AI heeft geen stekker meer

Over ethiek in het ontwerpproces

Rudy van Belkom
Stichting Toekomstbeeld der Techniek

AI HEEFT GEEN STEKKER MEER

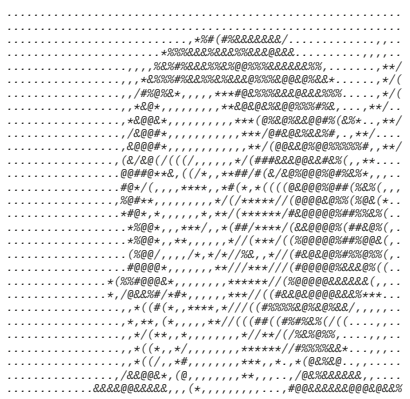
Over ethiek in het ontwerp-proces

Deel III in de reeks ‘De toekomst van artificiële intelligentie (AI)’
Keuzes maken in en voor de toekomst

Rudy van Belkom

Stichting Toekomstbeeld der Techniek (STT)

Stichting Toekomstbeeld der Techniek



Een handvat voor de toekomst van AI

Door Maria de Kleijn-Lloyd
Senior Principal, Kearney,
Voorzitter denktank STT
toekomstverkenning AI

Welke AI-toekomst willen we in Nederland? Een bijna onmogelijk te beantwoorden vraag. Want: wat is AI eigenlijk, welke mogelijke toekomst-scenario's zijn er en wie bepaalt op basis waarvan 'wat we willen'? Het derde deel van het STT-drieluik 'de toekomst van AI' gaat over deze derde, normatieve vraag. Waarbij de inzet is om over dit vraagstuk een breed maatschappelijk debat aan te zwengelen. We krijgen namelijk allemaal met AI te maken: direct als gebruikers van apps, maar ook indirect doordat personen en instanties AI gebruiken. Denk aan de dokter die een scan algoritmisch laat analyseren om een diagnose te kunnen stellen. Dit is geen toekomstmuziek; veel is allang mogelijk. De impact is vandaag al groot en zal naar verwachting steeds groter worden. Daarom is het goed om nadrukkelijk aandacht te besteden aan de bijbehorende ethische en maatschappelijke keuzes.

Er wordt al veel werk verzet. De EU heeft met een '*high level expert group*' de belangrijkste ethische AI-principes, zoals *explainability* en *fairness*, al uitgebreid beschreven. Maar daarmee zijn we er niet. Op het niveau van algemene principes is het namelijk relatief makkelijk om het met elkaar eens te zijn: natuurlijk willen we privacy, natuurlijk willen we eerlijke uitkomsten. In het denken over een vitale infrastructuur wordt ook wel gesproken over '*feel good principles*'. Natuurlijk zijn we het eens!

Het wordt pas spannend als we die principes gaan vertalen naar praktische toepassingen. Dan gaat het om twee uitdagingen. Ten eerste: het uitvoerbaar maken van het principe. Wat is bijvoorbeeld een transparant algoritme? Eentje waarvan de hele code – soms vele terabytes – is gepubliceerd, maar die alleen door een select groepje experts kan worden begrepen? Of eentje waarvoor een soort bijsluiter in lekentaal is geschreven met de belangrijkste ontwerpkeuzes, brondata, de werking en de bijwerkingen? Ten tweede: principes die in de uitvoering met elkaar botsen. Transparantie en privacy bijvoorbeeld. En ook hier is de context van belang: medische informatie is van een totaal andere orde dan Netflix-voorkeuren. Wie hakt de knoop door welk principe voorrang krijgt? We moeten met elkaar die spannende stap gaan zetten. Want ontwerpers van algoritmes maken deze keuzes al elke dag.

Dit voorwoord is geschreven middenin de '*intelligent lockdown*' van de coronacrisis, april 2020. Voor degenen die somberen dat Nederland de boot mist met digitalisering & AI: onze fysieke infrastructuur (van kabels tot cloud) blijkt het aan te kunnen en ook de meeste consumenten en bedrijven schakelen met relatief weinig moeite over naar thuiswerken. We ontwikkelen allemaal heel snel nieuwe digitale vaardigheden. Maar misschien nog wel interessanter is dat we ook massaal via de media deelnemen aan de dialoog over algoritmes en apps. Bijvoorbeeld de '*appathon*' die het ministerie van Volksgezondheid organiseerde omtrent de corona-app. Hoe maak je deze app zo dat die de privacy van burgers waarborgt, niet misbruikt kan worden én accuraat is? Kan dat allemaal tegelijk? En bedoelen we met accuraat dan 'geen coronagevallen missen' (geen *false negatives*) of 'niemand onnodig in quarantaine' (geen *false positives*)? Zo helpt de actuele situatie, hoe naar ook, ons om een aantal ethische keuzes in AI helder te krijgen. En we doen dat met 17 miljoen Nederlanders.

Ik hoop dat we, geholpen door deze studie en vooral ook door de interactieve online onderdelen, deze dialoog gericht en breed verder kunnen voeren om hem te kunnen vertalen in praktische handvatten. Zodat we een AI-toekomst krijgen in Nederland die we niet alleen accepteren, maar ook echt kunnen omarmen.

Aanpak onderzoek

Artificiële intelligentie (AI) staat hoog op de agenda bij politici, bestuurders, managers en beleidsmakers. Ondanks het feit dat er veel geschreven is over AI, is er nog veel onduidelijkheid over de wijze waarop de technologieën zich in de toekomst gaan ontwikkelen en wat de mogelijke impact op de samenleving gaat zijn. Het is daarom tijd voor een toekomstverkenning naar AI. Een groot verschil met andere technologieën is dat AI steeds meer autonoom wordt, waardoor we steeds meer beslissingen uit handen gaan geven. Dit heeft geleid tot de volgende centrale vraag:

Wat is de impact van AI op besluitvorming in de toekomst?

Menselijke besluitvorming komt tot stand door verschillende componenten. Naast feitelijke kennis spelen de beeldvorming en ambities van de betrokkenen in het besluitvormingsproces een rol. De invulling van deze componenten is sterk aan verandering onderhevig. Daarbij hebben deelnemers aan besluitvormingsprocessen vaak uiteenlopende ideeën over de werkelijkheid. Wat voor de één een onweerlegbaar feit is, is voor de ander discutabel. In dat opzicht zou je kunnen stellen dat *automated decision making* de nodige objectiviteit kan inbrengen. De vraag blijft echter in hoeverre AI menselijke besluitvormingsprocessen kan overnemen. En in hoeverre we dit kunnen en willen toestaan.

> De toekomst wordt vaak gezien als iets wat ons overkomt, maar het is eigenlijk iets waar we als mens zelf sturing aan kunnen geven. De ambitie van deze toekomstverkenning is om met een multidisciplinaire groep experts te werken aan een formulering van de gewenste toekomst van AI en daarbij in kaart te brengen wat ervoor nodig is om deze toekomst te kunnen bereiken. Dit heeft geleid tot het volgende drieluik:

Deel 1: Voorspellen

Wanneer het over de toekomst van AI gaat lijken er maar twee smaken te bestaan: de utopie en de dystopie. Vaak starten deze discussies met ethische vragen, terwijl de vraag of de technologieën deze scenario's in de toekomst daadwerkelijk kunnen waarmaken wordt overgeslagen. De focus in deel 1 ligt daarom op de technologieën:

> *Hoe verhoudt AI zich tot menselijke besluitvorming en op welke wijze gaat AI zich in de toekomst ontwikkelen?*

Hierbij is gebruik gemaakt van *technology forecasting*. Aan de hand van literatuurstudies en expertinterviews zijn de meest reële ontwikkelingsrichtingen van de technologieën in kaart gebracht (scope 0-10 jaar). Voor dit deel zijn 40 experts geraadpleegd, van AI-experts en neurowetenschappers tot psychologen en bestuurskundigen.

Deel 2: Verkennen

De vormen waarin AI in de toekomst zal worden ingezet, zijn afhankelijk van de maatschappelijke context. Naast de utopische en dystopische visie zijn er meerdere toekomstvisies te onderscheiden. De focus in deel 2 ligt daarom op de uitwerking van mogelijke toekomstscenario's:

> *Welke implicaties heeft de ontwikkelingsrichting van AI op besluitvorming in de toekomst en welke mogelijke toekomstscenario's kunnen aan de hand hiervan uiteengezet worden?*

Hierbij is gebruik gemaakt van *scenarioplanning*. Door middel van creatieve sessies zijn meerdere toekomstscenario's in kaart gebracht (scope 10-20 jaar). Voor dit deel zijn vier scenario workshops georganiseerd waaraan 30 experts uit verschillende vakgebieden deelnamen.

Deel 3: Normatief

AI kan de eerste technologie worden die zijn eigen ontwikkeling gaat bepalen. Het zou in dat geval belangrijker dan ooit tevoren zijn om de gewenste voorwaarden te bepalen voor de ontwikkelingsrichting. Welke toekomst willen we? De focus in deel 3 ligt daarom op ethische vraagstukken:

> *Welke ethische vraagstukken spelen een rol bij de impact van AI op besluitvorming in de toekomst en op welke wijze kunnen we ethisch verantwoorde AI-toepassingen ontwikkelen?*

Hierbij is gebruik gemaakt van *backcasting*. Aan de hand van een *online vragenlijst* is de gewenste toekomst van AI in kaart gebracht (scope 20-30 jaar). Vervolgens is onderzocht welke elementen noodzakelijk zijn om deze toekomst te kunnen bereiken. Voor dit deel is een vragenlijst uitgezet onder drie groepen, namelijk experts, bestuurders en studenten. Deze vragenlijst is meer dan 100 keer ingevuld.

Scope

Toekomstverkenningen van Stichting Toekomstbeeld der Techniek (STT) kennen doorgaans een scope van ongeveer 30 jaar. Deze scope heeft niet tot doel om concrete uitspraken te doen over een specifiek jaartal, maar het geeft aan dat de verkenningen zich niet beperken tot korte termijn ontwikkelingen. Het doel is juist om los te komen van de beperkingen in de huidige tijdsgeest. Hiervoor is het nodig om over grenzen heen te kijken en zo de horizon te verbreden en meer toekomstgericht te denken.

Deze verkenning borduurt voort op de inzichten uit eerdere STT-studies naar data, namelijk *Dealing with the data flood* (2002) en *Data is macht* (2017). Data vormt de grondstof voor hedendaagse AI-technieken. De vraag is of dit in de toekomst zo zal blijven of dat de technologieën zich in een andere richting gaan ontwikkelen.

» Technologie heeft vaak geen stekker meer. Die kun je er dus ook niet zomaar uittrekken. «

Denktank

Tijdens de verkenning is uitvoerig gebruik gemaakt van de expertise en ervaring van bestuurders en experts uit het werkveld. De volgende multidisciplinaire denktank is geformuleerd:

Marc Burger	Capgemini	CEO
Patrick van der Duin	STT	Directeur
Bernard ter Haar	Ministerie BZK	Buitengewoon adviseur
Frank van Harmelen	VU	Professor Knowledge Representation & Reasoning
Fred Herrebout	T-Mobile	Senior Strategy Manager
Marijn Janssen	TU Delft	Professor ICT & Governance
Maria de Kleijn-Lloyd*	Kearney	Senior Principal
Leendert van Maanen	Utrecht University	Assistant professor in Human-centred AI
Marieke van Putten	Ministerie BZK	Senior Innovation Manager
Jelmer de Ronde	SURF	Projectmanager SURFnet
Klamer Schutte	TNO	Lead Scientist Intelligent Imaging
Maarten Stol	BrainCreators	Principal Scientific Advisor

* voorzitter denktank STT-toekomstverkenning AI

Leeswijzer deel III

Artificiële intelligentie (AI) lijkt een van de meest besproken, maar misschien wel minst begrepen technologieën van dit moment. Het is op sommige vlakken ‘dommer’ dan veel mensen denken, maar op andere vlakken juist veel ‘slimmer’. Een computer als minister-president lijkt nog wat vergezocht, maar AI heeft nu al wel degelijk invloed op onze arbeidsmarkt. ‘*AI is here to stay*’. Deze intelligente systemen worden nu echter nog te vaak gezien als een doel op zichzelf. De vraag of AI de beste oplossing is voor het probleem wordt hierbij nauwelijks gesteld. Veel mensen lijken er vanuit te gaan dat AI een soort onvermijdelijke natuurwet is, waar we wel iets mee moéten. Koste wat het kost. Wat we hierbij niet moeten vergeten is dat AI zelf niet goed of slecht is, het gaat erom hoe het door mensen wordt ingezet. De vraag is dus wat voor samenleving we willen zijn, gegeven alle technologische ontwikkelingen. De maatschappij zal hoe dan ook fundamenteel veranderen en AI kan ons helpen om het juiste pad te vinden. We moeten alleen nog wel uitvinden waar we dan precies naartoe willen.

