Kenniscentrum
Data & Maatschappij

D2.2 RAAMWERK VOOR ETHISCHE VALIDATIE VAN AI

Kenniscentrum
Data & Maatschappij

DECEMBER 2019

© 2019, Kenniscentrum Data & Maatschappij

Dit werk valt onder een Creative Commons Naamsvermelding 4.0 Internationaal-licentie.

Je bent vrij om het werk te delen en te bewerken onder de volgende voorwaarden: naamsvermelding & geen aanvullende restricties. Voor elementen van het materiaal die zich in het publieke domein bevinden, en voor vormen van gebruik die worden toegestaan via een uitzondering of beperking in de Auteurswet, hoef je je niet aan de voorwaarden van de licentie te houden. Er worden geen garanties afgegeven. Het is mogelijk dat de licentie je niet alle gebruiksvrijheden geeft die nodig zijn voor het beoogde gebruik. Bijvoorbeeld, andere rechten zoals publiciteits-, privacy- en morele rechten kunnen het gebruik van een werk beperken. Je kan meer informatie vinden op de website van Creative Commons https://creativecommons.org/

Deze uitgave citeren als:

Ballon, P., Duysburgh, P., Fanni, R., Franck, G., Heyman, R., & Laenens, W. (2019). D2.2. Raamwerk voor ethische validatie van AI. Kenniscentrum Data & Maatschappij, Brussel, België. [auteurs in alfabetische volgorde]

www.data-en-maatschappij.ai

Inhoudstafel

Introductie

Deze deliverable van het Kenniscentrum Data & Maatschappij biedt een gecategoriseerd overzicht van tools om AI[1] ethisch te valideren. In dit document verkennen en vergelijken we tools die de mogelijkheid bieden om AI-toepassingen te beoordelen volgens een aantal criteria. Dit moet sectororganisaties, overheden en bedrijven in staat stellen de AI-toepassingen waar zij gebruik van willen maken of die zij op de markt willen brengen aan een ethische reflectie te onderwerpen, en de AI-toepassing meer 'trustworthy' te maken en dus ons vertrouwen in AI-toepassingen (terecht) vergroten.

Door het groeiende aantal AI-toepassingen en de verspreiding ervan in diverse domeinen, gaande van het bankwezen, human resources tot de openbare sector etc., wordt de impact van deze technologie steeds groter. Door dit laatste, wordt de nood voor een 'ethisch gebruik van AI' steeds groter. Maar het is onduidelijk hoe ethisch omspringen met AI-technologie zich vertaalt naar de praktijk. De vele en verschillende definities van AI en de diverse manieren waarop 'ethiek' wordt geïnterpreteerd, zorgen ervoor dat ethiek op heel verschillende manieren wordt geoperationaliseerd. Vraag is ook wie hier het voortouw in moet nemen, op welke manier deze personen moeten tewerk gaan en met welk doel.

Deze onduidelijkheid weerspiegelt zich in de veelheid ontwikkelde tools, instrumenten en kaders die beschikbaar zijn voor ethiek en AI. In dit rapport bieden we een overzicht van 18 ethische validatietools voor AI- en datagedreven toepassingen. Dit overzicht is niet exhaustief maar biedt een staalkaart voor sectororganisaties, overheden en bedrijven van de beschikbare raamwerken voor ethische validatie. Deze staalkaart zal niet alle onduidelijkheden wegnemen, maar wel aangeven welke instrumenten bestaan, hoe ze verschillen en op welke manier ze verschillend kunnen worden ingezet. Op basis van deze benchmarkoefening wil het Kenniscentrum Data & Maatschappij verder bouwen en in de toekomst de eigen visie rond ethiek en AI aanscherpen.

Dit document is als volgt opgebouwd. Eerst beschrijven we onze aanpak: hoe vonden we de instrumenten uit dit overzicht, op basis van welke criteria werden ze weerhouden, en hoe analyseerden we ze? Hierna volgen de resultaten: we beschrijven de gevonden tools en de variatie onder tools op basis van een aantal criteria. In de discussie reflecteren we op deze oefening en identificeren we mogelijke verbeterpunten bij bestaande tools. Vervolgens beschrijven we een aantal pistes die het Kenniscentrum zal verkennen om de bestaande raamwerken verder te verbeteren.

1 Artificiële intelligentie werd in deze deliverable niet verder gedefinieerd maar als zoekterm gebruikt om de veelheid aan instrumenten te categoriseren omtrent AI.

