hogeschool Windesheim

Machine Learning Ethics

Terugblik exit tickets

hogeschool Windesheim
Dichter bij jou

hogeschool Windesheim

Hoe ging het maken van opdrachten B en C?

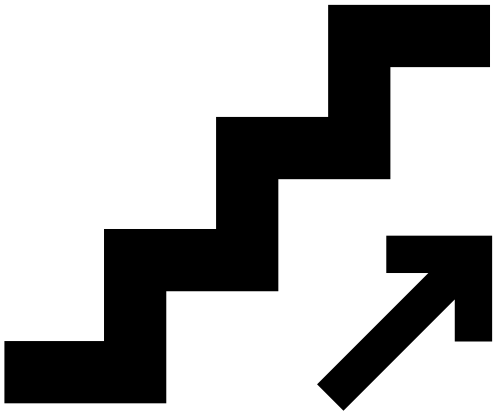
Workshop 3

Inhoud ethiek: morele intuïtie en persoonlijke integriteit – hoe kun je de morele grens bewaken?

Welke mate van foutgevoeligheid is acceptabel?

Waar kun je deze ontwikkeling (positief en negatief) nog meer voor gebruiken?

Vorige keer: belangen



Algemeen belang

Belang van derden

Welbegrepen eigenbelang

Geïsoleerd eigenbelang

Kwantitatieve bepaling van het meeste geluk

1. Intensiteit,
2. Duur,
3. Zekerheid/onzekerheid
4. Nabijheid/afstand,
5. Vruchtbaarheid (kans door ander genot van hetzelfde soort gevolgd te worden)
6. Reinheid (kans niet door een ervaring van tegenovergestelde aard te worden opgevolgd)
7. Extensie (aantal betrokkenen)

hogeschool Windesheim

Hoe komt het dat mensen dingen doen die zij zelf moreel onjuist vinden?

Sociaal-psychologische factoren (verantwoordelijkheid)

- Eigenbelang
- Gebrek aan bereidheid (te weinig moed)
- Microscopische visie (alleen verdiepen in eigen specialisatie)
- Gehoorzamen aan autoriteiten
- Toewijding (niet meer willen afwijken als er iets is besloten)
- Groepsdenken (mee doen met de meerderheid)
- Depersonalisatie (bijv. patiënten uitdrukken in bedden)
- Onzorgvuldig oordeelsvorming (werkelijkheid wordt vereenvoudigd tot model)

hogeschool Windesheim

Wat kun jij als professional doen om dat bij jezelf en anderen te voorkomen?

Morele intuïtie

- Automatische morele oordelen
- Dagelijks en vanzelf
- Bij gedrag anders dan verwacht

    - *Je kunt toch wel op tijd komen?!*
    - *Doe toch even normaal!*
    - *Hij laat ons voor al het werk opdraaien. Dat doe je toch niet?!*
    - *Ik vind het niet netjes dat ze zich niet even afgemeld heeft.*

- Objectief – feitenkennis (geen discussie)
- Subjectief – persoonlijk (voor ieder anders)
- Intersubjectief – gedeeld (normatief) kader

Intersubjectief – gedeeld (normatief) kader

- Tussen objectief en subjectief in

- Morele oordelen zijn dus NIET volledig subjectief (mensen delen vaak een oordeel)

Janine is 5 minuten te laat

- Feitelijk: Het is 9.05 uur en Janine is er nog niet, terwijl we dat wel hebben afgesproken.
- Emotioneel: Ik ben boos dat Janine er nog niet is (of: Fijn, dan kan ik nog even koffie halen)
- Moreel: Janine hoort op tijd te komen.

Pas op voor verwarring

- Emoties en morele oordelen
  - '*Fijn, kan ik koffie halen'* betekent niet dat diegene vindt dat je niet op tijd hoeft te komen.

- Morele oordelen en feiten
  - *'Het is nu 9.05 uur!'* (feitelijke uitspraak met doorklinkend moreel oordeel dat ze er had moeten zijn)

Verscholen moreel oordeel

- Het is 9.05 uur!

Duidelijker moreel oordeel zou zijn:

Je hoort op tijd hier te zijn als we een afspraak hebben.

Waarom spreken we morele oordelen zo impliciet uit?

- Twijfel aan juistheid van het standpunt
  (feitelijk: ‘nee hoor, je horloge loopt voor’ of moreel: ‘hoor je altijd stipt op tijd te zijn?’)
- Twijfel over acceptatie van het standpunt door anderen
  (‘5 minuten te laat vind ik nog wel acceptabel’)
- Twijfel over reactie van anderen
  (door het oordeel te verschuilen, hoeft de ander er niet expliciet op in te gaan)

  Gevolg: mensen spelen op safe, zijn bang om precies te zeggen wat ze vinden

Gevolgen van het bang zijn om te zeggen wat je vindt

- Ontkennen van verantwoordelijkheid
  - Gedachten zijn vrij en keuzes dus ook. Als we handelen, hebben we daarvoor gekozen.
  - Als je iets of iemand als excuus gebruikt voor jouw gedrag, dan ontken je je eigen vrijheid en verantwoordelijkheid.

