
Hoe komt dit?

Kevin Peachey, Sexist and biased? How credit firms make decisions, 18 November 2019, Sexist and biased? How credit firms make decisions - BBC News

Correlatie en causaliteit

De ‘black box’ zoekt allerlei verbanden, maar maakt geen onderscheid tussen correlatie en causaliteit.

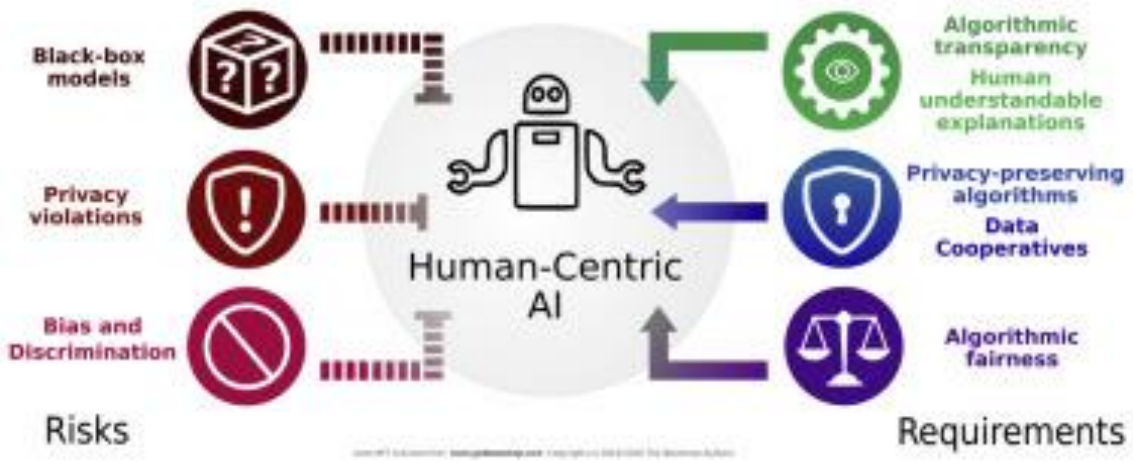
Daarom heel belangrijk dat de data input op geen enkele wijze vooroordelen bevat.