Terugblik deel I

Het overkoepelende karakter van AI maakt het een moeilijk te definiëren begrip. Een eenduidige en internationaal geaccepteerde definitie ontbreekt dan ook. In 2019 lanceerde het Nederlandse kabinet het Strategisch Actieplan voor Artificiële Intelligentie (SAPAI). Dit actieplan beschrijft de voornemens van het kabinet om de ontwikkeling van AI in Nederland te versnellen en internationaal te profileren. In het document wordt de definitie van de Europese Commissie gehanteerd: ‘AI verwijst naar systemen die intelligent gedrag vertonen door hun omgeving te analyseren en – met een zekere mate van zelfstandigheid – actie ondernemen om specifieke doelen te bereiken.’ Deze zin staat vol met multi-interpretabele containerbegrippen: Systemen? Intelligentie? Een zekere mate van zelfstandigheid? Specifieke doelen? Toch wordt op basis van deze holistische definitie een volledig actieplan ingericht. Om grip te krijgen op de ontwikkeling van AI is het van belang om begrip te krijgen van de werking en de toepassing van AI.

We zijn daarom in deel I van deze toekomstverkenning ‘Duikboten zwemmen niet’ uitvoerig ingegaan op de vragen wat AI is, hoe het werkt, op welke wijze het zich verhoudt tot menselijke intelligentie, in hoeverre menselijke besluitvorming geautomatiseerd kan worden, wat de meest dominante expertopinies zijn aangaande de ontwikkelingsrichting van AI en welke economische en politieke factoren deze ontwikkelingsrichting beïnvloeden. Het doel van dit eerste deel was om een beter geïnformeerde discussie over de toekomst van AI te kunnen voeren.

Terugblik deel II

Wanneer het over de toekomst van AI gaat wordt er vaak in extremen gedacht: belanden we in de utopie of in de dystopie? Naast het feit dat dergelijke uitersten veel beter werken in films en krantenkoppen, heeft deze tweedeling ook te maken met het idee dat hoewel de kans relatief klein is dat één van de twee ooit zal gebeuren, de mogelijke impact bij beide scenario’s zo groot kan zijn dat deze wel enige reflectie verdient. Dat geldt zowel voor de utopische visie (we hoeven nooit meer te werken) als de dystopische (we worden slaaf van de technologie). Er bestaan echter meerdere smaken.

We hebben daarom in deel II van deze toekomstverkenning ‘De computer zegt nee’ meerdere alternatieve realiteiten onderzocht en deze vertaald naar vijf toekomstscenario’s. Er is hierbij sprake van een oplopende intensiteit van AI in de verschillende scenario’s. De beperkte beschikbaarheid aan resources zorgt er in het scenario van *Game Over* voor dat de belofte die AI ooit was niet wordt waargemaakt. In het scenario van *The winner takes all* leidt de behoefte aan controle en regulatie ertoe dat

AI hoofdzakelijk wordt ingezet als tool om menselijke intelligentie te vergroten. Het streven naar automatisering zorgt er in het scenario van *Privacy for sale* voor dat de mens op verschillende gebieden wordt vervangen door AI. In het scenario van *Robot Rights* werken en leven mens en machine op een gelijkwaardige manier samen. AI overstijgt menselijke intelligentie op alle domeinen in het scenario van *The Singularity is here*. Deze toekomstscenario's helpen ons om de veranderende relatie tussen mens en technologie beter te leren begrijpen. Ook kunnen we hierdoor juist de wenselijke elementen in de toekomst identificeren, waar we vervolgens samen naartoe kunnen werken.

Waarom deze publicatie?

De wijze waarop AI in de toekomst zal worden ingezet is sterk afhankelijk van de maatschappelijke context. Dit gaat dus niet alleen over de prestaties van de technologieën en de beschikbaarheid van resources, maar ook over strategische belangen en maatschappelijke acceptatie. Wat vaak over het hoofd wordt gezien, is dat we deze context zelf creëren. De keuzes die we in het heden maken hebben grote invloed op de mogelijke toekomsten. Het is daarom van belang om ethische vraagstukken in kaart te brengen en op zoek te gaan naar antwoorden. Wie is er verantwoordelijk wanneer een AI-toepassing de fout ingaat? Kunnen we rechten verlenen aan technologieën? En hoe zorgen we ervoor dat AI-toepassingen vrij zijn van vooroordelen? Gelukkig komen er steeds meer ethische richtlijnen die de ontwikkeling van betrouwbare AI-systemen moeten waarborgen. De vraag is echter hoe je deze abstracte waarden kunt vertalen naar concrete praktijktoepassingen. Momenteel wordt er vooral veel gesproken over ethiek, maar concrete handvatten om het te integreren in het ontwerpproces ontbreken nog. Als we in de toekomst ethisch verantwoorde toepassingen willen gebruiken is het nu tijd om ethiek in praktijk te brengen.

» Many researchers will tell you that the heaven-or-hell scenarios are extremely unlikely. We're not going to get the AI we dream of or the one that we fear, but the one we plan for. Design will matter. «

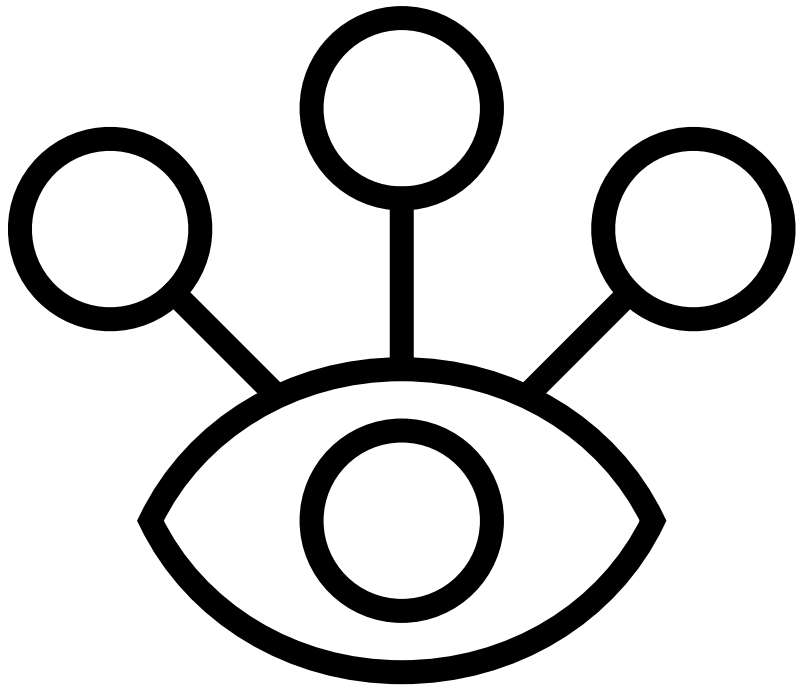
-- Stephan Talty

Voor wie is deze publicatie?

Deze publicatie is voor iedereen die geïnteresseerd is in de rol van ethiek in het ontwikkelingsproces van AI. Het doel is niet alleen om ethische vraagstukken in kaart te brengen, maar ook om te onderzoeken op welke wijze we in de toekomst ethisch verantwoorde AI-toepassingen kunnen ontwikkelen. Wil je weten op welke wijze we ethiek in praktijk kunnen brengen? En sta je open voor een nieuwe kijk op en aanpak van ethiek? Dan is deze publicatie geschikt voor jou.

Hoe is deze publicatie opgebouwd?

In het eerste hoofdstuk zoomen we verder in op AI & Ethiek. We proberen hierbij de opkomst van ethiek in AI te verklaren en kijken welke ethische vraagstukken er op dit moment spelen. Vervolgens kijken we in hoofdstuk 2 welke ethische richtlijnen er reeds zijn en wat de beperkingen daarvan zijn in de praktijk. In het derde en laatste hoofdstuk kijken we naar een nieuwe visie en aanpak voor de integratie van ethiek in de ontwikkeling van AI. Hierbij kijken we zowel naar het ontwerp als naar het proces.



1 ——————————
AI & Ethiek

Pagina’s: 46
Woorden: 10.403
Leestijd: ca. 1,5 uur

AI & Ethiek

Wanneer je de term artificiële intelligentie ↴ (AI) gebruikt, dan krijg je in de meeste gezelschappen wel de aandacht. We willen er allemaal 'iets' mee. Maar over de vraag wat AI precies is en wat we er precies mee kunnen bestaat geen consensus. Vaak wordt gesproken over 'de technologie', maar eigenlijk kun je niet van een technologie spreken. Het is de uitkomst van een aaneenschakeling van verschillende technologieën die gezamenlijk een vorm van intelligentie ↴ voortbrengen. Deze kunstmatige vorm van intelligentie kan daarbij op verschillende manieren bereikt worden. Denk bijvoorbeeld aan *whole brain emulation* (WBE), waarbij getracht wordt een volledig brein in een computer te uploaden. Wanneer het over AI gaat, dan gaat het in de meeste gevallen echter over toepassingen op het gebied van *machine learning*. Het gaat bij machine learning over een revolutie waarin mensen niet meer programmeren (als dit, dan dat), maar waarin machines zelf regels afleiden uit data. Zonder een grote hoeveelheid data ↴, rekenkracht en algoritmen ↴ geen AI.

Een vergelijkbaar principe lijkt voor ethiek te gelden; het is niet eenvoudig om tot een eenduidige opvatting erover te komen. Ethiek is een tak binnen de filosofie die zich bezighoudt met de systematische reflectie van wat als goed of juist handelen bestempeld wordt. Wanneer het over ethische vraagstukken gaat dan vinden we er allemaal wel 'iets' van. Maar vanuit welk perspectief we precies redeneren weten we vaak niet. 'De ethiek' bestaat niet. Er zijn verschillende ethische stromingen, zoals bijvoorbeeld de gevolgenethiek ↴ (waarbij de focus ligt op het resultaat van het handelen) en de beginselethiek ↴ (waarbij het uitgangspunt het beginsel is, ongeacht de gevolgen van het handelen). Denk bijvoorbeeld aan Robin Hood; hij steelt geld en goederen van de rijken om aan de armen te geven. De vraag of dit ethisch verantwoord is hangt af van het perspectief. In de gevolgenethiek zijn de acties van meneer Hood te verdedigen, het draagt immers bij aan meer gelijkwaardigheid. Maar vanuit de beginselethiek niet: stelen mag niet, ook niet als het voor een goed doeleinde is.

Het feit dat AI en ethiek beide overkoepelende begrippen zijn, waaronder verschillende stromingen en benaderingen vallen, maakt de discussie er niet eenvoudiger op. In discussies lopen verschillende AI-toepassingen en ethische stromingen dan ook vaak door elkaar heen. Het is daarom van belang om de opkomst van ethiek in AI nader te bekijken en verschillende ethische vraagstukken en stromingen te onderscheiden.

Ethiek in de spotlight

Ethiek ↴ is momenteel een *hot topic*. Zeker binnen de ontwikkeling van AI ↴ is het een veelbesproken onderwerp. Ethiek lijkt soms zelfs wel alleen voor AI gereserveerd te zijn. Maar dat is natuurlijk niet het geval. In verschillende domeinen spelen uiteenlopende ethische vraagstukken een belangrijke rol. Denk bijvoorbeeld aan de discussie over klonen in de medische biologie of over het gedrag van bankiers in de financiële sector. Toch komen we met de vraag wat 'het juiste handelen' precies is in relatie tot AI-toepassingen maar moeizaam vooruit. En dat is eigenlijk ook niet zo vreemd.

> Naargelang AI steeds autonomer kan opereren zullen we in steeds grotere mate de controle uit handen moeten gaan geven. Dit is nieuw terrein voor de mens en zorgt logischerwijs voor gevoelens van angst. Hollywood schroomt niet om deze angst verder te voeden. Scenario's over robots die in opstand komen vormen al bijna 100 jaar een populaire verhaallijn (ervan uitgaande dat 'Metropolis' uit 1927 van regisseur Fritz Lang de eerste echte sciencefiction film is waarbij een robot slechte intenties heeft). Daarbij komt dat fictie steeds vaker wordt ingehaald door de realiteit. Zo crashte in 2016 een Tesla op de automatische piloot, waarbij de bestuurder om het leven kwam. Fysiek letsel als gevolg van het gebruik van technologie is niet nieuw, maar dergelijke voorvallen zonder *human in the loop* (HITL) wel. Het volstaat niet langer om op afstand te filosoferen over wat goed of juist is. Ethiek is meer dan ooit praktische filosofie.

