Ethische aspecten bij het ontwikkelen en toepassen van AI

Een methode voor reflectie en deliberatie

Marc Steen*

De overheid moet allerlei taken uitvoeren en wil dat efficiënt en effectief doen. Daarom maakt zij gebruik van computers, data en software, van algoritmen en van wat tegenwoordig 'AI' wordt genoemd: artificiele intelligentie. Die term staat tussen aanhalingstekens omdat 'AI' dikwijls te pas en te onpas wordt gebruikt voor systemen die niet echt *artificieel* zijn, omdat ze zijn gebaseerd op enorm veel menselijke arbeid, bijvoorbeeld het handwerk van het labelen van trainingsdata en het finetunen en corrigeren van taalmodellen (Crawford 2021), en niet echt *intelligent*, althans niet intelligent op een menselijke manier (Runciman 2023); AI-systemen kunnen allerlei domme fouten maken omdat ze geen *common sense* hebben (Russell 2019).
In dit artikel zullen we ons richten op de systemen en algoritmen die de overheid op dit moment gebruikt (Van Veenstra e.a. 2021a). Dat zijn meestal relatief eenvoudige algoritmen, vergeleken met *state-of-the-art* systemen op basis van *deep learning*, waarbij een 'artificial neural network' wordt getraind op basis van enorme hoeveelheden data, zoals ChatGPT. Denk ook aan een algoritme dat het Centraal Justitieel Incassobureau (CJIB) gebruikt om in te schatten of iemand een boete wel of niet zal betalen, zodat het CJIB diegene kan helpen om niet (verder) in de schulden te komen.[1] Zo'n algoritme is gebaseerd op relatief eenvoudige *als-dan*-regels, bijvoorbeeld: *als* [vorige bekeuringen keurig betaald], *dan* [stuur standaard herinnering]. Die eenvoud heeft voordelen, bijvoorbeeld voor de transparantie. Zo kun je als ont-

* Dr. ir. M. Steen is senior onderzoeker Responsible Innovation bij TNO. Dit artikel is deels gebaseerd op resultaten van het NWO NWA-project VWData (www.nwo.nl/projecten/40017605-0), waarin TNO (o.a. Marc Steen), TU Delft (o.a. Ibo van de Poel en Paul Hayes) en het Ministerie van Justitie en Veiligheid (o.a. Remco Boersma) samenwerkten (zie Hayes e.a. 2020, 2023; Steen e.a. 2021a, 2021b).

1 Zie www.cjib.nl/minnelijke-schuldregeling.

doi: 10.5553/JV/016758502024050001008

wikkelaar, als uitvoerend ambtenaar, en zelfs als burger redelijk gemakkelijk inzicht krijgen in hoe het algoritme werkt; je kijkt daarvoor naar die *als-dan*-regels. Bij *deep learning* is dat een heel ander verhaal. In ChatGPT bepalen miljarden parameters in een 'artificial neural network' de werking van het systeem. Niemand kan precies volgen hoe het werkt in detail; 'als het maar werkt' is het motto. Echter, in een gevoelig domein zoals Justitie en Veiligheid is transparantie een belangrijk vereiste. Je moet kunnen uitleggen hoe de beslissingen die worden genomen op basis van een berekening van een AI-systeem tot stand zijn gekomen hoe het systeem (ongeveer) werkt, welke data erin gingen, hoe de berekening (ongeveer) verloopt, en wat daarom de uitkomst is.

In de volgende paragraaf behandelen we enkele ethische aspecten die spelen bij het ontwikkelen en toepassen van AI-systemen. Vervolgens bespreken we een methode waarmee mensen met verschillende achtergronden met elkaar in gesprek kunnen gaan over deze aspecten: een iteratief en participatief proces van reflectie en deliberatie. Tot slot wordt de meerwaarde besproken van transdisciplinaire samenwerking en deugdethiek voor het organiseren van dergelijke bijeenkomsten. Het doel van dit artikel is het ondersteunen van mensen die betrokken zijn bij het ontwikkelen en toepassen van AI-systemen bij het integreren van ethische aspecten in hun projecten.

Ethische aspecten

Er circuleren tientallen raamwerken met diverse ethische aspecten (Morley e.a. 2020; Prem 2023), deels gebaseerd op de traditie van *bioethics* (ethisch handelen in het biomedische domein) (Floridi e.a. 2018). Een bekend raamwerk is de *Ethics guidelines for trustworthy AI* van de High-Level Expert Group on Artificial Intelligence van de Europese Commissie (2019), een voorloper van de AI-verordening die onlangs werd aangenomen door het Europees Parlement.[2] Hierna zullen we enkele relevante aspecten kort bespreken, maar eerst iets over de terminologie. De High-Level Expert Group schrijft over *ethical prin-*

2 Zie www.digitaleoverheid.nl/nieuws/ai-verordening-aangenomen-door-het-europees-parlement/.

doi: 10.5553/JV/016758502024050001008

ciples en *key requirements*.[3] In dit artikel zullen we de term *aspecten* gebruiken, in aansluiting op de term ELSA, dat staat voor het meenemen van *Ethical, Legal, and Societal Aspects* (Van Veenstra e.a. 2021b). De term 'aspect' is relatief neutraal, zodat mensen gemakkelijker tot een gesprek komen, terwijl 'principe'onplezierig normatief kan overkomen. Ook drukt de term aspecten uit dat we de interacties tussen technologie en maatschappij als *wederzijds* begrijpen (Oudshoorn & Pinch 2003), terwijl bijvoorbeeld een term als impact de incorrecte suggestie kan wekken dat er een causale pijl in één richting wijst, *van* technologie *naar* maatschappij. Hierna bespreken we de volgende vijf aspecten: positieve bijdrage, menselijke autonomie, privacy, rechtvaardigheid en transparantie.[4]