Methode

De lijst van instrumenten en de analyse ervan is op een systematische manier tot stand gekomen. Een longlist van instrumenten werd eerst samengesteld op basis van een combinatie van Google zoekopdrachten en de sneeuwbalmethode. De Google zoekopdrachten bestonden uit een combinatie van zoektermen als "ethical tools", "ethical canvases", "ethics" en "AI", "artificial intelligence", "data-driven applications", "digital technology", etc. Op basis van de referenties van de gevonden instrumenten, en hun vermelding in andere documenten, konden ook andere tools gevonden worden. Het resultaat van deze oefening was een oplijsting van 34 instrumenten.

Deze lijst werd gereduceerd naar een lijst van 18 tools of instrumenten. De selectie gebeurde op basis van de volgende prioriteiten:

- de praktische bruikbaarheid van het gevonden instrument
- de mate waarin ethiek aan bod komt[2]
- de beschikbaarheid van de tool ter evaluatie

Deze lijst van instrumenten werd vervolgens geanalyseerd en vergeleken op basis van een aantal dimensies:

- de maker van de tool
- het land van oorsprong
- het doelpubliek
- de voorgestelde methode
- het verwachte resultaat van de tool
- de benodigde middelen voor gebruik
- de fase waarin het instrument kan worden gebruikt
- de geëvalueerde AI-componenten
- de manier waarop ethiek werd ingevuld

Het resultaat van deze analyse staat beschreven in het volgende hoofdstuk.

2 Rapporten, cursussen of programma's van accelerators waarin AI en ethiek slechts beperkt aan bod komen, zijn niet mee opgenomen.

Resultaat

Het belangrijkste resultaat van de vergelijking tussen de tools is te vinden in een spreadsheet (zie bijlage), waarin de selectie van tools is opgenomen en de analyse in detail te bekijken is.

Hieronder staan de dimensies toegelicht waarmee de tools werden vergeleken. Initieel hadden de onderzoekers een strikte categorisatie op elk van de dimensies voor ogen om de vergelijkbaarheid te maximaliseren. Maar door de grote diversiteit van instrumenten was een strikte categorisering niet altijd mogelijk. Voor tools die niet in een categorie passen werd een meer beschrijvende aanpak gekozen.

De tools werden vergeleken op basis van volgende dimensies:

- **Ontwikkeld door:** In deze categorie geven we een overzicht van de entiteiten die vandaag ethische AI beoordelingen en/of begeleidingen ontwikkelen. We merkten hier 4 type actoren op: overheden, academische instanties, industrie en non-profit organisaties. Specifiek zijn het de overheden van onder meer de UK, Canada en Dubai die praktische richtlijnen en beoordelingskaders aanbieden voor AI-ontwikkelaars binnen de organisatie volgens de specifieke 'ethische AI' vereisten van het land. In ons overzicht van tools is er een relatieve balans tussen de verschillende types van actoren.
- **Doelpubliek:** Aangezien verschillende mensen in verschillende contexten met AI in contact komen, is er een aanzienlijke variatie in de beoogde doelpublieken. Ofwel werden doelpublieken als leden van de organisatie bestempeld ofwel helemaal niet. In geval van het eerste, ligt de klemtoon op leden van de organisatie zelf; managers, AI-ontwikkelaars, databeheerders, gegevensverwerkers, etc. In het tweede geval is het doelpubliek niet duidelijk gedefinieerd in de tool, wat het minder evident maakt om te bepalen wie mogelijks voordeel kan halen uit het gebruik ervan.
- **Methode (beschrijvend en gecategoriseerd):** In deze dimensie bekeken we hoe de tool praktisch beoogt om tot ethische AI te komen. Bij 'beschrijvend' wordt de specifieke aanpak toegelicht, onder 'gecategoriseerd' maken we een indeling van de verschillende aanpakken. We identificeerden vijf categorieën om de tools in te delen, die hieronder in meer detail worden toegelicht:
    a. theoretische kaders en principes: deze tools stellen doorgaans een aantal principes of begrippen voor waar AI aan moet voldoen om ethisch te zijn (e.g. mag geen bias creëren, moet transparant zijn, etc). Vaak is er ook aandacht voor de context van de toepassing en het effectieve gebruik, toch zijn deze tools in de regel doorgaans abstracter van aard en zijn ze minder concreet dan de volgende categorieën.
    b. leidraden en stappenplannen: deze tools bieden een proces of stappenplan aan dat je als gebruiker kan volgen in bv. de ontwerpfase van de technologie, of om bij het gebruik van databestanden aandacht te hebben voor ethische aspecten.
    c. vragenlijsten: aan de hand van vragenlijsten kan je tot een inschatting komen van eth-

ische aandachtspunten met betrekking tot een specifieke AI-toepassing of datagebruik. Een voorbeeld hiervan is de toepassing van de Canadese overheid, die je toestaat de 'algoritmische impact' van een AI-toepassing te berekenen. In veel gevallen zijn deze toepassingen 'evaluatief', en kunnen ze enkel na de ontwikkeling en implementatie worden gebruikt. Doorgaans zijn ze ook gebaseerd op een eigen inschatting (zelfevaluatie).