Ethiek in opkomst

Wie is er verantwoordelijk bij een ongeluk met een zelfrijdende auto? Hoe beschermen we onze privacy ↴ bij deze datahongerige toepassingen? En hoe kunnen we een onrechtvaardige behandeling door AI-systemen voorkomen? Het is tegenwoordig moeilijk om een AI-gerelateerde conferentie bij te wonen zonder dat een deel van de programmering gewijd is aan ethiek. Een aantal jaren geleden vormden ethici slechts het voorprogramma, tegenwoordig zijn ze op steeds meer conferenties de hoofdact. Deze shift is niet alleen zichtbaar in de programmering van conferenties, maar ook in de organisatie ervan. Dit is goed te zien bij de organisatie van een van de grootste AI-conferenties ter wereld, de *Neural Information Processing Systems conference*. Uit een grootschalige enquête uit 2017 bleek dat de conferentie geen gastvrije omgeving bood aan vrouwelijke deelnemers. Respondenten maakten melding van seksuele intimidatie en seksistische of seksueel beledigende opmerkingen en grappen. De organisatie besloot daarom in 2018 een nieuwe *code of conduct* te introduceren om discriminatie tegen te gaan. Ze zetten zich verder in om het evenement inclusiever te maken door onder andere de ondersteuning van kinderopvang. Ook pasten ze de afkorting 'NIPS' aan in NeurIPS, om de associatie met '*nipples*' te voorkomen. Een ogenschijnlijk kleine, maar toch zeer belangrijke aanpassing. In een voorgaande editie droegen sommige mannelijke bezoekers van een workshop over vrouwen in *machine learning* een T-shirt met een 'grap' over tepels.

> Het blijft echter niet bij conferenties. Tientallen instanties, van bedrijven en wetenschappers tot overheden, hebben inmiddels ethische richtlijnen ↴ opgesteld om de ontwikkeling van betrouwbare AI-toepassingen te kunnen waarborgen. Denk bijvoorbeeld aan de *'Perspectives on Issues in AI Governance'* van Google, de *'Asilomar AI Principles'* van de Future of Life Institute en de *'Ethics guidelines for trustworthy AI'* van de High-Level Expert Group on Artificial Intelligence (AI HLEG) van de Europese Commissie. Het Amerikaanse ministerie van Defensie heeft zijn eigen *'AI Principles'* met aanbevelingen om het ethisch gebruik van AI binnen defensie te waarborgen. Zelfs het Vaticaan heeft in 2020 richtlijnen gepubliceerd voor de ontwikkeling en het gebruik van AI: *'Rome Call for AI Ethics'*. Tech-reuzen als IBM en Microsoft behoorden tot de eerste

ondertekenaars. Waarden als *privacy*, *transparency* ↴ en *fairness* ↴ worden in verschillende richtlijnen uitgebreid behandeld. Naast verschillende ethische principes worden er ook steeds vaker '*ethical boards*' in het leven geroepen, die moeten toezien op het ethisch handelen van tech-bedrijven. Volgens Gartner was '*Digital Ethics & Privacy*' een van de 10 '*Strategic Technology Trends*' van 2019.

Onethische ethiek

Toch lijkt deze 'ethiek hype' alleen besteed aan de *big tech*-organisaties. Ondanks het feit dat bedrijven over de hele wereld verwachten dat ze de komende jaren AI binnen hun bedrijf gaan toepassen, blijven ze achter bij de discussies omtrent ethiek, zo blijkt uit onderzoek van Genesys (2019). Meer dan de helft van de ondervraagde werkgevers zegt dat hun bedrijf momenteel geen geschreven beleid hebben op het gebied van het ethisch gebruik van AI. Opvallend hierbij is dat slechts 1 op de 5 respondenten uitgesproken bezorgd is dat hun bedrijf AI op een onethische manier zouden kunnen gebruiken. Dit percentage loopt zelfs af wanneer de leeftijd oploopt. 21% van de millennials maakt zich zorgen dat hun bedrijf AI onethisch zouden kunnen gebruiken, vergeleken met 12% van generatie X en slechts 6% van de babyboomers. De onderzoekers vragen zich af of deze mate van onbezorgdheid wel terecht is.

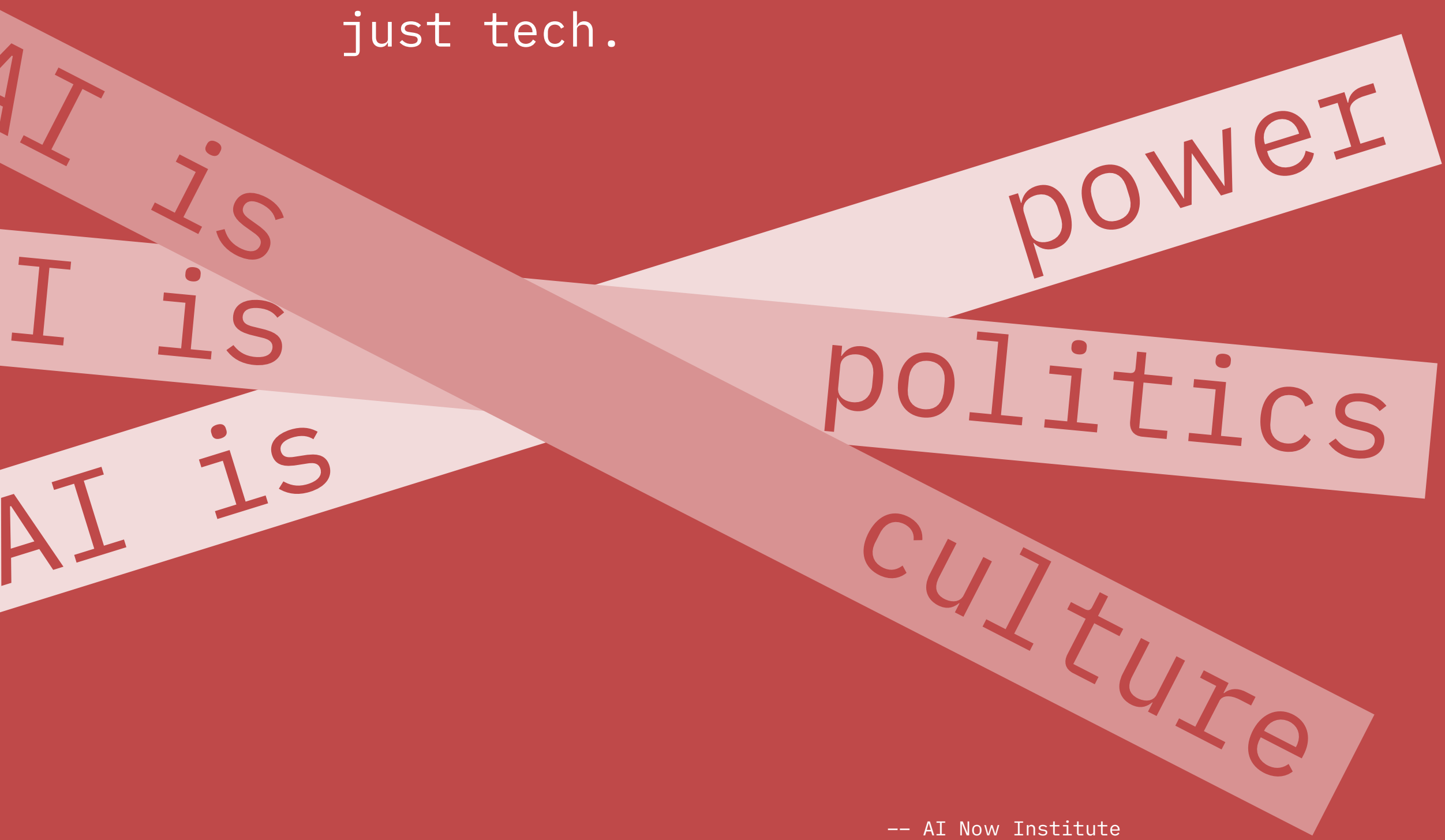
Het *AI Now Institute* bracht in 2018 een overzicht in kaart van belangrijk nieuws over de sociale implicaties van AI en de tech-industrie. Het overzicht bevestigde wat veel mensen al wel aanvoelden; het was nogal een bewogen jaar. Zomaar een greep uit 2018: een toename van het misbruik van data (met als dieptepunt de onthulling van het Cambridge Analytica schandaal, waarbij miljoenen persoonsgegevens werden ingezet om verkiezingen te beïnvloeden), een versnelling van de verspreiding van gezichtsherkenningssoftware (zoals een samenwerking van IBM en de *New York City Police Department* waarbij gezichten gezocht kunnen worden op basis van ras, met behulp van camerabeelden van duizenden politieagenten die zonder medeweten deelnamen) en een toename van schadelijke gevolgen door het testen van AI-systemen op levende populaties in domeinen met hoge risico's (zoals de dodelijke aanrijding van een voetganger door een zelfrijdende Uber). Dit is slechts een kleine selectie van een veel grotere database. Is het dan allemaal kommer en kwel? Nee, er zijn gelukkig ook positieve neveneffecten te onderscheiden. Het zet de discussie omtrent de noodzaak van regulering op scherp. Grote tech-bedrijven, zoals Microsoft en Amazon, hebben de Amerikaanse regering zelfs expliciet opgeroepen om technologieën zoals gezichtsherkenning te reguleren.

> *» Because AI isn't just tech. AI is power, and politics, and culture. «*
>
> -- AI Now Institute

Een groeiend verzet

Het *AI Now Institute* bracht ook in 2019 een overzicht in kaart van belangrijk nieuws over de sociale implicaties van AI en de tech-industrie. Dit gaf een ander beeld dan het jaar ervoor. Er bleek een toenemend verzet merkbaar te zijn tegen schadelijke AI-toepassingen. Een selectie uit de database van 2019: een versnelling van regulerende maatregelen (waaronder het verbod op gezichtsherkenning in San Francisco), een toenemende druk vanuit aandeelhouders (zoals de druk die aandeelhouders van Amazon opvoerden tegen het verkopen van gezichtsherkenningssoftware aan overheden) en een toename van het aantal protesten en stakingen van medewerkers van grote tech-bedrijven (zoals de klimaatstaking van duizenden medewerkers van onder andere Amazon, Google en Microsoft tegen de schadelijke impact van AI-toepassingen op het milieu). Het verzet in 2019 herinnert ons eraan dat er nog steeds ruimte is om te bepalen welke AI-toepassingen acceptabel zijn en op welke wijze we deze willen controleren. Maar ook hier lijkt de wenselijkheid ingehaald te worden door de werkelijkheid.

AI isn’t
just tech.



-- AI Now Institute

Door de snelle verspreiding van het coronavirus in 2020 zetten overheden datagedreven apps in om de verspreiding van het virus beter te kunnen monitoren en zo ‘*lockdowns*’ gerichter te kunnen inzetten. Mensen hebben met deze apps een *track and trace* systeem in hun broekzak; wanneer jij in de buurt bent geweest van een besmet persoon krijg je een melding. Dit schuurt natuurlijk tegen onze opvattingen over privacy en misschien nog wel belangrijker, tegen onze opvattingen over zelfbeschikking en gelijkwaardigheid. Wat als mensen onterecht verplicht in quarantaine worden gezet? In Nederland stuitte de introductie van deze apps dan ook op veel weerstand. Tegelijkertijd lijken veel mensen zich niet te beseffen dat deze apps op het gebied van privacy misschien wel vriendelijker zijn dan de apps van Google of Facebook die we al jaren zonder blikken of blozen gebruiken. Het is hoe dan ook van groot belang dat er vragen gesteld blijven worden over wie er baat hebben bij AI, wie er benadeeld worden door AI en wie daarover kunnen en mogen beslissen.