Positieve bijdrage

Als we een gesprek starten over de ethische aspecten van bijvoorbeeld een (nieuw) algoritme dat moet helpen om fraude met sociale uitkeringen op te sporen, dan kunnen we vragen stellen zoals: Wat is de bijdrage van dit algoritme? In welke opzichten is die bijdrage positief, heeft zij toegevoegde waarde? En in welke opzichten is die bijdrage negatief, brengt zij kosten of risico's mee? We kijken naar de huidige situatie *zonder* algoritme (of met de huidige versie van het algoritme) en vergelijken die met de situatie *met* het algoritme (of met een nieuwe versie van het algoritme). We kijken naar de kosten en de baten, zowel financieel als maatschappelijk. Bij een algoritme dat fraude opspoort kunnen we kijken naar de fouten die het kan maken: de *false positives* (iemand die niet fraudeert krijgt incorrect het label 'fraudeur') en de *false negatives* (iemand die fraudeert krijgt incorrect geen label 'fraudeur'). Vooral die *false positives* kunnen veel schade veroorzaken: voor de personen in kwestie, en ook om die schade te herstellen . Omgekeerd kunnen *false negatives* leiden tot maatschap-

3 Zij schrijven over vier *ethical principles* (*respect for human autonomy, prevention of harm*, *fairness* en *explicability*) en over zeven *key requirements* (*human agency and oversight*, *technical robustness and safety*, *privacy and data governance*, *transparency*, *diversity, non-discrimination and fairness*, *societal and environmental wellbeing* en *accountability*).

4 Deze aspecten corresponderen met de *ethical principles* en de *key requirements* van de High-Level Expert Group on Artificial Intelligence (2019, p. 8). Onder 'positieve bijdrage' vallen *prevention of harm*, *technical robustness and safety* en *societal and environmental wellbeing*, onder 'menselijke autonomie' *respect for human autonomy* en *human agency and oversight*, onder 'privacy' *privacy and data governance*, onder 'rechtvaardigheid' *fairness* en *diversity, non-discrimination and fairness* en onder 'transparantie' *explicability*, *transparency* en *accountability*.

doi: 10.5553/JV/016758502024050001008

pelijke kosten omdat die fraude buiten beeld blijft en er dus geld op een verkeerde plek terechtkomt. Het is nuttig om dit soort analyses te doen, ook om de kwaliteit van zo'n nieuw systeem of applicatie te evalueren. Ook belangrijk voor het evalueren van 'positieve bijdrage' zijn technische robuustheid en veiligheid. Je hebt graag een robuust en veilig systeem.
We kunnen ook kijken naar de kosten voor mensen, voor de maatschappij en voor onze planeet om zo'n systeem te laten werken: de arbeid, bijvoorbeeld de *click workers* in Kenia die werkten aan het finetunen van ChatGPT, de hardware, waarvoor zeldzame metalen moeten worden gedolven, meestal onder heel slechte omstandigheden, en de energie, bijvoorbeeld in een datacenter, om een model te trainen en te draaien. Die kosten blijven vaak buiten beeld, als 'externaliteiten'; ze zijn ook 'verder weg' in plaats en tijd, maar het zou goed zijn om die *wel* mee te nemen in de analyse.

Menselijke autonomie

Elk systeem zal fouten maken, zoals de *false positives* en *false negatives* in het voorbeeld hierboven. Daarom is het wenselijk om menselijke autonomie te waarborgen. Concreet betekent dat bijvoorbeeld dat een gebruiker of operator van een systeem enige vrijheid (professionele discretie) heeft in het wel of niet opvolgen van de output van het systeem. Menselijke autonomie verwijst ook naar *agency* en menselijke waardigheid (betekenisvol werk). Als je je werk als gebruiker van een bepaald systeem ervaart alsof je zelf slechts een radertje bent in een grote machine, dan ervaar je waarschijnlijk weinig *agency* en wellicht ook weinig menselijke waardigheid.
De term autonomie wordt overigens ook gebruikt om te verwijzen naar de autonomie van een systeem, zoals een robot die bepaalde taken kan uitvoeren. Dit roept interessante vragen op over de verhouding tussen de autonomie van de operator en de autonomie van dat systeem. Welke taken kan een operator gerust delegeren aan het systeem? Hoe kan de operator overzicht houden? En als het misgaat, hoe kan de operator dan de controle nemen? Of kan het systeem zichzelf op pauze zetten en vragen of de operator een bepaalde beslissing neemt?
Soms denken mensen aan uitruil: hoe meer autonomie van het systeem, hoe minder autonomie voor de gebruiker en andersom. In de

doi: 10.5553/JV/016758502024050001008

praktijk is dit iets ingewikkelder. Stel je een assenstelsel voor met van links naar rechts toenemende *computer automation*, en van onder naar boven toenemende *human control* (Shneiderman 2020). Zo ontstaan vier kwadranten met verschillende typen systemen: linksonder (lage *automation*, lage *control*) apparaten die eenvoudige taken uitvoeren, zoals een muziekspeler, rechtsonder (hoge *automation*, lage *control*) apparaten die automatisch werken en waar je niet wilt ingrijpen, zoals een airbag, linksboven (lage *automation*, hoge *control*) apparaten die veel sturing en vaardigheid vereisen van de gebruiker, zoals een muziekinstrument, en rechtsboven (hoge *automation*, hoge *control*) systemen die ingewikkelde taken kunnen uitvoeren en daarbij sturing en vaardigheid vanuit de gebruiker vereisen. Veel systemen in het domein van Justitie en Veiligheid vallen in dit vierde kwadrant. Dat betekent dat we processen moeten inrichten voor *human-machine teaming* en *meaningful human control* (Steen e.a. 2023). We kijken dan niet alleen naar het technische systeem, maar vooral ook naar het sociotechnische systeem: mensen, machines en organisatie (Steen e.a. 2021c). Kunnen operators het systeem bijsturen, zodat een lerend systeem ontstaat, met een *feedback loop*?

