**Ethische kwesties in Machine Learning**

*With great power*

*comes*

*great responsibility*

Machine Learning

ICT.ML.V21

Caroline Oosterkamp

Augustus 2024

**Inleiding**

Wie is er verantwoordelijk bij een ongeluk met een zelfrijdende auto? Hoe voorkomen we een onrechtvaardige behandeling door (bv. rechterlijke, ambtelijke) uitspraken of beslissingen (gezondheidszorg, onderwijs) gedaan door zelflerende algoritmes? De toepassing van machine learning heeft een enorme impact op hoe we leven of zouden willen leven. Het is daarom van belang bij de ontwikkeling van Machine Learning oplossingen ook na te denken over de morele of de wenselijke kant van de toepassing van ML.

Naarmate er meer gevorderde technieken gebruikt worden voor Machine Learning, wordt de code bovendien zo complex dat het vaak alleen nog als black box gebruikt wordt. Denk bijvoorbeeld aan frameworks zoals Tensorflow of Pytorch. Daardoor neemt het risico toe dat de machine onwenselijke zaken leert, wat tot onwenselijk output leidt, zonder dat dit direct herkend wordt. Bovendien is er steeds meer mogelijk met behulp van Machine Learning en *“with great power comes great responsibility.”* Het is ook daarom van belang bij de ontwikkeling van Machine Learning oplossingen na te denken over de morele kant ervan.

Tijdens de eerste periode van het semester Machine Learning is er bij elk thema een bijbehorende workshop over ethiek. Tijdens deze workshops zullen we ingaan op de morele kant van de onderwerpen die die week centraal staan bij dat thema. Gaandeweg leer je steeds beter morele kwesties op het gebied van Machine Learning te herkennen en te benoemen en denk je onder andere na over mogelijke morele argumenten om dingen wel of niet te doen en hoe je dergelijke argumenten boven tafel kunt krijgen.

In de tweede periode richten de contactmomenten zich op de morele kant van jullie eigen project en jullie persoonlijke visie daarop.

**Leeruitkomst**

In dit onderdeel staat de volgende leeruitkomst centraal:

*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.*

Dit is één van de leeruitkomsten van het semester. Ethiek vormt dus geen aparte module, maar is een integraal onderdeel van de toetsing van het semester.

**Periode 1: vier workshops rondom Machine Learning thema’s**

**Workshop 1**

Inhoud ethiek: herhaling basiskennis uit de module *Bedrijfsethiek voor ICT* en de ethiekworkshops bij Data Science, verdieping op tweezijdig argumenteren vanuit deontologie, utilisme en deugdethiek en reflectie op oordeel.

Inleiding morele kwesties in Machine Learning.

Denkvraag: Welke morele vragen kun je hierbij stellen?

**Opdracht A (individueel, na workshop 1 in portfolio):**

* 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.

**Workshop 2**

Inhoud ethiek: morele onderbouwing - hoe kun je (gezamenlijk) moreel redeneren inzetten om het goede na te streven? Daarnaast een uitleg van de belangenhiërarchie.

Nastreven waarden m.b.v. Machine Learning.

Denkvraag: Welke waarden kunnen hier botsen?

**Opdracht B (individueel, na workshop 2 in portfolio):**

Zoek een partner in de klas. Werk individueel de volgende stappen uit in een verslag:

* Benoem morele vragen bij de eerste themaopdracht (energie-efficiëntie huizen Assendorp) 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.
* Herzie je eigen uitwerking o.b.v. de ontvangen feedback (uit opdracht C).
* Neem ook de ontvangen feedback op en benoem van welke student dit afkomstig is.

Neem de uitwerking mee naar de volgende workshop.

**Opdracht C (individueel, na workshop 2 in portfolio):**

* Geef je medestudent feedback a.d.h.v. alle vragen in de *checklist bij het stappenplan ethiek ML*. Neem de gegeven feedback op in je eigen verslag.