1.2 ----------

Prangende ethische vraagstukken

Ethiek staat binnen de ontwikkeling van AI ↴ volop in de spotlight. Vraagstukken omtrent ‘*explainability* ↴’, ‘*privacy* ↴’ en ‘*biases* ↴’ vliegen je om de oren. Maar wat houden deze begrippen precies in? En hoe uiten deze zich in de praktijk? Om een oordeel te kunnen vellen over deze vraagstukken moeten we eerst in kaart brengen over welke ethische ↴ vraagstukken het precies gaat en welke afwegingen hierbinnen gemaakt kunnen worden.
Bij het in kaart brengen van ethische vraagstukken worden er vaak drie centrale begrippen gehanteerd, namelijk verantwoordelijkheid, vrijheden & rechten en rechtvaardigheid (van Dalen, 2013).

Verantwoordelijkheid

Wanneer het over de ontwikkeling van AI gaat wordt het begrip ‘verantwoordelijkheid’ vaak genoemd. De meeste mensen hebben bijvoorbeeld de vraag wie er verantwoordelijk is bij een ongeluk met een zelfrijdende auto wel eens voorbij horen komen. Is het de inzittende, de ontwikkelaar óf het systeem zelf? Machines als onafhankelijke actor kenden we eerder alleen uit sciencefictionfilms. Dit levert allerlei nieuwe vraagstukken op, zowel juridisch als sociaal.
De tijd van filosoferen vanaf de zijlijn is voorbij. Dergelijke vraagstukken zijn inmiddels de realiteit. Maar om te bepalen wie er verantwoordelijk is moeten we eerst bepalen waar zij dan precies verantwoordelijk voor zijn en welk gedrag wel en niet te verantwoorden is. Daarbij is het de vraag op welke wijze we de mate van verantwoordelijkheid kunnen herleiden.

Waarvoor zijn we verantwoordelijk?

Een bekend gedachte-experiment binnen de ethiek is het zogenaamde ‘trolleyprobleem’. De hamvraag hierbij is of het te verantwoorden is wanneer het leven van een persoon wordt opgeofferd om de levens van meerdere personen te redden. Om deze gedachte-exercitie voorstelbaar te maken wordt in het experiment gebruik gemaakt van een tram, oftewel *trolley*.

‘Een op hol geslagen tram rijdt op een groep van vijf spoorwerkers af. Je kunt nog ingrijpen door de spoorwissel om te zetten, zodat de tram op een ander spoor terecht komt. Op dit spoor staat echter nog een andere spoorwerker. Red je het leven van vijf personen door de wissel om te zetten, of spaar je het leven van één persoon door niets te doen?’

Dit roept uiteraard allerlei vragen op. Zijn vijf levens meer waard dan één leven? Ben je verantwoordelijk als je ingrijpt? En hoe zit dat wanneer je niet ingrijpt? Interessant hierbij is om te bekijken of de afweging die gemaakt wordt afhankelijk is van de identiteit van de vijf personen die eventueel gered worden en de ene persoon die daarvoor mogelijk opgeofferd wordt. Maken mensen dezelfde keuze wanneer die ene persoon hun liefdespartner is en de vijf andere personen onbekenden? Of wanneer die ene persoon een jonge dokter is en de vijf andere personen bejaarden? De vraag gaat dan niet langer over kwantiteit, maar over kwaliteit. En juist die vraag is in het kader van AI veel moeilijker om te automatiseren.

Met de komst van de zelfrijdende auto kunnen we eigenlijk nauwelijks nog van een gedachte-experiment spreken. Vergelijkbare verkeerssituaties kunnen zich nu daadwerkelijk voordoen. De semi-autonome auto’s die momenteel op de weg zijn toegestaan kunnen immers zelfstandig remmen en van baan wisselen. De vraag is dan wat een auto moet doen wanneer er bijvoorbeeld een groep mensen bij een zebrapad oversteekt en de auto niet meer op tijd kan remmen. Doorrijden om de inzittenden te sparen of uitwijken om de overstekende mensen te sparen? Door de uitwijkmanoeuvre komt de inzettende echter wel om het leven. De vraag hierbij is natuurlijk of het dan nog uitmaakt wie er oversteken en wie er in de auto zitten. MIT ontwikkelde hiervoor de *Moral Machine* om te kijken wat mensen in zo’n situatie zouden doen. Het experiment startte in 2016 en is wereldwijd inmiddels door meer dan 40 miljoen mensen ingevuld.

De resultaten zijn in 2018 gepubliceerd in *Nature*. Uit een analyse van de resultaten blijkt dat er op sommige vlakken algemeen geldende voorkeuren bestaan. Denk hierbij aan het sparen van jongeren boven ouderen en het sparen van mensen boven dieren. Andere oordelen zijn juist cultureel bepaald. Deelnemers uit Midden- en Zuid-Amerika blijken vrouwen te sparen boven mannen en mensen met een atletisch lichaam boven mensen met overgewicht. Bij deelnemers uit landen met een hoge inkomensongelijkheid bleken mensen met een hoge sociale status gespaard te blijven ten opzichte van mensen met een lagere sociale status.

Wat we als ‘verantwoordelijk’ beschouwen lijkt daarmee niet volledig te objectiveren. Toch kun je je afvragen of dit de ontwikkeling van de zelfrijdende auto moet tegenhouden. Hoe vaak komen dit soort keuzesituaties in de praktijk daadwerkelijk voor? En hoe verhoudt zich dit tot het aantal ongelukken dat voorkomen kan worden met autonome voertuigen? Juist wanneer alles met elkaar verbonden is (dus niet alleen de auto’s onderling, maar álle weggebruikers en dus ook voetgangers en fietsers) kan een autonoom systeem nog veel sneller dan de mens vooruitkijken en anticiperen op mogelijk gevaarlijke situaties. Het trolleyprobleem werkt goed om te filosoferen en om de complexiteit van ethische dilemma’s inzichtelijk te maken, maar het is en blijft een gedachte-experiment. Ontwikkelaars houden zich in de praktijk eerder bezig met de vraag op welke wijze de zelfrijdende auto zo veilig mogelijk gemaakt kan worden. Je kunt dan ook kijken naar de weg die het voertuig heeft afgelegd vóórdat het dilemma zich voordoet. Bij het trolleyprobleem gaat men te veel uit van de bestaande context: waarom zouden we bij de zelfrijdende auto een omgeving creëren waarbij voertuigen en voetgangers elkaar überhaupt kunnen kruisen? Elon Musk werkt bijvoorbeeld met ‘The Boring Company’ aan een ondergrondse vorm van mobiliteit. Mensen kunnen dan in een ondergrondse tunnel door autonome voertuigen verplaatst worden. Daar bestaat het trolleyprobleem in principe helemaal niet.

» *De discussie over het trolley-probleem zou niet over de gedwongen keuze moeten gaan, maar over het optimaliseren van veiligheid.* «

-- Arjen Goedegebure, OGD ict-diensten

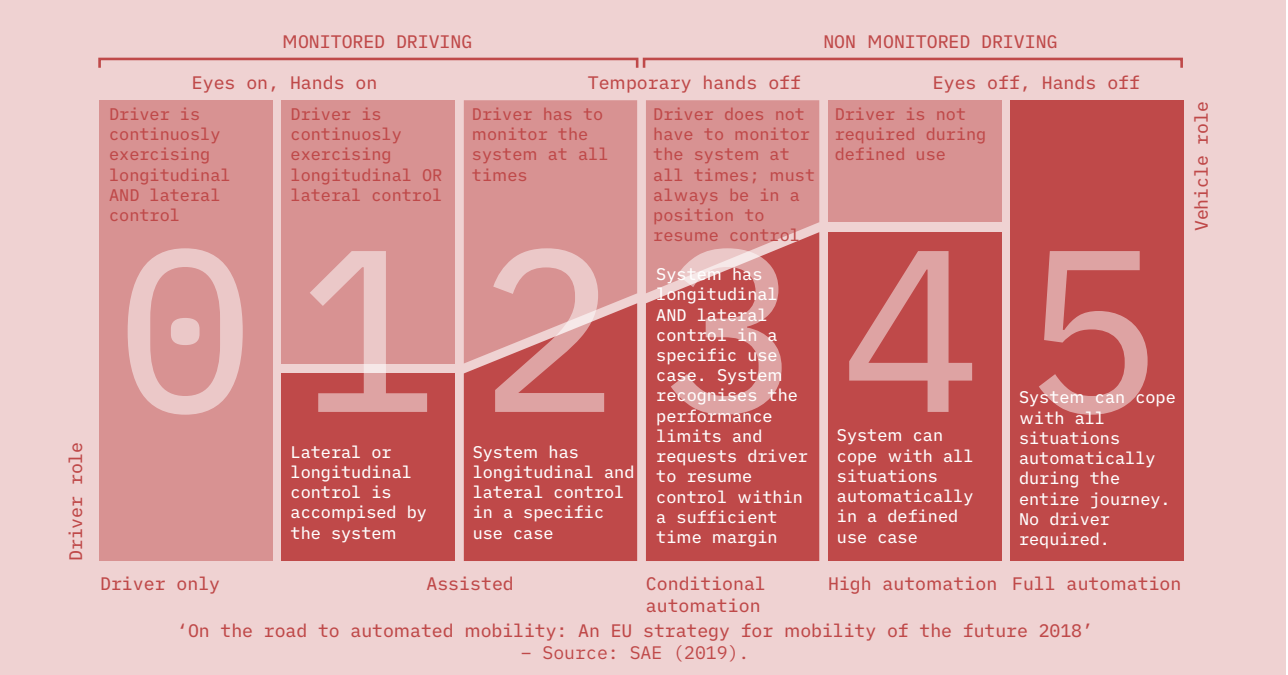
Wie is er verantwoordelijk?

Zelfs al doet het trolleyprobleem zich niet voor, dan nog kunnen er ongelukken ontstaan als gevolg van systeemfouten van autonome voertuigen en blijft het de vraag wie er volgens de wet verantwoordelijk is bij bijvoorbeeld een botsing. Verantwoordelijkheden die door de wet worden voorgeschreven noemen we aansprakelijkheid. Het aansprakelijkheidsvraagstuk is overigens niet volledig nieuw voor AI. Volgens de Nederlandse wet bestaat er namelijk het principe van 'productaansprakelijkheid'. In artikel 6:185 van het Burgerlijk Wetboek wordt productaansprakelijkheid als volgt omschreven: 'De producent is aansprakelijk voor de schade veroorzaakt door een gebrek in zijn product'.

> Wanneer een bestuurder van een reguliere auto kan aantonen dat deze letsel heeft opgelopen en dat het letsel is veroorzaakt door het gebrekkige product dan is de producent van de auto dus aansprakelijk.

Volgens informatie van de Nederlandse Rijksoverheid gelden de regels voor aansprakelijkheid die gelden voor normale auto's ook bij testen met zelfrijdende auto's. 'De bestuurder is verantwoordelijk, als hij zelf het voertuig bestuurt. Maar als het systeem niet goed werkt, is de producent verantwoordelijk.' Deze wetgeving is volgens de overheid voor de testfase voldoende.
Bij de semi-autonome auto's die momenteel op de markt zijn wordt van de bestuurder immers verwacht dat hij nog steeds blijft opletten en direct kan ingrijpen als er iets misgaat. Verschillende autoproducenten hebben echter aangekondigd binnen enkele jaren volledig zelfrijdende auto's op de markt te gaan brengen. In dat geval zal productaansprakelijkheid in de toekomst een veel grotere rol gaan spelen. Autonome auto's nemen dan immers de taken van de bestuurder volledig over en daarmee ligt de verantwoordelijkheid voor het waarborgen van de veiligheid ook in steeds hogere mate bij het systeem. Ondanks de redenatie dat autonome voertuigen de verkeersveiligheid drastisch kunnen vergroten, zullen de systemen hun eigen uitdagingen hebben. Denk hierbij aan situaties waarbij: de hardware faalt, er een bug in de software zit, het systeem wordt gehacked, de interactie tussen mens en machine hapert of wanneer het systeem niet goed anticipeert op andere deelnemers aan het verkeer of onverwachte verkeerssituaties (Van Wees, 2018).

> Momenteel lijken de richtlijnen omtrent product-aansprakelijkheid echter nog onvoldoende geschikt te zijn voor de introductie van autonome voertuigen. Zo is het momenteel niet duidelijk of software ook onder de richtlijnen valt. Het principe van product-aansprakelijkheid richt zich namelijk op roerende zaken (zoals een zelfrijdende auto), maar het is nog de vraag of software hier juridisch gezien ook onder geschaard wordt. Wanneer autonome voertuigen vanaf niveau 3 worden ingevoerd treedt daarbij een paradoxale situatie op. Volgens de richtlijnen van de *Society of Automotive Engineers* (SAE) kunnen voertuigen op niveau 3 onder specifieke omstandigheden autonoom rijden en is de bestuurder niet nodig voor het in de gaten houden van de omstandigheden. De bestuurder dient echter wel op ieder moment in staat te zijn om in te grijpen, wanneer het voertuig aangeeft dat dit nodig is. Maar een bestuurder die niet op hoeft te letten, lijkt ook niet in staat te zijn om in te grijpen.