Privacy

Als het over data, algoritmen en AI gaat, gaat het al snel over privacy, en dan vaak over privacy in de zin van *information privacy* en *data protection*. We kunnen ook kijken naar privacy in bredere zin, bijvoorbeeld zoals verwoord in artikel 8 van het Europees Verdrag voor de Rechten van de Mens (EVRM): 'Een ieder heeft recht op respect voor zijn privéleven, zijn familie- en gezinsleven, zijn woning en zijn correspondentie.' In de uitspraak van de rechtbank Den Haag inzake SyRI (Systeem Risico Indicatie) speelden beide betekenissen.[5] SyRI was een instrument waarmee instanties van de overheid gegevens van burgers konden delen en analyseren, voor het opsporen van fraude met uitkeringen, toeslagen en belastingen. In de engere zin van privacy toetste de rechter SyRI aan de beginselen van de Algemene verordening gegevensbescherming (AVG), zoals transparantie, doelbinding en datami-

5 Rb. Den Haag 5 februari 2020, ECLI:NL:RBDHA:2020:1878; zie ook www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Rechtbanken/Rechtbank-Den-Haag/Nieuws/Paginas/SyRI-wetgeving-in-strijd-met-het-Europees-Verdrag-voor-de-Rechten-voor-de-Mens.aspx.

doi: 10.5553/JV/016758502024050001008

nimalisatie. De rechter vond met name de transparantie onder de maat, ook omdat geen inzicht werd gegeven in de werking van het onderliggende algoritme. In de bredere zin van privacy toetste de rechter ook aan artikel 8 EVRM en aan de voorwaarden waaronder een overheid zou mogen ingrijpen in het privéleven van burgers (proportionaliteit, geschiktheid en subsidiariteit). Ook dit aspect vond de rechter onder de maat: (voorlopers van) SyRI leverden relatief veel *false positives*, wat leidt tot onterechte verdachtmaking van burgers, een enorme inbreuk op hun leven en onevenredig veel schade.

Rechtvaardigheid

Dat het toepassen van algoritmen kan leiden tot allerlei vormen van discriminatie is uitgebreid gedocumenteerd (Benjamin 2019; Buolamwini 2023; Eubanks 2017; Noble 2018; O'Neil 2016). Meestal is de oorzaak *bias* in de trainingsdata. Als je een model traint op data die discriminerende inhoud bevatten, dan zal het model discriminerende output produceren. Vaak heeft deze algoritmische discriminatie betrekking op huidskleur, geslacht of inkomen, en op diverse intersecties. Berucht is het voorbeeld van Google Photos in 2015; dat systeem gaf een foto van twee *African-American* tieners het label 'gorillas'.[6] Google paste de software aan, zodat het systeem de labels *gorilla*, *chimp*, *chimpanzee* en *monkey* niet meer kon produceren.[7] Een oppervlakkige *quick fix*. Hoe hadden ze het anders kunnen oplossen? Ze hadden een meer diverse of inclusieve dataset kunnen gebruiken om de software te trainen. Maar dat kost geld. Hier speelt ook een filosofische vraag: wil je de wereld die nu bestaat weergeven, of de wereld die je zou willen hebben? Als je 'directeur' of 'professor' in een zoekmachine voor afbeeldingen intypt, dan krijg je vooral witte mannen. Dat komt doordat veel directeuren en professors witte mannen zijn, en in de trainingsdata zitten. Echter, veel mensen zouden het toejuichen als meer mensen van kleur of vrouwen dergelijke functies zouden invullen. Dan zou je die diversiteit dus ook willen zien in de zoekresultaten.
Een ander bekend voorbeeld is COMPAS: een algoritme in de Verenigde Staten dat recidivisme van veroordeeleden voorspelt.[8] Rechters gebruikten die voorspellingen om de borgtocht te bepalen. De *false*

6 Zie www.businessinsider.com/google-tags-black-people-as-gorillas-2015-7.
7 Zie www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.
8 Zie www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

doi: 10.5553/JV/016758502024050001008

positives gingen vooral naar mensen met een donkere huidskleur (zij werden onterecht verdacht van hoge recidive) en de *false negatives* vooral naar mensen met een lichte huidskleur (zij werden onterecht *niet* verdacht van hoge recidive). Het is lastig om hiervoor een rechtvaardige oplossing te vinden. Dit is ook wiskundig erg lastig, omdat misdaad niet random is verdeeld over de bevolking (Fry 2018, p. 77-80; Lagioia e.a. 2023).

Hier spelen ook maatschappelijke en politieke vragen. Als we zien dat een algoritme discriminerende output levert, dan kunnen we dat algoritme proberen te *fixen*. Echter, zolang dat algoritme opereert binnen een maatschappij met allerlei discriminatie zal ook het algoritme discrimineren. Een grondiger aanpak zou kunnen helpen, door het onderwerp misdaad in een breder kader te plaatsen, tegen een achtergrond van armoede, opleiding, huisvesting, gezondheid en gezinssamenstelling. Als je de problemen in die achtergrond gaat oplossen, werkt dat als preventie tegen de daaraan gerelateerde misdaad. Er is echter veel maatschappelijke en politieke inspanning nodig om discriminatie tegen te gaan (Binns 2018), het fixen van één algoritme is niet voldoende.