Neem de uitwerking mee naar de volgende workshop.

**Workshop 3**

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

Waar kun je deze ontwikkeling (positief en negatief) nog meer voor gebruiken? Uitleg risico ‘function creep’.

Denkvraag: Welke mate van foutgevoeligheid is acceptabel?

**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 *checklist bij het stappenplan ethiek ML.*
* Markeer de wijzigingen die hieruit voortvloeien zodat de verschillen goed zichtbaar worden (bijvoorbeeld met kleur).

**Workshop 4**

Inhoud ethiek: de toepassing van handelingsopties bij en na de keuze om iets moreel goed te keuren

Introductie begeleidingsethiek

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

**Opdracht F (individueel, na workshop 4 in portfolio):**

* Benoem morele vragen bij de derde themaopdracht (chatbot).
* Kies een relevante morele vraag en maak daarvoor een individuele uitwerking van de themaopdracht o.b.v. het stappenplan uit de bijlage.
* Werk hierbij ook de stap ‘handelingsopties’ uit.
* Licht je eigen morele grens bij dit thema toe.
* Gebruik bij deze opdracht de *checklist bij het stappenplan ethiek ML.*

**Inleveren tussentijds portfolio:**

Na de vier workshops wordt het individueel portfolio (waar ethiek een onderdeel van is) op de ELO ingeleverd voor feedback.

* Kies opdracht D/E of F als showcase om de leeruitkomst aan te tonen.
* Voeg in één bijlage alle uitwerkingen van de ethiekopdrachten A t/m F toe.

**Periode 2: eigen project**

In de tweede periode ga je bezig met het herkennen en benoemen van diverse morele vragen bij een eigen praktijkopdracht binnen een Machine Learning context en het onderbouwen van een moreel oordeel met argumenten vanuit verschillende ethische invalshoeken. Daarbij denk je ook na over je eigen morele grens (wat ben je wel of niet bereid om te doen en waarom).

**Workshop 5: feedback showcase in tussentijds portfolio**

In deze workshop ontvangen jullie feedback op de ingeleverde showcases uit periode 1. Jullie dienen deze feedback te verwerken voor het definitief in te leveren portfolio en er ook van te leren voor de uitwerking van de morele vraag bij het project.

**Opdracht G (individueel, voor workshop 6):**

* Formuleer meerdere morele vragen bij het project. Zorg dat dit morele vragen zijn die over jullie eigen morele dilemma’s gaan.

**Workshop 6: morele vragen**

De diverse morele vragen bij jullie projecten worden uitgewisseld en besproken in een workshop, zodat jullie van elkaar kunnen leren en elkaar feedback kunnen geven.

Kies als groep één morele vraag die relevant is voor jullie project en wat jullie eigen morele dilemma is. Neem deze vraag als uitgangspunt voor opdrachten H en I.

**Opdracht H (individueel, voorafgaand aan workshop 7):**

* Maak een individuele uitwerking (op basis van het stappenplan) van de morele vraag die jullie als groep gekozen hebben.

**Opdracht I (groepsuitwerking, voorafgaand aan workshop 7):**

* Ga met elkaar in gesprek over verschillen van morele grenzen.
* Kom als groep tot een gezamenlijke uitwerking van een morele vraag bij jullie project aan de hand van het stappenplan (zie bijlage).
* Mail de uitwerking van opdracht I drie werkdagen voor de feedbacksessie naar de docent.

**Workshop 7: Feedbacksessie per groep**

De groepsuitwerking wordt in een feedbacksessie met de docent besproken.

**Opdracht J (groepsuitwerking):**

* Verwerk daarna als groep de ontvangen feedback in de groepsuitwerking.
* Markeer de wijzigingen die hieruit voortvloeien zodat de verschillen goed zichtbaar worden (bijvoorbeeld met kleur).

**Eindverslag: individueel portfolio**

Ethiek is onderdeel van het eindportfolio voor het semester Machine Learning. Op de ELO staan het sjabloon en de beoordeling nader toegelicht.