Gevolgen van het bang zijn om te zeggen wat je vindt

- Aanpassen aan wat anderen vinden
  - Gewoonte om te doen wat de groep doet
  - Niet jezelf zijn
  - Makkelijk: niet zelf nadenken over wat moreel juist is
  - Drogreden: iedereen doet het zo
  - Risico hellend vlak en vervagende normen

  Belangrijk: bewust zijn van groepsnormen en kritisch over blijven nadenken. Blijf jezelf vragen: waarom doe ik hieraan mee?

hogeschool Windesheim

Eigen voorbeelden

Ontkennen van verantwoordelijkheid

Aanpassen aan anderen

Morele intuïtie ontwikkelen

- Blijf alert op je taalgebruik en argumenten (wijs je naar anderen als excuus voor jouw handelen?)
- Wees bewust van groepsdenken en –gedrag (doe je dingen omdat anderen het ook doen?)
- Sta aandachtig stil bij wat er in je omgaat (welke ervaringen liggen ten grondslag aan jouw morele oordeel?)

Bewust worden

- Van het morele oordeel dat je intuïtief vormt
- Van de mate waarin je bang bent om zelf verantwoordelijk te zijn voor je keuzes
- Wanneer je geneigd bent mee te lopen met anderen

Als je daar bewust van bent, kun je erover nadenken, leren van je ervaringen en andere keuzes maken.

Function creep

Wanneer het doel van een systeem steeds verder opgerekt wordt

Bijv. verkeerscamera’s boven snelwegen:

- Oorspronkelijk voor verkeersstromen en overtredingen
- Nu ook voor traceren onverzekerde auto’s en volgen verdachte auto’s

Voorbeelden function creep

hogeschool Windesheim
Dichter bij jou

Hoe voorkom je function creep?

Hoe zie je het aankomen?

Inschatten leeftijd

- Waar kun je dat allemaal voor gebruiken?
- Waar zou iemand het nog meer voor kunnen (gaan) gebruiken?

Leeftijd inschatten

Wat als 1 op de 10.000 afbeeldingen verkeerd wordt herkend?

Wat kunnen daar de gevolgen van zijn? *(bijv. bij leeftijdsinschatting van vluchtelingkinderen om recht op asiel te bepalen)?*

Wanneer wegen baten nog op tegen kosten (zoals mogelijke nare gevolgen van gemaakte fouten)?

Waar zou je nog aan meewerken?

En wat als je niet wilt meewerken?

Welke mate van foutgevoeligheid is nog acceptabel?

Uitwerken volgens stappenplan

- Morele vraag (met leeftijd inschatten en mate van foutgevoeligheid)

- Normen, waarden en deugden

- Argumenten voor en tegen

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.

- Oordeel en reflectie

hogeschool Windesheim

Hoe programmeer je ethiek in een systeem?

Oefening Golem Genie

“**The Golem Genie**: Imagine a superpowerful golem genie materialises in front of you one day. It tells you that in 50 years time it will return to this place, and ask you to supply it with a set of moral principles. It will then follow those principles consistently and rigidly throughout the universe. If the principles are faulty, and have undesirable consequences when followed consistently by a superpowerful being, then disaster could ensue. So it is up to you to ensure that the principles are not faulty. Adding to your anxiety is the fact that, if you don’t supply it with a set of moral principles, it will follow whichever moral principles somebody else happens to articulate to it.”

The Golem Genie and Unfriendly AI , 31 januari 2013. Philosophical Disquisitions: The Golem Genie and Unfriendly AI (Part One)

- Schrijf de principes op
- Presenteer aan de anderen

Exit ticket

Waar ligt voor jou een morele grens:

welke (mogelijke) ontwikkeling op het gebied van Machine Learning wil je niet aan meewerken?

Wat maakt dat dat voor jou een grens is? (welke ervaringen liggen daaraan ten grondslag)

Na workshop 3 in portfolio (individueel):

Opdracht D (individueel, na workshop 3 in portfolio):
- Benoem morele vragen bij de tweede themaopdracht (pakketjes ‘herkennen’ inpakrobot) en geef aan in welke fase van de CRISP-DM cyclus de vraag speelt.
- Kies een relevante morele vraag en maak daarvoor een individuele uitwerking van de themaopdracht o.b.v. het stappenplan uit de bijlage. De stap ‘handelingsopties’ mag je hierbij nog even overslaan.
- Licht je eigen morele grens bij dit thema toe.
Neem de uitwerking mee naar de volgende workshop.

Opdracht E (individueel, na workshop 3 in portfolio):

- Evalueer zelf je uitwerking van opdracht D aan de hand van de checklist bij het stappenplan ethiek ML.
- Markeer de wijzigingen die hieruit voortvloeien zodat de verschillen goed zichtbaar worden (bijvoorbeeld met kleur).