We kunnen ook kijken naar *material* en *procedural fairness*. *Material fairness* gaat over de inhoud van een beslissing of handeling en *procedural fairness* over de wijze waarop die beslissing of handeling tot stand komt (Steen e.a. 2021c). Hoe effectief is in de praktijk het proces waarmee burgers inzicht kunnen krijgen in de werking van een bepaald algoritme en bezwaar kunnen maken tegen de uitkomsten? Deze procedurele rechtvaardigheid is vaak een punt van zorg. Aristoteles noemde rechtvaardigheid de belangrijkste deugd. Niet verwonderlijk, want rechtvaardigheid werkt door in andere aspecten. Hoe wordt bijvoorbeeld die *positieve bijdrage* verdeeld over verschillende groepen of individuen? En hoe worden de kosten en risico's verdeeld? Dergelijke vragen spelen ook voor autonomie en privacy. Welke groepen of individuen krijgen meer autonomie, en welke minder? En op de privacy van welke groepen of individuen wordt vooral inbreuk gemaakt?

Transparantie

Een programma met eenvoudige *als-dan*-regels is transparant: je kunt begrijpen en voorspellen hoe het werkt. Als je eigen risico voor zorg-

doi: 10.5553/JV/016758502024050001008

kosten 385 euro is, en je krijgt een rekening van 500 euro, dan moet je de eerste 385 euro zelf betalen en zal de zorgverzekeraar de resterende 115 euro aan je uitkeren. Dat is helder. Maar transparantie kan ook flink tekortschieten, en dat gebeurt vooral bij geavanceerde en complexe systemen. Denk aan systemen op basis van *deep learning*, de bekende Large Language Models bijvoorbeeld, zoals ChatGPT. Zulke systemen, en ook de processen eromheen, worden vaak gekarakteriseerd als een *black box* (Pasquale 2016), met een negatieve betekenis: 'Computer says no', zonder verdere toelichting. Maar hoe moet het dan wel?

Transparantie gaat niet over een dikke stapel papier waarop de software is uitgeprint. Zelfs *computer scientists* en *software engineers* zullen het knap lastig vinden om daar iets van te begrijpen. In plaats daarvan kunnen we streven naar transparantie in pragmatische betekenis: is een bepaald aspect van een systeem helder voor een bepaald persoon en kan hij of zij er een bepaalde taak (beter) mee uitvoeren (Hayes e.a. 2023). Een systeem is transparant als een auditor de *fairness* van dat systeem in praktische zin kan evalueren. Of als mensen kunnen begrijpen wanneer hun gedrag het label 'fraude' krijgt, zodat ze hun gedrag daarop kunnen aanpassen. Deze vorm van transparantie wordt vaak 'uitlegbaarheid' of 'begrijpbaarheid' (*explicability*) genoemd. Verder hangt transparantie samen met *accountability*, dat ook een juridisch begrip is en een bepaalde vorm en mate van transparantie vereist.

Reflectie en deliberatie

Hoe kunnen we aan de slag met dergelijke ethische aspecten? Hoe kunnen mensen die betrokken zijn bij het ontwikkelen en toepassen van algoritmen of AI-systemen ethische aspecten meenemen? Uiteraard zijn er allerlei methoden om dat te doen (Reijers e.a. 2018). Hierna zal ik één methode beschrijven, bij de ontwikkeling waarvan ik betrokken was en die ik meerdere keren heb toegepast, onder andere in de context van Justitie en Veiligheid (Steen e.a. 2021a). Deze methode bouwt voort op de tradities van *human-centred design*, *value sensitive design* en *responsible innovation*, en gaat over het organiseren van een iteratief en participatief proces van reflectie en deliberatie. Om aan te sluiten bij methoden die in de praktijk worden gebruikt,

doi: 10.5553/JV/016758502024050001008

noemden we deze methode *Rapid Ethical Deliberation* (*Rapid* om weg te komen van associaties met ethiek als langzaam en stroperig; en *Deliberation* om aan te geven dat het niet gaat om het éénmalig doorlopen van een checklist). De methode bestaat uit drie (iteratieve) stappen (proces) en vier ethische perspectieven (inhoud). De methode helpt een groep mensen om samen te werken en ethische aspecten systematisch en zorgvuldig mee te nemen in een concrete casus.

De eerste stap is het gezamenlijk vaststellen van aspecten of issues die kritisch zijn of zouden kunnen worden. De tweede stap gaat over het organiseren van gesprekken over deze aspecten of issues, bijvoorbeeld in het projectteam of, liever nog, met mensen die het systeem gaan gebruiken in de praktijk en mensen die verstand hebben van bepaalde aspecten of issues, zoals mensenrechten of discriminatie. In de derde stap vertaalt het projectteam de inzichten uit deze gesprekken naar het project. Op basis van de uitkomsten kunnen ze beslissingen nemen en dingen uitproberen in de praktijk. Dit is natuurlijk de belangrijkste stap; zonder deze stap zou ethiek een theoretische exercitie zijn. Inhoudelijk bestaat de methode uit het expliciet maken van vier verschillende ethische perspectieven op het systeem dat wordt ontwikkeld:[9]