‘On the road to automated mobility: An EU strategy for mobility of the future 2018’ – Source: SAE (2019).

Daarbij komt dat wanneer mensen niet langer zelf hoeven te rijden, de rijvaardigheid in de loop der tijd zal afnemen. De vraag is of mensen dan nog in staat zijn om in te grijpen wanneer het systeem daarom vraagt. In de meeste gevallen zal dit om complexe situaties gaan. Een zelfrijdende auto vraagt in dat opzicht dus juist om een meer bekwame, in plaats van een minder bekwame chauffeur. Dit principe geldt ook in andere situaties. Wanneer bijvoorbeeld een menselijke arts moet ingrijpen bij een robotchirurg zal deze ook zijn chirurgische skills op peil moeten hebben en kennis moeten hebben van het complexe systeem. Dit maakt het lastig om het principe van *human in the loop* (HITL) te kunnen waarborgen.

» The curse of automation: the need of higher skilled operators. «

-- Edgar Reehuis, Hudl

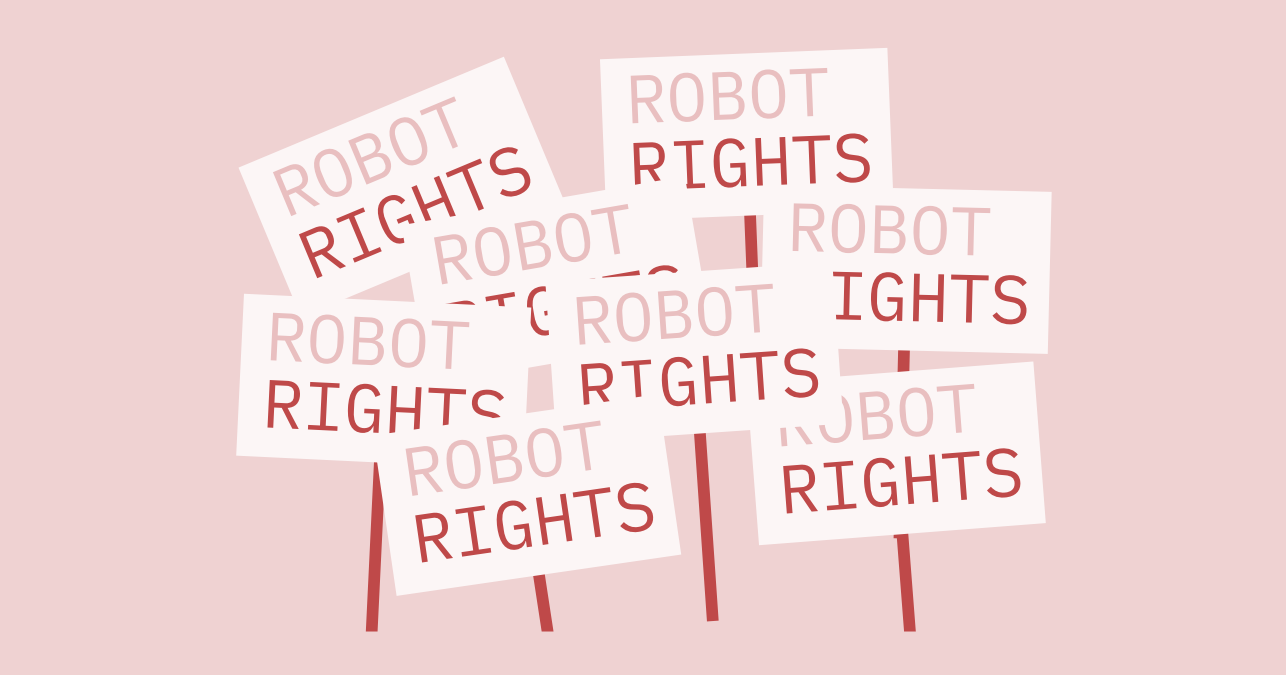
Toch dendert de ontwikkeling van autonome systemen voort. Techgiganten zoals Google en Alibaba zijn naar eigen zeggen al druk bezig met de ontwikkeling van voertuigen op niveau 4 en zelfs al op niveau 5, waarbij geen menselijke bestuurder meer aanwezig hoeft te zijn. Waymo (het voormalige *self-driving project* van Google)

experimenteert al vanaf 2017 met volledig autonome auto’s in enkele buitenwijken van Phoenix. Vooralsnog zat er altijd een mens achter het stuur die kon ingrijpen als het mis ging, maar in 2019 kondigde het bedrijf aan dat ze betaalde taxiritten gaan aanbieden zónder menselijke bestuurder. Uiteraard is het verkeer in de buitenwijken van Phoenix wat voorspelbaarder dan in het centrum van Amsterdam, maar het is alsnog een doorbraak.

» Binnenkort op de weg: zelfrijdende robottaxi’s zonder reservemens achter het stuur «

-- Bard van de Weijer, Volkskrant

Aangezien het om zelflerende systemen gaat (waarbij keuzes niet volledig geprogrammeerd zijn, maar gebaseerd worden op data en ervaringen van het systeem zelf) kan de producent niet volledig aansprakelijk gesteld worden. En zonder menselijke bestuurder kan de inzittende ook niet alle aansprakelijkheid dragen. In dat geval zou het systeem zelf tot de verantwoording geroepen moeten kunnen worden. Dit klinkt als toekomstmuziek, maar dat is het niet. In 2016 werd het zelfrijdende autosysteem van Google in de VS officieel erkend als ‘*driver*’ door de *National Highway Traffic Safety Administration* (NHTSA). Hierdoor wordt het AI-systeem dus geclassificeerd als de bestuurder van de auto. In hetzelfde jaar introduceerde de Europese Commissie voor het eerst de term ‘*e-personhood*’. De term werd in een rapport gebruikt om de juridische status van de meest geavanceerde autonome robots te beschrijven. Volgens het rapport zouden deze systemen specifieke rechten en verplichtingen kunnen krijgen om zo aangesproken te kunnen worden op gemaakte schade (Delvaux, 2016). Inderdaad, naast verplichtingen zouden de systemen dan ook rechten verkrijgen. Zouden autonome systemen in de toekomst dan ook mensen aansprakelijk kunnen stellen? Lees hier meer over in het scenario ‘Robot Rights’.



Hoe kunnen we verantwoordelijkheid herleiden?

Naarmate AI steeds autonomere beslissingen gaat nemen (en daardoor steeds vaker aansprakelijk gesteld zal moeten gaan worden) is het van belang dat het inzichtelijk is op welke wijze de beslissingen van dergelijke systemen tot stand zijn gekomen. Algoritmen ↴ maken hierbij steeds vaker zélf de clusters ↴, zonder voorgeprogrammeerde labels. Het wordt hierdoor steeds lastiger voor mensen om te achterhalen waarop de uitkomsten gebaseerd zijn. AI wordt daarom nu nog vaak vergeleken met een *black box*. Een belangrijke vraag hierbij is of het besluitvormingsproces van de toekomst voldoende transparant is en of de uitkomst voldoende uitlegbaar is. Er wordt daarom veel aandacht besteed aan Explainable AI ↴ (XAI).

Er moet hierbij onderscheid gemaakt worden tussen transparantie, de uitleg en uitlegbaarheid. Transparantie gaat voornamelijk over het proces en de opgestelde criteria vooraf, terwijl de uitleg en uitlegbaarheid over de toelichting op en herleidbaarheid van het besluit achteraf gaan. Uitlegbaarheid is daarbij erg subjectief en contextafhankelijk, terwijl transparantie en de uitleg veel meer objectief zijn. Het in de praktijk brengen van uitlegbaarheid is daardoor enorm lastig. Je moet van een blackbox naar een '*glassbox*'. In de praktijk betekent dit dat je dan achteraf een redenatie opzet. Systemen die momenteel ontwikkeld worden met behulp van bijvoorbeeld *deep learning* ↴ zijn echter dermate complex, dat deze voor mensen nauwelijks nog te begrijpen zijn. Het is niet zo dat je een systeem om uitleg vraagt en er dan netjes een A4'tje uitrolt met een chronologisch overzicht van de belangrijkste overwegingen en beslissingen. Het is een *multi-layered* web van verbindingen. Net als het menselijk brein overigens. Ook bij menselijke besluitvorming kunnen we niet onder de motorkap ↴ kijken. Wanneer iemand in het verkeer een ongeluk veroorzaakt moeten we ons ook beroepen op een geregisseerde verklaring achteraf. We kunnen ook dan niet in de hersenpan kijken om zo de exacte aanleiding van een keuze te achterhalen.

» Het agenderen van transparantie en uitlegbaarheid is eenvoudig, maar het realiseren ervan is enorm complex. «

-- Marijn Janssen, TU Delft

Transparantie is daarbij überhaupt een misleidend begrip; het is namelijk altijd een beperkte weergave van de realiteit. Denk bijvoorbeeld aan een raam. Door het glas kun je naar buiten kijken en lijkt het transparant. Maar je kunt slechts zoveel zien als het venster van het raam je laat zien. Dit levert een *governance* probleem op. Alles moet gecontroleerd en *geaudit* kunnen worden; niet alleen de output, maar ook de beslisregels en de input. Het is echter bijna onmogelijk te herleiden wat de precieze oorsprong van de data is. Data wordt namelijk al opgeschoond en gereduceerd voordat het gebruikt wordt. Deze zogenaamde *data cleansing* maakt het lastig om het proces te kunnen reproduceren zelfs voordat het gestart is.

In de huidige discussies omtrent de toepasbaarheid van AI bij besluitvormingsprocessen wordt er onvoldoende onderscheid gemaakt tussen de verschillende typen besluiten en de impact die de besluiten hebben op de betrokkenen. Het maakt nogal een verschil of het gaat om een aanbeveling voor een film, een medische diagnose op basis van longfoto's of een uitwijkmanoeuvre van een autonoom voertuig. De mate van impact is dus onder andere afhankelijk van de mogelijke risico's. Er zijn daarom '*levels of explainability*' nodig. Bij chatbots is uitlegbaarheid bijvoorbeeld minder noodzakelijk dan bij zelfrijdende auto's en oorlogsdrones. De noodzaak van explainability en de gevolgen van foutieve beslissingen zijn dus afhankelijk van de context en het type besluit.

Dit hangt nauw samen met de mate van autonomie die we aan het systeem zullen toekennen en daarmee de relatie die we met de technologie hebben. Is het 'slechts' een tool of maakt het volledig autonome beslissingen? AI heeft in dat opzicht nog een lange weg te gaan. Uit onderzoek van de UvA naar geautomatiseerde besluitvorming door AI uit 2018 blijkt dat veel Nederlanders bezorgd zijn dat AI leidt tot manipulatie, risico's of onaanvaardbare resultaten. Alleen bij meer objectieve besluitvormingsprocessen, zoals de aanvraag van een hypotheek, worden er kansen gezien voor AI. Menselijke controle, menselijke waardigheid, eerlijkheid en nauwkeurigheid ↴ worden als belangrijke waarden genoemd bij het nadenken over besluitvorming door AI.

» Er komt misschien wel een dag dat wij tegen AI zeggen 'verklaar jezelf' maar dat de AI zegt 'you wouldn't understand it anyway'. «

-- Maarten Stol, BrainCreators

Vrijheden en Rechten

Vrijheid gaat in de basis over de vrijheid die mensen hebben om zelf te kunnen bepalen hoe zij hun leven inrichten. Dit is zelfs een recht, het zelfbeschikkingsrecht. Dit recht wordt vervolgens ingeperkt door het verbod om bij anderen schade aan te richten. Je bent dus bijvoorbeeld niet vrij om bij andere mensen thuis in te breken. Bij de formulering van rechten staat de vrijheid van het individu centraal. Bij sommige rechten gaat het erom dat anderen zich niet met je bemoeien of zelfs dingen moeten nalaten om je vrijheden te beschermen (vrijheidsrechten), terwijl het er bij andere rechten juist om gaat dat anderen – meestal overheden – zich moeten inspannen voor je om je vrijheden te kunnen realiseren (sociale rechten). Vrijheidsrechten en sociale rechten worden internationaal onderkend en zijn opgenomen in de Universele Verklaring van de Rechten van de Mens. Met de komst van AI is het de vraag in hoeverre wij nog volledig vrij zijn en in hoeverre bepaalde rechten worden aangetast. Zo worden we steeds meer in de gaten gehouden door bijvoorbeeld gezichtsherkenningssoftware.