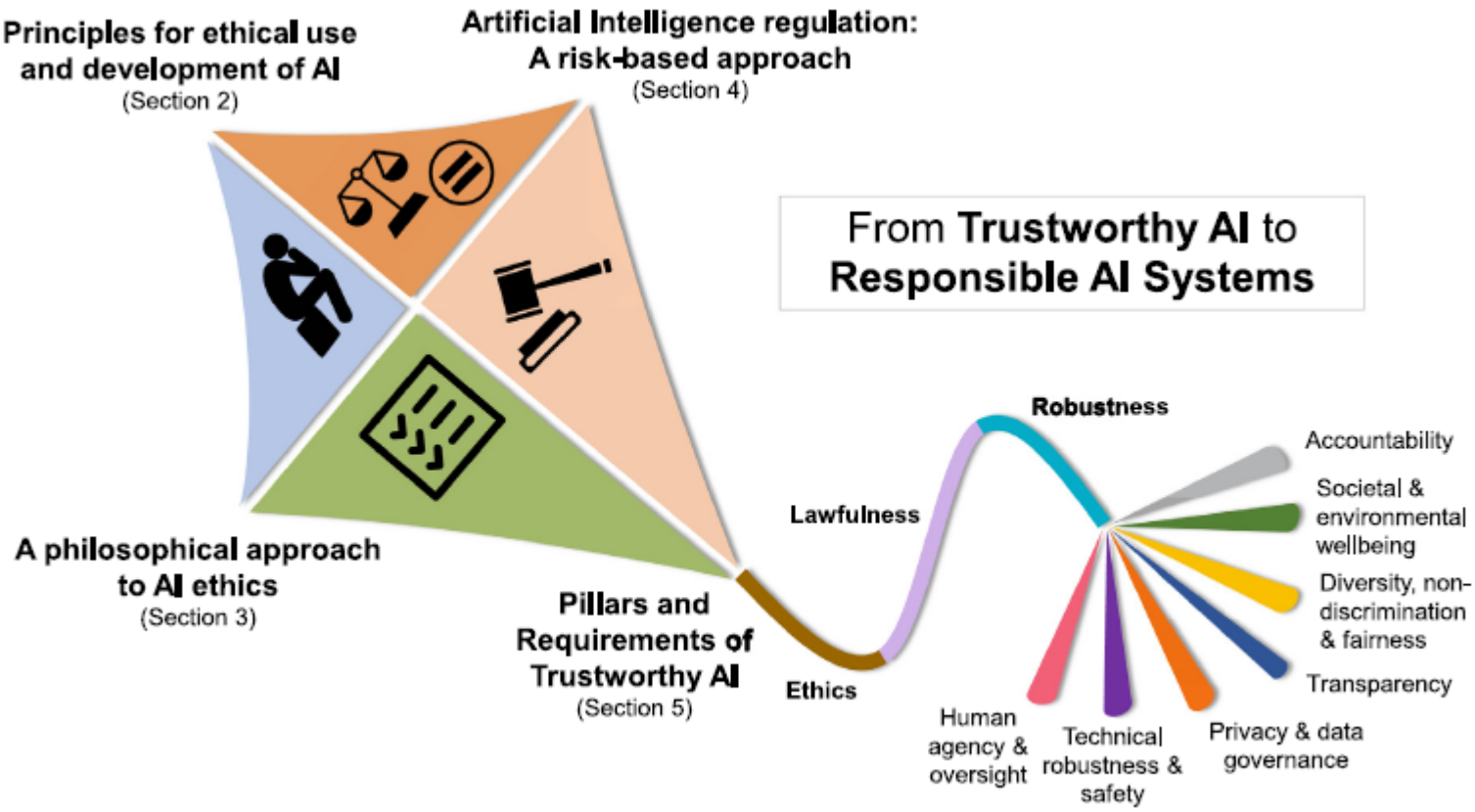


'(…)the underlying, fundamental problem at the heart of AI, which is that even a thoughtfully designed algorithm must make decisions based on inputs from a flawed, imperfect, unpredictable, idiosyncratic real world'

Jonathan Shaw, 'Artificial Intelligence and Ethics: Ethics and the dawn of decision-making machines', *Harvard Magazine*, jan/febr 2019.

Andere voorbeelden ethiek in Machine Learning?


Black-box models
Privacy violations
Bias and Discrimination
Risks
Human-Centric AI
Algorithmic transparency
Human understandable explanations
Privacy-preserving algorithms
Data Cooperatives
Algorithmic fairness
Requirements



Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation☆

Natalia Díaz-Rodríguez [a,*,1], Javier Del Ser [b,c,**,1], Mark Coeckelbergh [d], Marcos López de Prado [e,f,g], Enrique Herrera-Viedma [a], Francisco Herrera [a]

5 principes

- Beneficence: promoting well-being, preserving dignity, and sustaining the planet
- Non-maleficence: privacy, security and ‘capability caution’
- Autonomy: the power to decide (to decide)
- Justice: promoting prosperity, preserving solidarity, avoiding unfairness
- Explicability: enabling the other principles through intelligibility and accountability

Luciano Floridi en Josh Cowls, A Unified Framework of Five Principles for AI in Society, HDSR (Harvard Data Science Review), 2 juli 2019, A Unified Framework of Five Principles for AI in Society · Issue 1.1, Summer 2019 (mit.edu)

AI

- Tegenover de mens (bijv. schaakwedstrijd)
- In combinatie met de mens (bijv. politie bij foto matching)