- gevolgenethiek: hier kijken we naar de gevolgen van een bepaalde technologie of toepassing en drukt deze uit in voordelen (plussen) en nadelen (minnen); het doel is het maximaliseren van voordelen en het minimaliseren van nadelen, en het voorkomen van ongewenste gevolgen;
- plichtethiek: hier kijken we enerzijds naar de plichten van onder andere de ontwikkelaar of aanbieder van een bepaald systeem, en anderzijds naar de rechten van degenen die te maken krijgen met dit systeem; dit perspectief gaat ook over menselijke waardigheid en autonomie;
- relatie-ethiek: hier kijken we naar de wijze waarop technologie de interacties tussen mensen (en tussen mensen en natuur) kan veranderen; en ook naar de wijzen waarop de inzet van een bepaalde technologie bestaande machtsverhoudingen kan veranderen;
- deugdethiek: hier kijken we naar hoe mensen technologie kunnen gebruiken om bepaalde deugden te cultiveren, met als doel goed

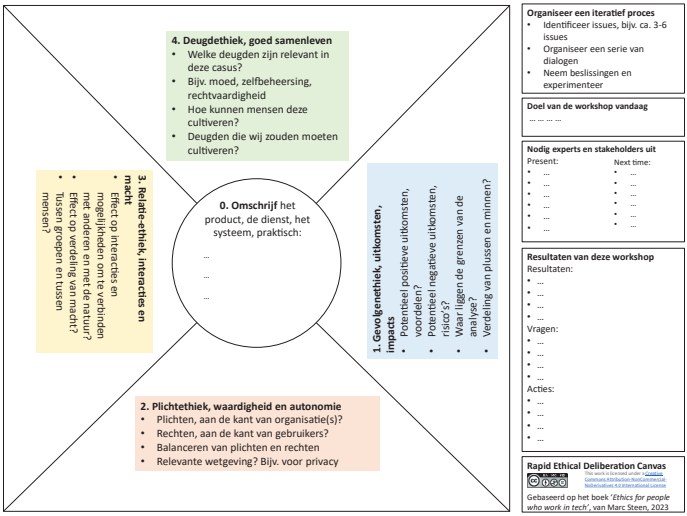
9 Vanuit deze vier perspectieven kun je kijken naar een project of naar bepaalde ethische aspecten.

doi: 10.5553/JV/016758502024050001008

met elkaar samen te leven (Vallor 2016), en naar deugden die mensen nodig hebben bij het ontwikkelen of toepassen van technologie (Steen 2022).

Deze vier perspectieven worden vaker naast elkaar gebruikt, want elk perspectief legt andere accenten; daarom is het nuttig om ze parallel te gebruiken (Steen 2022; Van de Poel & Royakkers 2011). Hiervoor is een *Rapid Ethical Deliberation Canvas* beschikbaar, een hulpmiddel om zo'n bijeenkomst of workshop te faciliteren (zie figuur 1). Hierna volgen ter illustratie drie voorbeelden van de toepassing van deze methode op projecten met de politie. Het gaat om relatief kleinschalige workshops met steeds ongeveer drie mensen van het projectteam.[10]

Figuur 1 **Rapid Ethical Deliberation Canvas, voor het faciliteren van ethische reflectie en deliberatie (https://ethicsforpeoplewhoworkintech.com)**



10 De voorbeelden komen uit Rapid Ethical Deliberation-workshops die TNO heeft uitgevoerd in verschillende projecten met de politie. In de workshops hebben we voor ieder project alle vier de ethische perspectieven gebruikt; hier noemen we steeds één perspectief ter illustratie (voor details: Steen e.a. 2021a).

doi: 10.5553/JV/016758502024050001008

Voorbeelden

Het eerste voorbeeld betreft een project voor een systeem in het Real-Time Intelligence Centre (RTIC) van de meldkamer van de noodhulpdiensten. Als een melding van een incident binnenkomt bij 112, dan kan een RTIC-operator naar aanvullende informatie zoeken, parallel aan het inschakelen van hulp. Dit moet snel gebeuren en daarom wordt overwogen om delen van deze taak te automatiseren. Maar welke delen kan het systeem automatisch uitvoeren, en welke delen moet een operator doen? Tijdens de workshop praatten de deelnemers vooral over menselijke autonomie, rechtvaardigheid en transparantie. Onder andere het perspectief van gevolgenethiek was hierbij nuttig. Wat is de mogelijke meerwaarde van zo'n systeem en wat zijn mogelijke nadelen? Wat zijn de sterke punten van wat operators kunnen doen, en wat zijn hun tekortkomingen? Daarbij kwamen cognitieve *biases* ter sprake, zoals *confirmation bias*, een voorkeur voor bevestigende informatie. Zo ontstond het idee voor een systeem dat operators bewust maakt van hun cognitieve *biases*, zodat ze hun zoekstrategieën kunnen aanpassen en ook kritisch kunnen kijken naar de output van het systeem. Dit systeem zou passen in het kwadrant van hoge *computer automation* en hoge *human control*.
Een tweede voorbeeld: een algoritme dat de kwetsbaarheid van individuen voor gewelddadige radicalisering inschat (Multi-Agency Vulnerability Assessment Support Tool, MAVAST). De vraag was of (een aangepaste versie van) dit algoritme kan worden gebruikt voor het inschatten van risico's dat jongeren afglijden naar ondermijnende criminaliteit.[11] Het project startte met het formuleren van enkele uitgangspunten, bijvoorbeeld dat het algoritme geen getallen geeft als uitkomst, zoals vaak wordt gedaan met algoritmen, een getal tussen 0 en 1 voor waarschijnlijkheid. In plaats daarvan geeft het indicatoren, en experts interpreteren die indicatoren. Hierbij spelen menselijke autonomie, rechtvaardigheid en transparantie. De experts kunnen hun oordeelsvermogen gebruiken en een stukje *in* het algoritme kijken. Onder andere relatie-ethiek was hier behulpzaam. Hoe verandert zo'n algoritme de relaties en interacties tussen de betrokken mensen?