Hoeveel vrijheid hebben we om zelf te kiezen?

In de discussie omtrent AI gaat het vaak over het schrikbeeld ↴ dat intelligente robots de controle kunnen overnemen en we elke vorm van zelfbeschikking kwijtraken. Hierbij wordt er vanuit gegaan dat een afzonderlijk systeem de capaciteit zal hebben om menselijke intelligentie ↴ te overstijgen. Wat hierbij echter over het hoofd wordt gezien is dat als hele krachtige specialistische systemen aan elkaar worden gekoppeld, er ook een intelligent systeem ontstaat. In plaats van '*General AI ↴ gone bad*', is het dan '*Narrow AI ↴ everywhere*'.

Steden hangen in steeds hogere mate vol met sensoren. Dit stelt intelligente systemen in staat om steeds meer geautomatiseerde keuzes te maken, ook zónder tussenkomst van de mens. Elke zelf respecterende stad noemt zichzelf tegenwoordig een *Smart City*. Oftewel een stad die verschillende vormen van data inzet om processen te optimaliseren en problemen op het gebied van bijvoorbeeld vervoer, veiligheid en

milieu aan te pakken. Dit gaat veel verder dan slimme stoplichten. Denk bijvoorbeeld aan camera's met gezichtsherkenningssoftware waarmee voetbalhooligans uit stadions kunnen worden geweerd of het monitoren van sociale media om de toeristenstroom in de stad inzichtelijk te maken en te kunnen sturen.

» In de stad van de toekomst praten lantaarnpalen mee en burgers niet. «

-- Maurits Martijn & Sanne Blauw, De Correspondent

De democratie kan dan veranderen in een zogenaamde 'algocratie', waarbij steden (en dus mensen) worden bestuurd door data. Steeds meer experts waarschuwen hierbij voor een *black box society*, waarin keuzes van slimme algoritmen niet meer te herleiden zijn. Burgers kunnen dan in steeds mindere mate zelf beslissen of ze onderdeel uitmaken van deze datagedreven samenleving. Met behulp van de technologie kan zelfs een totalitaire staat gecreëerd worden; *Big Brother is watching you*. Zo rolt China stap voor stap een 'sociaal kredietsysteem' uit. Chinese burgers krijgen op basis van hun gedrag een bepaalde score. Op basis van deze score kunnen mensen op een zwarte lijst worden geplaatst en allerlei (voor)rechten verliezen, zoals de mogelijkheid geld te lenen of naar het buitenland te reizen. In 2018 werden al 23 miljoen Chinezen verboden een trein- of vliegticket te kopen. Ook worden wereldwijd op steeds meer plekken restricties gelegd op het internetgebruik. Zo keurden autoriteiten in Pakistan in 2020 ingrijpende nieuwe regels goed die het gebruik van sociale media beperken, waarmee de vrijheid van meningsuiting in gevaar wordt gebracht.

De vrijheid om zelf keuzes te maken kan uiteraard ook vanuit het omgekeerde standpunt worden benaderd. AI wordt (of is) op verschillende vlakken beter dan de mens. Met name wanneer het gaat om zeer specialistische toepassingen. Dit gaat verder dan extreem krachtige schaakcomputers. Zo blijken algoritmes nu al beter in

staat om vormen van kanker te herkennen op longfoto's dan artsen. Je kunt dan de vraag stellen of het nog wel verantwoord is om bepaalde taken door mensen uit te laten voeren. In dat opzicht zouden mensen zelf de keuze moeten krijgen om een artificieel systeem te verkiezen boven de mens.

Hebben we het recht vergeten te worden?

AI-systemen dringen alsmaar verder door in ons leven en schuren tegen de Rechten voor de Mens. Denk bijvoorbeeld aan het Systeem Risico Indicatie (SyRI) dat de Nederlandse overheid gebruikte om fraude te bestrijden op het gebied van toeslagen, belastingen en uitkeringen. Op basis van gegevens over onder andere werk, inkomen, pensioenen en schulden berekende het systeem wie er mogelijk zou kunnen frauderen. Met name in kwetsbare wijken. Een groot bezwaar bij een dergelijk systeem is dat de gegevens van álle bewoners in een wijk geanalyseerd kunnen worden, ook als ze onschuldig zijn. Dit maakte mensen Bij Voorbaat Verdacht. Verschillende burgerrechten- en privacyorganisaties vonden dan ook dat het systeem niet door de beugel kon en klaagde de Nederlandse Staat aan. En met succes. In 2020 oordeelde de rechtbank dat de wetgeving die de inzet van SyRI regelt in strijd is met artikel 8 van het Europees Verdrag voor de Rechten voor de Mens (EVRM), namelijk het recht op respect voor het privéleven. Deze bepaling vereist een eerlijke balans tussen het maatschappelijk belang dat de wetgeving dient en de inbreuk op het privéleven die de wetgeving maakt. De rechtbank oordeelde dat het voorkomen en bestrijden van fraude onvoldoende opwoog tegen de inbreuk op het privéleven.

Het Rathenau Instituut pleitte in 2017 voor een nieuw Europees verdrag waarmee mensenrechten een update zouden krijgen en daarmee aangepast zouden worden aan de digitale samenleving. In het rapport worden zelfs nieuwe mensenrechten geopperd, waaronder het recht om niet gemeten, geanalyseerd of beïnvloed te worden (het kunnen weigeren van online profilering, tracking en

beïnvloeding). En niet zonder reden. Toepassingen op het gebied van bijvoorbeeld gezichtsherkenning ↴ zetten ons recht op privacy steeds verder onder druk. Denk bijvoorbeeld aan de ophef rondom het bedrijf *Clearview*, dat voor het trainen van hun gezichtsherkenningssoftware foto's van Facebook en miljoenen andere websites 'schraapt' en hun diensten aanbiedt aan verschillende opsporingsdiensten. Of aan zangeres Taylor Swift die tijdens haar concerten gezichtsherkenningstechnologie inzette om mogelijke stalkers in het publiek te kunnen identificeren. Camera's hoeven niet eens meer in de buurt te zijn. Er wordt nu gewerkt aan gezichtsherkenningssystemen voor militaire doeleinden waarmee gezichten tot op een kilometer afstand herkend kunnen worden.

Een variant op het recht om niet gemeten te worden is in 2018 realiteit geworden in de Europese *General Data Protection Regulation* (GDPR) of Algemene Verordening Gegevensbescherming (AVG). In deze regelgeving is onder andere het '*Right to be Forgotten*' opgenomen, oftewel het 'Recht op Vergetelheid'. Het doel van deze verordening is om gebruikers de controle te geven over hun persoonlijke gegevens en zo te kunnen inzien wat bedrijven precies met deze gegevens doen. Dit betekent dat alle organisaties expliciet om toestemming moeten vragen om persoonsgegevens te mogen verzamelen en gebruiken. Het betekent ook dat gebruikers het recht hebben om inzicht te krijgen in welke persoonsgegevens een organisatie verwerkt en op welke wijze deze gegevens worden beveiligd. Gebruikers hebben daarbij het recht om organisaties te vragen om alle persoons-gegevens te verwijderen. Organisaties die hier niet aan voldoen riskeren hoge boetes. De Zweedse Autoriteit voor Gegevensbescherming heeft in 2019 de eerste GDPR-boete van het land uitgedeeld aan een scholengemeenschap voor het gebruik van gezichtsherkenningstechnologie om de aanwezigheid van studenten op school te controleren. Het ging om een boete van 200.000 Zweedse Kronen, wat neerkomt op bijna 19.000 euro.

Er zijn echter uitzonderingen. Bijvoorbeeld als een organisatie wettelijk verplicht is om gegevens te gebruiken. Denk hierbij aan gegevens die noodzakelijk zijn voor een rechtsvordering of het waarborgen van de volksgezondheid. Dit zet de discussies omtrent apps die worden ingezet om epidemieën te monitoren, zoals tijdens de coronacrisis in 2020, verder op scherp. Het recht om vergeten te worden reikt daarbij niet verder dan de grenzen van de Europese Unie, zo oordeelde het Europese Hof van Justitie in 2019. Het evenwicht tussen het recht op privacy enerzijds en de vrijheid van informatie van internetgebruikers anderzijds, zal volgens een verklaring van de rechter aanzienlijk variëren over de hele wereld. Dit betekent dat zoekmachinegiganten niet verplicht zijn om informatie buiten de EU-landen te verwijderen. Privacy zou daarmee in de toekomst een luxeproduct kunnen worden dat slechts voor een kleine groep mensen toegankelijk is. Lees hier meer over in het scenario 'Privacy for Sale'.

Hoe kunnen we het recht in eigen handen nemen?

Steeds vaker komen 'spionagepraktijken' door overheden aan het licht en we komen daardoor steeds meer in een 'surveillancesamenleving' terecht. Voormalig CIA-medewerker Edward Snowden luidde in 2013 de klokken

over massale afluisterpraktijken van de Amerikaanse *National Security Agency* (NSA). *'Big brother is watching you'* werd realiteit. Geïnspireerd op deze ontwikkelingen en de quote van George Orwell uit zijn boek 1984, lanceerde een Australisch kledingmerk in 2014 een speciaal ontwikkelde kledinglijn waarmee je telefoon van de radar verdwijnt. De belangrijkste eigenschap van de 1984-kledinglijn is de zogenaamde *'UnPocket'*. Een canvas zakje, gewoven met metalen stoffen, waarmee onder andere WiFi- en GPS-signalen geblokkeerd worden. Of je hiermee veilig bent voor de NSA is nog maar de vraag, maar het beveiligt mensen wel eenvoudig tegen locatietracking.

» I don't know why people are so keen to put the details of their private life in public; they forget that invisibility is a superpower. «

-- Banksy

In plaats van te wachten op herziene wetgeving (die niet altijd afdoende is) ontwikkelen steeds meer creatieve ondernemers apps en gadgets waarmee we zelf onze eigen privacy kunnen waarborgen. Ook uit andere voorbeelden blijkt dat de fashion-industrie erg actief is op dit vlak. Denk bijvoorbeeld aan grafische prints op kleding waardoor surveillancetechnologie in de war wordt gebracht of gestileerde gezichtsmaskers waarmee je gezicht niet herkend wordt door gezichtsherkennings-software. Ook om te voorkomen dat we te pas en te onpas op de foto worden gezet (en vaak onbedoeld, maar wel ongevraagd op social media belanden) is er een kledingstuk. De van oorsprong Nederlandse ontwerper, Saif Siddiqui, ontwikkelde *The ISHU scarf*. Een speciale sjaal waarmee de flits van smartphones direct wordt teruggekaatst door hele kleine kristallen balletjes die in de sjaal verwerkt zijn. Hierdoor kan de foto dus niet gebruikt worden.

Rechtvaardigheid

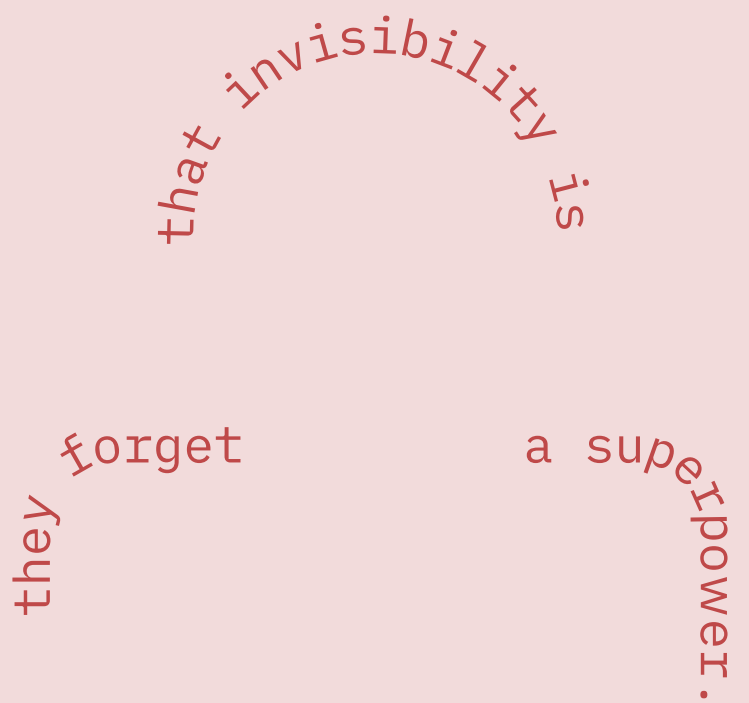
Wanneer het over rechtvaardigheid gaat, dan gaat het in de basis over de gelijkwaardigheid van mensen. Mensen dienen gelijk behandeld te worden én gelijke kansen te krijgen. Dit betekent niet dat er geen verschillen tussen mensen kunnen of mogen zijn. Maar wanneer mensen verschillend behandeld worden moet hier een duidelijk aanwijsbare reden voor zijn om het verschil te kunnen rechtvaardigen. Denk bijvoorbeeld aan verschillende salarisschalen die gebaseerd zijn op ervaring en opleiding. De vraag is of een gelijkwaardige verdeling en behandeling met de opkomst van AI gewaarborgd kan blijven.