   d. digitale algoritme en data audit toolkits: dit zijn interfaces die organisaties toestaan hun algoritmes of datasets statistisch door te lichten en te laten nakijken op mogelijke interne biases, betrouwbaarheid, etc.
   e. opleidingen: deze opleidingen en cursussen trainen deelnemers in ethische reflectie met betrekking tot algoritmes en datagedreven toepassingen.

- **Beschikbaarheid / prijs:** Heel wat van de beschreven tools zijn online gratis beschikbaar. Bepaalde cursussen en specifiek consultancywerk is betalend. Er zijn ook tools die enkel contextgebonden gebruik toelaten en enkel door leden van een bepaalde organisatie kunnen worden gebruikt (eg. de Canadese tool voor overheidsinstanties die met AI- en datagedreven toepassingen aan de slag gaan, is enkel binnen die organisatie bruikbaar).
- **Benodigde effort:** De personeelsinzet die vereist is om deze tools te gebruiken, is vaak moeilijk in te schatten. Het vraagt om projectmanagers of andere profielen die bedreven zijn in het gebruik van deze tools en het gebruik ervan kunnen inpassen in de organisationele workflows. Soms zijn ook specifieke profielen nodig zoals business analisten, dataspecialisten en technische experten. Er zijn weinig indicaties van de werklast die het gebruik van deze tools creëert en het aantal uren / dagen dat hiervoor in rekening moet worden gebracht. Het is ook nodig om hierbij een onderscheid te maken tot de eigenlijke analyse op basis van de tool, en de impact van gevolggeving aan de conclusies die uit de analyse kunnen resulteren. Wel is duidelijk dat er grote verschillen zijn. Terwijl de inspanning van het gebruik van sommige tools zich beperkt tot het invullen van een vragenlijst, gaat het in andere gevallen om het uitwerken en documenteren van een uitgebreid stappenplan, een veel arbeidsintensievere oefening.
- **Proces:** De ethische aspecten van AI kunnen op verschillende momenten in het proces van een totstandkoming van een AI-systeem worden onderzocht. We identificeren vijf stadia of momenten, die allen te maken hebben met het AI-systeem op zich: ethische aspecten van AI kunnen worden onderzocht a) voor de ontwikkeling van het AI-systeem (probleemanalyse & ideatie), b) in de ontwerpfase (ontwerp), c) tijdens de ontwikkeling van het systeem (ontwikkeling), of d) tijdens de implementatie van het systeem (implementatie). Tenslotte kan ook e) nadat het project is opgeleverd, een ethische reflectie over de toepassing worden gemaakt, die mogelijks resulteert in een aanpassingen (evaluatie en iteratie).
- **Systeemniveau:** Het systeemniveau voor de meeste ethische tools is wel een belangrijk aspect om te bestuderen. Tools kijken naar datagebruik en de analytische methoden die gebruikt worden in de verwerking, de AI-technologie (eg. machine learning), naar de applicatie in z'n geheel, etc. Daarnaast kijken sommige tools ook buiten het systeem om naar de manier waarop

de tool wordt ingezet, de gebruikersnoden en de context van het AI-systeem.

- **Ethische principes:** Naar welke ethische aspecten kijkt deze tool? In welke mate zijn deze ethische principes ook vertaald in operationele doelstellingen en praktijkregels? Dit kan gaan over vooroordelen (bias), privacy, robuustheid, betrouwbaarheid, etc. Heel vaak merkten we dat hier geen verdere specificering is gebeurd van de specifieke ethische invalshoek die is gekozen, of ook, dat de termen die gehanteerd worden, niet verder zijn toegelicht, terwijl het vaak gaat om abstracte begrippen die op verschillende wijzen kunnen worden geïnterpreteerd.
- **Verwachte uitkomsten:** Afhankelijk van de hierboven beschreven methode kan de uitkomst van het gebruik van een tool sterk verschillen. In veel gevallen worden verschillende ethische aspecten van technologie belicht en wordt in hoofdzaak het bewustzijn vergroot van ethische aspecten van de technologie of het datagebruik. Doorgaans op basis van zelfrapportering stijgt het inzicht in de mate waarin aan een bepaalde standaard of norm wordt voldaan. In bepaalde gevallen kan de tool dienen om weloverwogen en gemotiveerde beslissingen te nemen. Externe evaluatie of 'bewijs' is zeldzaam.