European Commission
High Level Expert Group on AI

1. Accountability
2. Data Governance
3. Design for all (by all -include diversity)
4. Governance of AI Autonomy (Human oversight)
5. NonDiscrimination
6. Respect for Human Autonomy
7. Respect for Privacy
8. Robustness
9. Safety
10. Transparency

https://ec.europa.eu/digital-single-market/en/news/draft-ethicsguidelines-trustworthy-ai

PRINCIPLED ARTIFICIAL INTELLIGENCE

A Map of Ethical and Rights-Based Approaches to Principles for AI

Authors: Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, Madhulika Srikumar

Designers: Arushi Singh (arushisingh.net) and Melissa Axelrod (melissaaxelrod.com)

HOW TO READ:

Date, Location

Document Title

Actor

COVERAGE OF THEMES:

Higher — Lower

Not referenced

References International Human Rights

Explicitly Adopts Human Rights Framework

The size of each dot represents the percentage of principles in that theme contained in the document. Since the number of principles per theme varies, it's informative to compare dot sizes within a theme but not between themes.

The principles within each theme are:

Privacy:
Privacy
Control over Use of Data
Consent
Privacy by Design
Recommendation for Data Protection Laws
Ability to Restrict Processing
Right to Rectification
Right to Erasure

Accountability:
Accountability
Recommendation for New Regulations
Impact Assessment
Evaluation and Auditing Requirement
Verifiability and Replicability
Liability and Legal Responsibility
Ability to Appeal
Environmental Responsibility
Creation of a Monitoring Body
Remedy for Automated Decision

Safety and Security:
Security
Safety and Reliability
Predictability
Security by Design

Transparency and Explainability:
Explainability
Transparency
Open Source Data and Algorithms
Notification when Interacting with an AI
Notification when AI Makes a Decision about an Individual
Regular Reporting Requirement
Right to Information
Open Procurement (for Government)

Fairness and Non-discrimination:
Non-discrimination and the Prevention of Bias
Fairness
Inclusiveness in Design
Inclusiveness in Impact
Representative and High Quality Data
Equality

Human Control of Technology:
Human Control of Technology
Human Review of Automated Decision
Ability to Opt out of Automated Decision

Professional Responsibility:
Multistakeholder Collaboration
Responsible Design
Consideration of Long Term Effects
Accuracy
Scientific Integrity

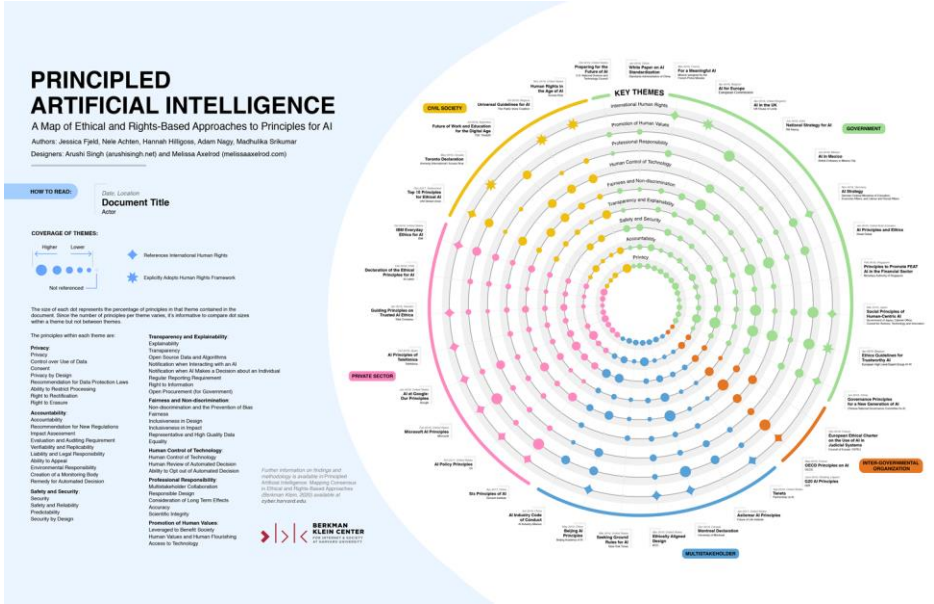
Promotion of Human Values:
Leveraged to Benefit Society
Human Values and Human Flourishing
Access to Technology

Further information on findings and methodology is available in Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches *(Berkman Klein, 2020) available at* ***cyber.harvard.edu***.