Wat is een rechtvaardige verdeling?

AI wordt steeds vaker ingezet bij kredietbeoordelingen, fraudedetecties en sollicitatieprocessen. Het uitgangspunt is dat mensen hierbij gelijke kansen en mogelijkheden krijgen. In de praktijk blijkt dit echter vaak niet op te gaan. Zo heeft Amazon in 2014 een AI-applicatie ontwikkeld om sollicitanten te evalueren om zo tot een selectie van de beste kandidaten te komen. Pas een jaar later kwamen ze erachter dat deze seksistisch was. Het probleem was dat het algoritme werd getraind met data van sollicitanten die de afgelopen 10 jaar bij Amazon gesolliciteerd hadden. Aangezien dit in de techbranche voornamelijk mannen zijn, kreeg het algoritme een voorkeur voor mannelijke kandidaten. In 2017 is de ontwikkeling van de applicatie stopgezet. Om de ongelijkwaardige verdeling van algoritmen zelf te kunnen ervaren ontwikkelden vier alumni van *NYU Abu Dhabi* het spel 'Survival of the Best Fit.' Het educatieve spel legt de vooroordelen van AI ↴ in het sollicitatieproces bloot.

I DON'T KNOW WHY PEOPLE ARE SO KEEN TO PUT THE DETAILS OF THEIR PRIVATE LIFE IN PUBLIC

-- Banksy



Bij dergelijke toepassingen worden vooroordelen – en daarmee digitale discriminatie – dus juist versterkt. Het probleem is dat data niet divers genoeg is en dat de technologieën daardoor niet neutraal zijn. Uit onderzoek van het *Georgia Institute of Technology* blijkt dat zelfs de beste beeldherkenningssystemen minder nauwkeurig zijn in het detecteren van voetgangers met een donkere huidskleur dan voetgangers met een lichte huidskleur. De onderzoekers gaven aan dat deze bias voornamelijk werd veroorzaakt door het feit dat er te weinig voorbeelden van voetgangers met een donkere huidskleur in de *trainingssets* worden gebruikt. Wanneer de input niet volledig is, dan is de output dat ook niet. 'Garbage In, Garbage Out'.

De vraag hierbij is of er überhaupt voldoende data beschikbaar is van de meer kwetsbare groepen. Het lijkt erop dat vooroordelen tegen gehandicapten nog moeilijker te bevechten zijn dan discriminatie op basis van geslacht of ras. Bijvoorbeeld bij zelfrijdende auto's. De algoritmen worden onder andere getraind om voetgangers te kunnen herkennen, zodat de voertuigen er niet overheen rijden. Wanneer de trainingsgegevens bijvoorbeeld geen mensen in een rolstoel bevatten, kan de technologie die mensen in levensbedreigende situaties brengen. Handicaps komen weliswaar relatief vaak voor, maar er zijn enorm veel verschillende soorten handicaps. En deze zijn niet altijd zichtbaar. Een auto kan bijvoorbeeld wel claxonneren om naderende voetgangers te waarschuwen, maar een doof persoon hoort dit niet. Informatie over handicaps is daarnaast ook erg gevoelig. Mensen zijn veel minder geneigd om deze informatie te delen dan informatie over geslacht, leeftijd of ras. In sommige situaties is het zelfs illegaal om deze informatie te vragen.

Virginia Eubanks schreef in 2018 een veelgeprezen boek over het automatiseren van ongelijkheid; '*Automating Inequality*'. Hierin beschrijft ze op welke wijze geautomatiseerde systemen – in plaats van mensen – bepalen welke buurten worden gecontroleerd, welke gezinnen de benodigde middelen krijgen en wie er wordt onderzocht op fraude. Juist mensen met minder middelen en kansen worden door deze systemen benadeeld. Dit gaat niet zozeer over privacy, maar over de verregaande gevolgen van keuzes die gebaseerd zijn op onbetrouwbare data.

» Databescherming is niet een privaat, maar een algemeen belang en ligt in het hart van de rechtsstaat. «

-- Maxim Februari, jurist en schrijver

In hoeverre is er sprake van een rechtvaardige behandeling?

Dat mensen gelijk behandeld dienen te worden en gelijke kansen behoren te hebben is vastgelegd in de Nederlandse Grondwet. Paradoxaal genoeg wordt het steeds meer duidelijk dat AI-toepassingen die gebruikt worden binnen het rechtssysteem een ongelijkwaardige behandeling juist bevorderen. Denk bijvoorbeeld aan ontwikkelingen binnen *predictive policing*, waarbij crimineel gedrag wordt voorspeld door middel van grootschalige monitoring en data analyses. Het risico is hierbij dat mogelijk de verkeerde mensen worden opgepakt. Wat is in dit geval rechtvaardig? Zet je mensen mogelijk onterecht vast, of riskeer je dat ze een misdaad begaan? Gebruikte technologieën, zoals gezichtsherkenningssoftware, zijn daarbij verre van foutloos. In een test die de *American Civil Liberties Union* (ACLU) in 2018 heeft uitgevoerd bleek dat de software van Amazon 28 leden van het Congres ten onrechte identificeerde als mensen die zijn gearresteerd voor een misdaad.

In Amerika is veelvuldig gebruik gemaakt van software om te voorspellen hoe groot de kans is dat een veroordeelde opnieuw de fout in gaat. Uit onderzoek van *ProPublica* in 2016 blijkt dat deze software bevooroordeeld is in het nadeel van mensen met een donkere huidskleur. Het is daarbij bijzonder lastig om in te schatten op welke wijze autonome systemen zich gaan gedragen wanneer deze aan elkaar gekoppeld worden.

Denk bijvoorbeeld aan het ontstaan van negatieve *feedbackloops*. Wanneer uit data bijvoorbeeld blijkt dat zwarte mannen een grotere kans hebben om na vrijlating weer terug te komen in de gevangenis, dan zal het algoritme bepalen dat zwarte mannen langer in de gevangenis moeten blijven. Dit heeft vervolgens weer invloed op de gevolgcijfers: zwarte mannen zitten dan daadwerkelijk langer in de cel, wat vervolgens weer wordt versterkt door het algoritme, etc. Hierbij wordt geen rekening gehouden met het feit dat de cijfers gebiased zijn als gevolg van menselijk politiewerk, waarbij mogelijk sprake is van profilering. Biases van algoritmen worden dus voornamelijk veroorzaakt door biases ↴ van mensen. Het is dus maar de vraag of een robotrechter daadwerkelijk objectiever is.

» Algoritmes zijn even bevooroordeeld als de mensen die ze maken. «

-- Sanne Blauw, De Correspondent

Hoe waarborgen we rechtvaardige procedures?

De vraag wie er kunnen en mogen beslissen over wat wel en niet ethisch verantwoord is in relatie tot de ontwikkeling en toepassing van AI wordt nog te weinig gesteld. Vooralsnog zijn het voornamelijk de grote tech-bedrijven die de touwtjes in handen lijken te hebben. Zij beschikken over verreweg de meeste resources, zoals data, rekenkracht, geld én kennis. Met name die laatste wordt nog weleens onderschat. Denk bijvoorbeeld aan het moment dat Mark Zuckerberg in 2018 werd verhoord in het Amerikaanse Congres over de privacy-lek bij Facebook. Senatoren stelden hem vragen waaruit duidelijk bleek dat ze geen idee hadden van wat Facebook precies is en doet. De vraag van Utah Senator Orrin Hatch spande toch wel de kroon. Hij vroeg zich af op welke wijze Facebook een verdienmodel in stand kon houden, terwijl gebruikers niet voor de service hoeven te betalen.

» How do you sustain a business model in which users don't pay for your service? «

-- Orrin Hatch, US Senator

Het antwoord van de verbaasde Zuckerberg was '*Senator, we run adds*'. Deze vorm van digitale ongeletterdheid maakt het lastig om complexe technologieën te kunnen reguleren. Toch lijken overheden nog steeds een flinke vinger in de pap te hebben. Zo heeft de Russische president Vladimir Poetin in 2019 een controversieel wetsvoorstel ondertekend waardoor het een misdaad is om de staat te 'minachten' en 'nepnieuws' online te verspreiden. In hetzelfde jaar legde de Iranese overheid het internet plat tijdens protesten tegen de toenemende inflatie. Het werd hierdoor voor demonstranten lastiger om zich te organiseren en voor journalisten lastiger om informatie te krijgen over de situatie. Landen lijken hiermee vooral hun eigen digitale infrastructuur en machtspositie te willen beschermen.

Ook wanneer het op ethiek aankomt lijken landen zich nog te veel naar binnen toe te keren. Zo lanceerden verschillende overheden hun éigen ethische richtlijnen ↴. Van Duitsland en Oostenrijk tot Australië en de Verenigde Staten. Het is hierbij niet de vraag welk land de béste ethische richtlijnen heeft, het gaat er juist om op welke vlakken overeenstemming gevonden kan worden. Ethiek is geen wedstrijd, maar een teamsport. Zonder mondiale aanpak is het vrijwel onmogelijk om betrouwbare AI-toepassingen te kunnen ontwikkelen. Juist de verschillende, vaak cultureel bepaalde, invalshoeken moeten aansluiting met elkaar vinden. Deze invalshoeken ontbreken nu nog te vaak in het globale debat over AI en ethiek. Momenteel lijkt de ontwikkeling van AI nog te veel op een wedstrijd ↴. De financiële belangen lijken zwaarder te wegen dan de morele belangen. Lees hier meer over in het scenario 'Winner takes all'.

Naast de onethische implicaties van AI-toepassingen moet dus ook worden nagedacht over de totstandkoming van ethische richtlijnen, anders blijft er alsnog sprake van 'onethische ethiek'.

1.3

Een kwestie van ethisch perspectief

Binnen de ontwikkeling van AI ↴ hebben we te maken met uiteenlopende ethische ↴ vraagstukken. Wie is er verantwoordelijk bij een botsing met een zelfrijdende auto? Hoe zwaar weegt het recht op privacy ↴ ten opzichte van de vrijheid van informatie van internet-gebruikers? En in hoeverre kan een robotrechter objectief oordelen? Deze vraagstukken zijn allesbehalve eenduidig en daardoor enorm complex. Het is bijvoorbeeld sterk de vraag of het überhaupt wenselijk is dat een robotrechter objectief is. Wat de mens als rechtvaardig bestempelt is niet altijd in een formule te vangen en hangt sterk af van de context en eventuele verzachtende omstandigheden.

> Om te kunnen oordelen over dergelijke vraagstukken is het van belang om te beseffen dat er verschillende perspectieven zijn. Er bestaat vaak geen universeel aangenomen waarheid. Het is daarom ook lastig om consensus te vinden over wat ethisch gezien de beste oplossing is. We zijn het er in het algemeen wel over eens dat bijvoorbeeld privacy belangrijk is. Maar op de vraag of Apple de FBI toegang moet geven tot de gegevens van de iPhone van een terrorist moest in 2016 de rechter nog aan te pas komen. Dergelijke vraagstukken zijn niet alleen contextafhankelijk, het oordeel is tevens afhankelijk van de ethische stroming die wordt gehanteerd.