11 Zie www.om.nl/onderwerpen/ondermijnende-criminaliteit: 'Georganiseerde ondermijnende criminaliteit is misdaad die maatschappelijke structuren of het vertrouwen daarin schaadt.' Ondermijnende criminaliteit is iets anders dan gewelddadige radicalisering, en het proces van daarnaartoe afglijden ook, vandaar de vraag.

doi: 10.5553/JV/016758502024050001008

Zo kwam het gesprek op onderliggende aannames in het project. Is dit algoritme bedoeld om 'kwetsbare mensen' te vinden en hun hulp aan te bieden, of om 'mogelijke criminelen' te vinden en hen in de gaten te houden? En hoe denken de organisaties die betrokken zijn bij dit project hierover? Hebben ze verschillende beelden, dan is het nuttig om die aannames en beelden met elkaar te bespreken.
Het derde voorbeeld betreft een project over de rol van burgers in buurtpreventie, de bekende WhatsApp-buurtgroepen. Vanuit de politie gezien is het niet wenselijk, en ook niet haalbaar, om actief deel te nemen aan zulke groepen. De politie neemt wel deel in zogenaamde beheerdersgroepen; daarin zitten de moderators van buurtgroepen. Als er in een buurtgroep iets geks gebeurt, dan kan de beheerder 'opschalen' naar een beheerdersgroep. Hierbij spelen vragen over privacy en rechtvaardigheid. In zulke groepen bestaat het risico dat mensen slordig omgaan met persoonlijke gegevens en anderen stigmatiseren en discrimineren, bijvoorbeeld op basis van vooroordelen over uiterlijk. Hier was onder andere deugdethiek waardevol.[12] We startten met de centrale deugden van de politie: 'dienstbaar' en 'waakzaam', te lezen op elke politieauto. Die deugden kunnen richting geven aan wat ze doen in die beheerdersgroepen. Bijvoorbeeld het vinden van het juiste midden: dienstbaar, maar niet overdreven dienstbaar en voortdurend in actie komen, en waakzaam, maar niet overdreven waakzaam en alles in de gaten houden. Voor burgers bespraken we deugden als zelfbeheersing en rechtvaardigheid, om te voorkomen dat elke melding een hoop gedoe en verdachtmaking oplevert. Burgers zouden bijvoorbeeld niet te snel conclusies moeten trekken (zelfbeheersing) en naar beide kanten van een verhaal moeten luisteren (rechtvaardigheid). Beheerders van WhatsAppgroepen zouden dat soort aanbevelingen kunnen meenemen in hoe ze die groepen beheren.

Aan de slag

We kunnen dus bijeenkomsten organiseren waarin de mensen die betrokken zijn bij het ontwikkelen en toepassen van een AI-systeem die drie (iteratieve) stappen doorlopen: expliciet maken van ethische aspecten, met elkaar in gesprek gaan over die aspecten vanuit ver-

12 Zie ook https://ccv-secondant.nl/platform/article/ethische-vragen-rondom-whatsapp-buurtpreventie.

doi: 10.5553/JV/016758502024050001008

schillende perspectieven, en praktische beslissingen nemen. Natuurlijk komt er uit één bijeenkomst geen kant-en-klaar antwoord. Wel komen er acties uit naar voren om dingen verder uit te zoeken of verder uit te werken.
Er zijn allerlei factoren die het proces, de uitkomsten en het succes van zulke bijeenkomsten kunnen beïnvloeden, zoals: Wie neemt het initiatief? Wie bepaalt wie wel of niet wordt uitgenodigd? Wie coördineert of faciliteert de bijeenkomst? Wie bepaalt de agenda? En wie is verantwoordelijk voor acties die voortkomen uit zo'n bijeenkomst? Wie wel eens zulke bijeenkomsten heeft meegemaakt, heeft kunnen ervaren dat samenwerking niet altijd vanzelf gaat. Mensen rondom een tafel zetten is niet voldoende. Vaak begrijpen deelnemers elkaar niet goed, gebruiken ze andere woorden, hebben ze andere doelen, enzovoort. Hierna worden twee manieren besproken om zo'n samenwerking te faciliteren: transdisciplinair samenwerken en deugdethiek.

Transdisciplinair samenwerken

Sinds een paar jaar wordt de term ELSA gebruikt voor het organiseren van samenwerking tussen mensen met expertise op Ethical, Legal en Societal Aspects (dus niet alleen ethische aspecten) in het ontwikkelen en toepassen van technologie, bijvoorbeeld AI-systemen (Ryan & Blok 2023; Van Veenstra e.a. 2021b). ELSA gaat over *transdisciplinair* samenwerken (McPhee e.a. 2018); dit gaat verder dan multidisciplinair of interdisciplinair samenwerken en richt zich expliciet op het begrijpen en oplossen van maatschappelijke vraagstukken. Zo'n ELSA-aanpak is een beloftevolle ontwikkeling. Twee opmerkingen zijn echter op hun plaats. Ten eerste komt de potentie van ELSA pas uit de verf als het wordt uitgevoerd als iteratief proces. Het zou jammer zijn om slechts één keer, bijvoorbeeld helemaal aan het begin of helemaal op het einde van een project, één ELSA-bijeenkomst te organiseren. ELSA heeft juist meerwaarde als het wordt geïntegreerd in het proces van ontwikkelen en toepassen van een specifieke applicatie. Dan kunnen de inzichten die voortkomen uit die bijeenkomsten immers worden toegepast in het project: bij het praktisch verder ontwikkelen en praktisch toepassen van die specifieke applicatie. Ten tweede is het nuttig om de rol van technologie te verhelderen. Er kan soms een onhandige dynamiek ontstaan als de mensen die aan de technologie werken en de mensen met verstand van ELSA tegenover elkaar staan: de 'tech-