Voor de leeruitkomst ethiek gaat het om het herkennen en benoemen van diverse morele vragen bij een eigen praktijkopdracht binnen een Machine Learning context en het onderbouwen van een moreel oordeel met argumenten vanuit verschillende ethische invalshoeken.

Het moreel oordeel dat hieruit volgt, staat niet op zichzelf. Het betreft jullie eigen project en het oordeel en de afweging daarbij kan dus ook gevolgen hebben voor hoe je het project wilt doen of zelfs de vraag oproepen of je het project wel wilt doen. Daarom is het belangrijk om het oordeel vervolgens te koppelen aan je morele grens: wat betekent dit oordeel nu voor de realiteit? Wil je het project nog wel op deze manier doen? Benoem dus ook jouw persoonlijke morele grens, licht deze toe en geef aan wat dit betekent voor jullie project.

In het portfolio laat je met twee showcases zien hoe en in welke mate de leeruitkomst ontwikkeld is. Dit wordt individueel onderbouwd met twee showcases. De eerste showcase is de uitwerking van **opdracht D/E of F**. Hiervoor kan de eerder ingeleverde showcase opnieuw ingebracht worden, waar nodig herzien op basis van de ontvangen feedback. De tweede showcase is de uitwerking van **opdracht H.** In de bijlage neem je **alle uitwerkingen (A t/m J)** op.

Voorwaarden voor beoordeling:  
1. Zijn de volledige uitwerkingen van opdrachten A t/m J in één bijlage opgenomen? (dus niet in het portfolio zelf)  
2. Is de 1e showcase de uitwerking van opdracht D/E of F?  
3. Is de 2e showcase de uitwerking van opdracht H?  
4. Indien er gebruik gemaakt is van ChatGPT (of soortgelijke hulpmiddelen): is dit helder vermeld én kun je nog steeds zelf de leeruitkomst voldoende aantonen?

Na het inleveren van het portfolio volgt er (naast het reguliere assessment voor de andere leeruitkomsten) een assessment specifiek gericht op de leeruitkomst ethiek. Dit groepsassessment wordt afgenomen door twee ethiekdocenten. Na dit ethiekassessment wordt voor elke student individueel de beoordeling van de ethiekleeruitkomst vastgesteld. Deze leeruitkomst vormt een integraal onderdeel van de toetsing van het semester. Een onvoldoende beoordeling voor ethiek betekent dus ook een onvoldoende voor het gehele semester.

**Bijlage: stappenplan**

* Geef een korte toelichting op de opdracht, waarin alle relevante feiten benoemd worden.
* Benoem een morele vraag passend bij jullie opdracht.
* Benoem de morele uitgangspunten (normen, waarden en deugden) die je relevant vindt. Plaats deze tegenover elkaar en werk voor beide kanten morele uitgangspunt(en) uit. Leg uit hoe en waarom deze uitgangspunten een rol spelen.
* Ga in op de kwestie van verantwoordelijkheid. Wie is in deze kwestie waarvoor verantwoordelijk en waarom?
* Benoem diverse argumenten vóór en tegen het moreel dilemma in kwestie vanuit
  + De plichtethische invalshoek
  + De gevolgenethische invalshoek
  + De deugdethische invalshoek
* Benoem je moreel oordeel. Dit oordeel is een afgewogen antwoord op je morele vraag.
* Reflecteer kritisch op het oordeel.
* Handelingsopties (indien van toepassing): voeg concrete handelingsopties toe m.b.t. de technologie, omgeving en gebruiker.
* Terugblik stappenplan: eigen morele grens

*Alleen bij groepsuitwerking project:* ga bij deze stap in op verschillen tussen jullie morele grenzen.

* Consequenties voor werken aan opdracht / project (kan en wil je ermee doorgaan, moeten er dingen anders aangepakt, moet er iets aangekaart of geadviseerd worden, etc.)