Discussie

De systematische vergelijking van ethische tools voor AI maakt duidelijk dat een relatief grote variatie aan tools al beschikbaar is. De tools hebben gemeenschappelijk dat ze allen beogen voorbij het abstracte denken over ethiek en AI te gaan, en een werkwijze aanbieden om het denken hierover te kanaliseren en te concretiseren.

Terwijl we met dit overzicht geen volledigheid nastreven, is uit de selectie van tools wel al af te leiden dat er grote variatie is op vlak van instanties die de tools maken (academisch, industrie, non-profit, overheid), beoogde doelgroepen (project- en productmanagers, databeheerders, user experience specialisten, beleidsmakers, burgers, etc), gebruikte methodes (vragenlijsten, stappenplannen, theoretische frameworks), het systeemonderdeel waarop wordt gefocust, en de te verwachten uitkomsten. Deze diversiteit toont ook de veelheid aan manieren waarop ethiek in relatie met technologie in het algemeen en artificiële intelligentie in het bijzonder kan worden benaderd.

Zonder afbreuk te doen aan de verdienstelijkheid van deze tools, kwamen uit deze vergelijking ook een aantal tekortkomingen of mogelijke verbeterpunten voor deze tools in het algemeen, of sommige in het bijzonder naar boven. Een aantal punten die we willen aanhalen:

- Een aantal tools blijven vaag over het type AI-technologie dat ze precies beogen. De term 'artificiële intelligentie' is een brede waaier van technologieën[3] en de tools in kwestie verduidelijkten zelden in welk AI-domein of paradigma zij het ethisch denken faciliteren. Zeer weinig tools hebben effectief een definitie van AI opgenomen in hun richtlijnen voor gebruik. In de meeste gevallen dient de tool niet om ethiek te onderzoeken in het kader van AI, maar om in te zetten bij alle algoritme-gebaseerde tools of digitale technologie in het algemeen. Ook als volgens de ontwikkelaars van de tool geen verdere technologische specificatie nodig zou zijn, wordt dit zelden geëxpliciteerd in de gebruiksaanwijzing van de tool. Om echt bruikbaar te zijn voor ontwikkelaars, projectmanagers en andere, is een verduidelijking van de technologische reikwijdte van belang. Hoe vager de afbakening van de technologie, en hoe minder concreet zaken als bv. het doelpubliek zijn gedefinieerd, hoe vager de tool vaak blijft en hoe beperkter de inzetbaarheid en vindbaarheid ervan.
- Hieraan gerelateerd is het evenmin duidelijk hoe 'ethiek' wordt verstaan door de ontwikkelaars van de tool. Sommige auteurs kijken heel specifiek naar één aspect van ethiek (bv. bias) terwijl anderen niet specifiëren hoe ze ethiek invullen, welke specifieke waarden hiermee worden be-

3 Er bestaan veel verschillende classificaties van de technologieën die onder 'artificiële intelligentie' worden verstaan. Terwijl hier nog geen eensgezindheid rond is, vertrekken de auteurs van dit document graag van de AI knowledge Map zoals samengesteld door Francesco Corea: https://medium.com/@Francesco_AI/ai-knowledge-map-how-to-classify-ai-technologies-6c073b969020

doeld of welke invulling zij specifiek geven aan deze ethische begrippen.