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

[PrincipledAI_FinalGraphic.jpg (5100×3300) (harvard.edu)]

Key themes



- Privacy
- Accountability
- Safety and security
- Transparency and explainability
- Fairness and non-discrimination
- Human control of technology
- Professional responsibility
- Promotion of human values

Rondom die thema’s (en andere waarden) kunnen morele vragen in elke fase ontstaan.

	Business understanding	Data understanding	Data preparation	Modelling	Evaluation	Deployment
Privacy						
Accountability						
Safety & Security						
Transparency & Explainability						
Fairness & non-discrimination						
Human Control of Technology						
Professional Responsibility						
Promotion of Human Values						
Overig						

hogeschool Windesheim

Herhaling basiskennis Bedrijfsethiek

hogeschool Windesheim

Wat is een moreel dilemma?

hogeschool Windesheim

Wat is het verschil tussen een zuiver en een onzuiver moreel dilemma?

hogeschool Windesheim

Hoe formuleer je een morele vraag?

hogeschool Windesheim

Wat is een waarde?

hogeschool Windesheim

Wat is een norm?

hogeschool Windesheim

Hoe verhouden normen en waarden zich tot elkaar?

hogeschool Windesheim

Wanneer ben je moreel verantwoordelijk?

hogeschool Windesheim

Wat is een moreel oordeel?

hogeschool Windesheim

Hoe kun je besluiten wat nu moreel juist is om te doen? (wat voor soort argumenten kun je gebruiken)

Het formuleren van een morele vraag

- Mag je…. / Is het moreel verantwoord om…

- Dilemma duidelijk maken

Bijvoorbeeld: Mag je van alle burgers ongericht data opslaan en analyseren, zonder aanleiding, om daarmee de veiligheid beter te kunnen waarborgen?

Oefening

- Bekijk de eerder geformuleerde voorbeelden van ethiek in Machine Learning
- Kies een voorbeeld en formuleer mogelijke morele vragen hierbij

Kies 1 morele vraag

hogeschool Windesheim

Hoe formuleer je een plichtethisch argument?

hogeschool Windesheim

Hoe formuleer je een gevolgenethisch argument?

hogeschool Windesheim

Hoe formuleer je een deugdethisch argument?

Wat is nu goed handelen / wat is juist ?

- Plichtsethiek (deontologie) – de aard van de handeling

…., want je hoort (niet) ….<norm>

- Gevolgenethiek (utilisme)- het resultaat / doel (maximaliseren nut voor alle betrokkenen)

…., want gevolgen <a, b, …> voor betrokkene <x> en gevolgen <c, d, …> voor betrokkene <y> wegen zwaarder dan gevolgen <e, f, …> voor betrokkene <z>, omdat ….<afweging>

- Deugdethiek – de intentie, bedoeling, op basis van positieve karaktereigenschap

…., want als persoon wil ik graag <deugd> zijn.

Exit ticket

- Formuleer een morele vraag op het gebied van Machine Learning waaruit duidelijk een moreel dilemma blijkt

- Noteer daaronder welke waarden er botsen

Ethiek in portfolio na workshop 1 (individueel)

➔ Drie voorbeelden van morele dilemma's in Machine Learning die niet in de workshop genoemd zijn. Vermeld hierbij de bron.

➔ Geef voor elk voorbeeld aan wat maakt dat dit een moreel dilemma is. Welke onderliggende waarden zie je botsen in dit dilemma?

➔ Geef voor elk voorbeeld aan waar voor jou persoonlijk de morele grens ligt. Licht je antwoord toe.

Neem de uitwerking mee naar de volgende workshop.