doi: 10.5553/JV/016758502024050001008

neuten' tegenover de 'critici'. Het kan behulpzaam zijn als in hun samenwerking niet de techniek centraal staat, maar een bepaald vraagstuk. De diverse experts kijken dan naar dat vraagstuk vanuit verschillende perspectieven en gaan met elkaar in gesprek: experts vanuit ELSA en vanuit technologie, en ook experts vanuit de praktijk, operators die het systeem gaan gebruiken, of burgers die te maken krijgen met de uitkomsten van het systeem. Zo kunnen ze 'schouder aan schouder' werken, zowel aan het beter begrijpen van het probleem als aan het ontwikkelen van mogelijke oplossingen.

Deugdethiek

Deugdethiek is op twee manieren nuttig: om te begrijpen hoe mensen technologie kunnen gebruiken om bepaalde deugden te cultiveren, en om te begrijpen welke deugden mensen nodig hebben wanneer ze technologie ontwikkelen of toepassen. We zagen dat ook in het voorbeeld van de WhatsApp-buurtgroepen. Voor het stimuleren van samenwerking tussen verschillende experts en professionals kan het nuttig zijn om te kijken naar de deugden die ze nodig hebben.
Als je als ontwikkelaar werkt aan een AI-systeem en wilt dat dat AI-systeem niet gaat discrimineren, dan zul je de deugd van rechtvaardigheid moeten cultiveren, bijvoorbeeld door goed te kijken naar de diverse plussen en minnen (en *false positives*) en naar hoe die plussen en minnen worden verdeeld over verschillende groepen. Als je als operator wilt dat aspecten zoals non-discriminatie of transparantie goed worden uitgewerkt in zo'n systeem, dan zul je moed moeten cultiveren om je vinger op te steken en vragen te stellen. Anderen kunnen dat als irritant ervaren, want er liggen al genoeg issues op tafel: technische dingen, juridische vereisten, een deadline en een budget. En als projectleider heb je praktische wijsheid nodig: op welk moment laat ik wie aan het woord, en hoe kunnen we voldoende ethische reflectie en deliberatie combineren met voldoende voortgang in het project?
Ook hier twee opmerkingen. Als eerste een mogelijk misverstand. Sommige mensen associëren deugdethiek met iets dat zich binnen één persoon afspeelt. Dit klopt deels: het cultiveren van een bepaalde deugd speelt zich af binnen in een persoon. Dat gebeurt echter niet in een vacuüm. Mensen cultiveren deugden in interactie met andere mensen. Aristoteles, een van de grondleggers van de westerse deugdethiek, gaf les in politiek bestuur, het creëren van omstandigheden

doi: 10.5553/JV/016758502024050001008

waarin mensen goed kunnen samenleven, in een *polis*. Hij doceerde geen zelfhulp voor individueel welzijn.
Als tweede opmerking een vraag: welke deugden hebben de mensen nodig bij transdisciplinair samenwerken, bijvoorbeeld in zo'n ELSA-aanpak? Een goed startpunt bieden de vier klassieke deugden: rechtvaardigheid, moed, zelfbeheersing en praktische wijsheid. Daarnaast hebben we hedendaagse deugden nodig, zoals eerlijkheid en burgerschap. Eerlijkheid heb je nodig als je werkt aan generatieve AI, bijvoorbeeld om ervoor te zorgen dat de output van zo'n systeem zo veel mogelijk eerlijk en betrouwbaar is. En burgerschap heb je bijvoorbeeld nodig als je werkt aan AI in social media, bijvoorbeeld om constructieve interacties te stimuleren en polarisatie tegen te gaan. Het is overigens geen probleem als verschillende mensen verschillende deugden cultiveren; dan kunnen ze elkaar aanvullen. Een voorbeeld: stel dat Rachel en Jaap werken aan een AI-systeem dat fraude opspoort; Rachel cultiveert bijvoorbeeld compassie en nieuwsgierigheid (hoe zou een operator of een eindgebruiker dit systeem gebruiken en ervaren?) en Jaap zelfbeheersing en creativiteit (die extra feature zou leuk zijn, maar welk effect kan die hebben op fairness en op transparantie?) (Steen e.a. 2021b).
Kortom: ethiek is geen kijksport – je kunt niet vanaf een afstand toekijken en dingen roepen – het is een teamsport; het gaat om het organiseren van praktische kritische reflectie en deliberatie, en om het nemen van beslissingen over het praktisch ontwikkelen en toepassen van technologie.

Literatuur

Benjamin 2019
R. Benjamin, *Race after technology: abolitionist tools for the new Jim Code*, Cambridge: Polity 2019.

Binns 2018
R. Binns, 'Fairness in machine learning: lessons from political philosophy', *Proceedings of Machine Learning Research* (81) 2018, p. 149-159.

doi: 10.5553/JV/016758502024050001008

Buolamwini 2023
J. Buolamwini, *Unmasking AI: my mission to protect what is human in a world of machines*, Londen/New York: Penguin Random House 2023.