- Slechts een klein aantal tools zijn in het Nederlands opgesteld. Terwijl het gebruikte jargon in de IT-sector algemeen gesproken Engels is, kan de anderstaligheid van deze instrumenten toch een barrière vormen in het gebruik door personen die Nederlands als moedertaal hebben. Voor ethische reflecties is een andere woordenschat nodig dan deze die IT'ers doorgaans hanteren, en in de vertaling gaan vaak nuances verloren die vanuit ethisch oogpunt van groot belang kunnen zijn. Aangezien het doelpubliek vaak bestaat uit product- en projectmanagers, gegevensbeheerders, ontwikkelaars, etc. gaat het doorgaans niet om personen die ethische begrippen op dagelijkse basis hanteren, en bijgevolg is verduidelijking van de begrippen van groot belang. Zoals eerder gezegd, worden ethische begrippen bij de richtlijnen voor de tools helaas vaak niet gedefinieerd. Samen met anderstaligheid kan dit een mogelijke barrière vormen.
- Een aantal tools kijken niet of slechts in beperkte mate naar de context van de technologie: het gebruik van het tool, de gebruikersnoden of de maatschappelijke meerwaarde die de technologie kan bieden. Zo zijn er technologie- of datasetspecifieke tools die interne 'biases' (vooroordelen) onderzoeken maar niet verder kijken naar de verschillende contexten of domeinen waarin de technologie kan worden ingezet. Terwijl deze tools een waardevol inzicht kunnen bieden, zouden ze idealiter ook kaderen hoe ze passen binnen een bredere denkoefening rond ethiek. Zo kan het doelpubliek dat met de tool aan de slag gaat goed inschatten welke aspecten nog verdere aandacht nodig hebben. Voor dit rapport hebben we de tools zo goed als mogelijk proberen te kaderen en te tonen op welk aspect van het proces ze focussen, en welke laag van het systeem ze onderzoeken. Met deze indeling hopen we anderen wegwijs te maken in de veelheid van tools, te tonen welke aspecten de tool specifiek van naderbij bekijkt en of de context al mee in rekening wordt gebracht.
- Cursussen zoals deze aangeboden door de Universiteit van Utrecht, of de cursus verbonden aan het Data Ethics Canvas, bieden vermoedelijk een waardevolle methodologie om ethische aspecten te verkennen. Niettemin zijn deze cursussen vaak betalend, wat ze zeker voor kleinere bedrijven minder toegankelijk maakt.
- Bijna alle tools vertrekken van een zelfbeoordeling of zelfinschatting door de beoogde doelgroep. Terwijl dit een waardevol vertrekpunt kan zijn, is het risico aanzienlijk dat het wereldbeeld van de betrokken doelgroep het oordeel sterk kan kleuren.
- Zeer weinig tools gaan dieper in op de bruikbaarheid van de tools. Verschillende tools bieden een stappenplan, richtlijnen of een theoretisch kader, maar in welke mate zijn die voldoende uitgewerkt om te kunnen worden ingezet door de beoogde doelgroep en in welke mate leveren ze resultaten op die de technologie en de context ervan substantieel beïnvloeden?

Ondanks deze kritische bedenkingen bieden de meeste van deze tools toch een waardevolle bijdrage aan de zoektocht naar hoe we over ethiek kunnen nadenken in het kader van AI, en hoe we dat denken kunnen vormgeven en de resultaten van de oefening zo waardevol mogelijk kunnen maken.

Het Kenniscentrum Data & Maatschappij beoogt met zijn activiteiten hier op verschillende manieren toe bij te dragen. We overwegen op korte en middellange termijn volgende activiteiten in dit kader:

- Cocreatie workshops met stakeholders. Om de bruikbaarheid van deze tools en hun mogelijke verbeterpunten verder te kunnen bepalen, plannen we deze tools in nog meer detail en samen met stakeholders te bestuderen. In dialoog met de stakeholders, willen we ook verkennen hoe deze tools meer inzetbaar kunnen worden gemaakt, en zo verder kunnen worden verbeterd.
- Op basis van dit document, workshops met stakeholders, en nieuw onderzoek wil het Kenniscentrum ook eigen tools ontwikkelen die het ethische denken in het kader van AI kunnen kaderen en faciliteren.
- Een verhoogde kennis over AI, niet enkel bij de doelgroepen van deze tools, maar ook bij de ethici, techniekfilosofen en andere personen die ze ontwikkelen, kan helpen om het ethische denken over deze technologieën concreter te maken. AI-technologieën moeten gedemystificeerd worden, zodat ook de ethische gevoeligheden van de technologieën beter kunnen worden aangeduid. Het Kenniscentrum Data & Maatschappij zal hiertoe bijdragen door onder meer 'fact sheets' te ontwikkelen die telkens meer duidelijkheid brengen over specifieke elementen van AI.
- Aangezien de meeste tools momenteel vertrekken van zelfbeoordeling en zelfinschatting, kan het Kenniscentrum helpen om richtlijnen op te stellen waarmee ontwikkelaars de afwegingen en de beslissingen die ze maken helder te documenteren. Ook kunnen er richtlijnen komen over de manier waarop ontwikkelaars hiervoor transparantie kunnen creëren bij gebruikers van de technologie (of andere externe partijen).
- Het Kenniscentrum zal ook onderzoeken in welke mate certifiëren een oplossing kan bieden om het ethische karakter van AI-technologie al zeker voor gebruikers en andere externe stakeholders transparanter te maken.

Kenniscentrum
Data & Maatschappij

Kenniscentrum Data & Maatschappij
Pleinlaan 9
1050 Brussel

info@data-en-maatschappij.ai
www.data-en-maatschappij.ai