Crawford 2021
K. Crawford, *Atlas of AI: power, politics, and the planetary costs of artificial intelligence*, New Haven/Londen: Yale University Press 2021.

Eubanks 2017
V. Eubanks, *Automating inequality*, New York: St. Martin's Press 2017.

Floridi e.a. 2018
L. Floridi e.a., 'AI4People. An ethical framework for a good AI society: opportunities, risks, principles, and recommendations', *Minds and Machines* (28) 2018, p. 689-707.

Fry 2018
H. Fry, *Hello world. How to be human in the age of the machine*, Londen: Transworld 2018.

Hayes e.a. 2020
P. Hayes, I. van de Poel & M. Steen, 'Algorithms and values in justice and security', *AI & Society* (35) 2020, p. 533-555.

Hayes e.a. 2023
P. Hayes, I. van de Poel & M. Steen, 'Moral transparency of and concerning algorithmic tools', *AI and Ethics* (3) 2023, p. 585-600.

High-Level Expert Group on Artificial Intelligence 2019
High-Level Expert Group on Artificial Intelligence, *Ethics guidelines for trustworthy AI*, Brussel 2019.

Lagioia e.a. 2023
F. Lagioia, R. Rovatti & G. Sartor, 'Algorithmic fairness through group parities? The case of COMPAS-SAPMOC', *AI & Society* (38) 2023, p. 459-478.

McPhee e.a. 2018
Ch. McPhee, M. Bliemel & M. van der Bijl-Brouwer, 'Editorial: transdisciplinary innovation', *Technology Innovation Management Review* (8) 2018, p. 3-6.

Morley e.a. 2020
J. Morley, L. Floridi, L. Kinsey & A. Elhalal, 'From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices', *Science and Engineering Ethics* (26) 2020, p. 2141-2168.

Noble 2018
S.U. Noble, *Algorithms of oppression: now search engines reinforce racism*, New York: New York University Press 2018.

O'Neil 2016
C. O'Neil, *Weapons of math destruction*, Londen: Penguin 2016.

doi: 10.5553/JV/016758502024050001008

Oudshoorn & Pinch 2003
N. Oudshoorn & T. Pinch, *How users matter: the co-construction of users and technology*, Cambridge, MA/Londen: MIT Press 2003.

Pasquale 2016
F. Pasquale, *The black box society. The secret algorithms that control money and information*, Cambridge, MA: Harvard University Press 2016.

Van de Poel & Royakkers 2011
I. van de Poel & L. Royakkers, *Ethics, technology, and engineering: an introduction*, Chichester: John Wiley and Sons 2011.

Prem 2023
E. Prem, 'From ethical AI frameworks to tools: a review of approaches', *AI and Ethics* (3) 2023, p. 699-716.

Reijers e.a. 2018
W. Reijers e.a., 'Methods for practising ethics in research and innovation: a literature review, critical analysis and recommendations', *Science and Engineering Ethics* (24) 2018, afl. 5, p. 1437-1481.

Runciman 2023
D. Runciman, *The handover. How we gave control of our lives to corporations, states and AIs*, Londen: Profile Books 2023.

Russell 2019
S. Russell, *Human compatible: AI and the problem of control*, Londen: Allen Lane 2019.

Ryan & Blok 2023
M. Ryan & V. Blok, 'Stop reinventing the wheel: or how ELSA and RRI can align', *Journal of Responsible Innovation* (10) 2023, afl. 1, https://doi.org/10.1080/23299460.2023.2196151.

Shneiderman 2020
B. Shneiderman, 'Human-centered artificial intelligence: reliable, safe & trustworthy', *International Journal of Human-Computer Interaction* (36) 2020, afl. 6, p. 495-504.

Steen 2022
M. Steen, *Ethics for people who work in tech*, Boca Raton, FL: Routledge/CRC Press 2022.

Steen e.a. 2021a
M. Steen, M. Neef & T. Schaap, 'A method for Rapid Ethical Deliberation in research and innovation projects', *International Journal of Technoethics* (12) 2021, afl. 2, p. 72-85.

Steen e.a. 2021b
M. Steen, M. Sand & I. van de Poel, 'Virtue ethics for responsible innovation', *Business and Professional Ethics Journal* (40) 2021, afl. 2, p. 243-268.

doi: 10.5553/JV/016758502024050001008

Steen e.a. 2021c
M. Steen, T. Timan & I. van de Poel, 'Responsible innovation, anticipation and responsiveness: case studies of algorithms in decision support in justice and security, and an exploration of potential, unintended, undesirable, higher-order effects', *AI and Ethics* (1) 2021, afl. 4, p. 501-515.

Steen e.a. 2023
M. Steen, J. van Diggelen, T. Timan & N. van der Stap, 'Meaningful human control of drones: exploring human-machine teaming, informed by four different ethical perspectives', *AI and Ethics* (3) 2023, afl. 1, p. 281-293.

Vallor 2016
S. Vallor, *Technology and the virtues: a philosophical guide to a future worth wanting*, New York, NY: Oxford University Press 2016.

Van Veenstra e.a. 2021a
A.F. van Veenstra, F. Grommé & S. Djafari, 'The use of public sector data analytics in the Netherlands', *Transforming Government: People, Process and Policy* (15) 2021, afl. 4, p. 396-419.

Van Veenstra e.a. 2021b
A.F. van Veenstra, L. van Zoonen & N. Helberger (red.), *ELSA labs for human centric innovation in AI*, NL AI Coalitie 2021.

doi: 10.5553/JV/016758502024050001008