Verschillende vormen van ethiek

In de discussies lopen ethiek en moraliteit vaak door elkaar heen. Toch is er een duidelijk verschil. Moraliteit is het geheel van meningen, beslissingen en acties waarmee mensen (individueel of collectief) uitdrukking geven aan wat zij denken dat goed of juist is. Ethiek is vervolgens de systematische reflectie op wat moreel is (Van de Poel & Royakkers, 2011). Verschillende ethische richtlijnen ↴, zoals de *guidelines* van de Europese Commissie zijn dus eigenlijk morele richtlijnen. Het moraal gaat over het handelen zelf,

terwijl de ethiek de studie ervan is. Binnen de studie over wat moreel is kunnen verschillende deelgebieden onderscheiden worden. Grofweg kan er onderscheid worden gemaakt tussen twee benaderingen, namelijk de normatieve benadering en de non-normatieve benadering.

De normatieve benadering

Binnen de normatieve ethiek worden duidelijke morele posities ingenomen. ‘Goed’ en ‘fout’ worden hierbij vastgelegd in algemene basisprincipes, oftewel beginselen. Deze principes moeten mensen houvast geven over hoe te handelen en dienen als basis voor het reguleren van gedrag. Deze benadering wordt ook wel prescriptieve ethiek genoemd, omdat het mensen regels en principes voorlegt over hoe zij zich zouden moeten gedragen.

De non-normatieve benadering

Binnen de non-normatieve benadering worden geen morele posities ingenomen. Hierbij wordt onderscheid gemaakt tussen descriptieve ethiek, meta-ethiek en toegepaste ethiek. Descriptieve ethiek gaat over het beschrijven en begrijpen van wat mensen als 'het goede' bestempelen. Het wordt daarom ook wel beschrijvende ethiek genoemd. Wanneer de nadruk ligt op het bestuderen van centrale begrippen uit de ethiek (verantwoordelijkheid, vrijheden & rechten en rechtvaardigheid) wordt gesproken van meta-ethiek. Hierbij wordt tevens gekeken of het mogelijk is om een ethisch raamwerk te creëren dat in elke situatie toegepast kan worden, onafhankelijk van onze eigen opvattingen. Toegepaste ethiek richt zich op specifieke domeinen, zoals bio-ethiek, bedrijfsethiek of medische ethiek. Hierbij staan specifieke problemen uit de praktijk centraal. De vraag is in hoeverre principes uit de normatieve ethiek toepasbaar zijn en daarmee antwoord geven op de begrippen uit de meta-ethiek.

Aangezien AI-toepassingen in verschillende domeinen worden gebruikt, is het niet eenvoudig om duidelijke ethische kaders op te stellen. Wat ‘goed’ of ‘juist’ is hangt steeds af van de specifieke context. Veel mensen zijn er zich daarbij niet van bewust dat naast de verschillende contexten ook verschillende stromingen en uitgangspunten vaak door elkaar heen lopen. Dit maakt ethische discussies omtrent de toepassingen van AI vaak rommeling.

Verschillende stromingen binnen ethiek

Mensen vinden verschillende dingen belangrijk, afhankelijk van de context. In sommige gevallen ligt de nadruk op de handeling zelf en in andere gevallen meer op de consequenties die het handelen heeft. En soms ligt de nadruk juist op de intentie van degene die de handeling uitvoert. Binnen de ethiek bestaan er dan ook verschillende stromingen en opvattingen die soms lijnrecht tegenover elkaar staan. Wat ‘het goede’ is en hoe ‘juist te handelen’ zijn vragen die de mens al eeuwen bezighoudt. Met de komst van AI zijn deze vragen relevanter dan ooit. Technologieën krijgen een steeds grotere mate van autonomie en zouden dus ook in steeds grotere mate in staat moeten zijn om dergelijke vraagstukken zelfstandig te behandelen.

Beginselethiek

Bij beginselethiek wordt steeds een beginsel als uitgangspunt genomen. Denk bijvoorbeeld aan eerbied voor het leven en menselijke waardigheid. Bij de oplossing van een ethisch probleem moet recht gedaan worden aan een of meer van deze beginselen of principes. Het beginsel dient hierbij te allen tijde toegepast te worden, onafhankelijk van de gevolgen. Iemand handelt dus moreel juist als zijn daad overeenkomst met het beginsel. Sommige daden kunnen als goed worden beschouwd, zelfs als de gevolgen ervan negatief zijn. En vice versa. Deze gedragsregels, oftewel normen, zijn afspraken over hoe wij met elkaar omgaan. Hoewel er tegenover het schenden van deze normen vaak

geen concrete sancties staan, worden de gedragsregels door een samenleving als geheel wel in stand gehouden. Veel religies hanteren bijvoorbeeld dergelijke gedragsregels, zoals de 10-geboden in het Christendom. Sommige normen zijn als wetten geformuleerd, zoals het verbod om te discrimineren.

Beginselethiek wordt ook wel plichtethiek, of deontologie genoemd. De bekendste aanhanger van de deontologische theorie is de Duitse filosoof Immanuel Kant. Hij formuleerde in zijn '*Kritik der praktischen Vernunft*' in 1788 het 'categorische imperatief'. De bekendste formulering hieruit is 'Handel zodanig dat je zou willen dat je stelregel verheven werd tot een universele natuurwet'. Wanneer je jezelf bijvoorbeeld afvraagt of het is toegestaan om afval uit het autoraam te gooien, is het eenvoudig om te bedenken dat wanneer iedereen dit zou doen (omdat het de wet is) er een onleefbare wereld ontstaat. Afval op straat gooien is dan ook niet de norm.

» Handel zodanig dat je zou willen dat je stelregel verheven werd tot een universele natuurwet. «

-- Immanuel Kant

Gevolgenethiek

De gevolgenethiek stelt dat de gevolgen van een bepaalde handeling bepalen of er sprake is van 'juist handelen'. Het gedrag moet dus positieve gevolgen hebben, zelfs wanneer hierbij bepaalde beginselen ondermijnd worden. De handeling zelf wordt dus niet ter discussie gesteld, alleen de gevolgen ervan. Om de gevolgen van een handeling te beoordelen worden waarden gebruikt. Een waarde is een doel dat we als samenleving willen bereiken door middel van ons gedrag. Denk bijvoorbeeld aan rechtvaardigheid of vrijheid. Dit worden ook wel eindwaarden genoemd. Eigenschappen die mensen helpen om deze eindwaarden te bereiken, worden ook wel instrumentele waarden genoemd, zoals behulpzaamheid en verantwoordelijkheid.

Gevolgenethiek wordt ook wel consequentialisme of utilitarisme genoemd. De grondlegger van het utilitarisme is de Britse filosoof Jeremy Bentham. Binnen de utilitaristische theorie wordt de morele waarde van een handeling afgemeten aan de bijdrage die deze handeling levert aan het algemeen nut. De vraag is dus in hoeverre de handeling bijdraagt aan de maximalisatie van geluk. Het doel van een handeling is altijd het zorgdragen voor de grootst mogelijke hoeveelheid geluk voor de grootst mogelijke groep mensen. Als in ruil daarvoor een kleine groep mensen negatieve gevolgen moet ondervinden, dan wordt dit voor lief genomen.

» Het doel heiligt de middelen «

Deugdethiek

Bij de deugdethiek staan niet de regels of bepaalde principes centraal binnen het moreel oordelen, maar het karakter van de persoon die handelt. Ook hier staat het handelen los van de expliciete gevolgen die het heeft. Om goed te kunnen handelen zijn bepaalde karaktereigenschappen nodig, oftewel deugden. Een deugd is een positieve karaktereigenschap die het gedrag aanstuurt. Wanneer deugden worden gebruikt bij het oordelen van het handelen, wordt niet zozeer gekeken naar de incidentele handeling, maar naar de persoon zelf en zijn of haar bedoelingen.

Aristoteles wordt gezien als de grondlegger van de deugdethiek. In tegenstelling tot de beginsel- en gevolgenethiek wordt hierbij de mens zelf in overweging genomen. Goede daden zijn volgens Aristoteles daden waarvan je een beter mens wordt. De mens moet volgens deze opvatting dus aan zichzelf blijven werken. Deugden kunnen zich ontwikkelen. Een deugd wordt gezien als een soort gulden middenweg tussen meer extremere gedragskenmerken. Zo is dapperheid een deugd, die tussen overmoed en angst in ligt.

» Een goede daad verrichten is makkelijk; de gewoonte ontwikkelen om dat altijd te doen niet. «

-- Aristoteles

Niet elke grote denker valt in te delen binnen de drie stromingen. Zo stond de Duitse filosoof Friedrich Nietzsche bekend als 'ethiek-criticus'. Volgens hem waren ethische opvattingen in die tijd (1887) gebaseerd op een 'slavenmoraal'. Mensen waren volgens hem volgzaam en dachten niet meer zelf na. Nietzsche wilde dat de mens zich losmaakte van dit (destijds voornamelijk Christelijke) moraal. De mens moest zijn eigen ethiek gaan ontwerpen en hierbij zijn eigen normen en waarden scheppen. Ethiek is volgens Nietzsche geen kwestie van het volgen van een plicht of deugd, maar van persoonlijke smaak.

Verschillende opvattingen over ethisch verantwoord

Om over de verschillende ethische discussies binnen de ontwikkeling van AI-systemen te kunnen oordelen is het van belang om te bepalen vanuit welke ethische stroming geredeneerd wordt. De vraag of het verantwoord is wanneer het leven van een persoon wordt opgeofferd om de levens van meerdere personen te redden (trolleyprobleem) hangt sterk af van het uitgangspunt. Er moet hierbij onderscheid worden gemaakt tussen de handeling (beginselethiek), de gevolgen van de handeling (gevolgenethiek) en de persoon – afhankelijk van de uitkomst van het aansprakelijkheidsvraagstuk ↴ in dit geval de inzittende, de fabrikant of het systeem zelf – die handelt (deugdethiek).

Wat moet een auto doen wanneer er een groep mensen bij een zebrapad oversteekt en de auto niet meer op tijd kan remmen? Doorrijden om de inzittende te sparen of uitwijken om de overstekende mensen te sparen? Door de uitwijkmanoeuvre komt de inzettende echter wel om het leven.

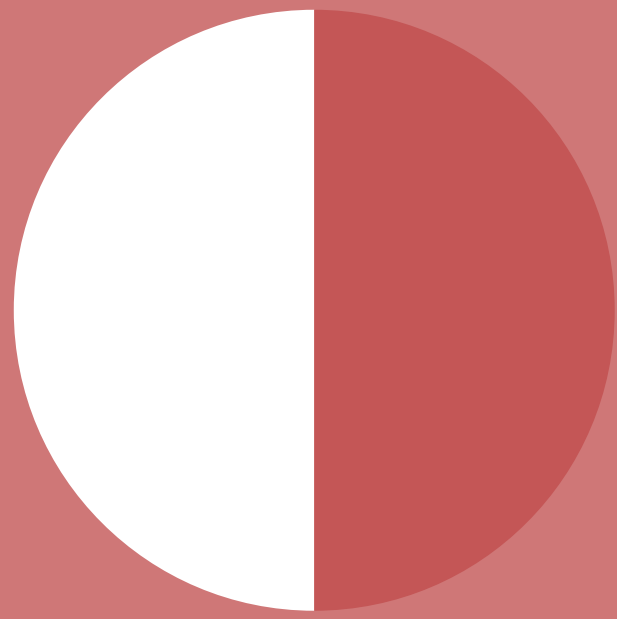
De vraag is dus of het moreel te verantwoorden is wanneer de auto niet ingrijpt en het leven van de inzittende beschermt, in plaats van het leven van de groep voetgangers. Vanuit de gevolgenethiek is dit niet te rechtvaardigen; het gevolg is immers dat er een groep voetgangers om het leven komt, in plaats van alleen de inzittende. Vanuit de beginselethiek kan gesteld worden dat het doorrijden wel te rechtvaardigen is; door in te grijpen zou de auto van zijn natuurlijke koers afwijken en is het systeem verantwoordelijk voor de dood van de inzittende. En doden mag niet, ook niet als je daarmee de levens van andere mensen spaart. Vanuit de deugdethiek kan vanuit een vergelijkbaar vertrekpunt juist het tegenovergestelde worden geredeneerd. Wanneer de auto de mogelijkheid heeft om in te grijpen maar dit niet doet, is er sprake van ondeugdelijkheid en handelt de auto immoreel. Je dient te allen tijde rekening te houden met anderen en verantwoordelijk te handelen.

Toch zijn dergelijke vraagstukken niet altijd eenduidig. Denk bijvoorbeeld aan de vraag of Apple de FBI toegang moet geven tot de gegevens van de iPhone van een terrorist. In 2016 stond deze vraag centraal in een rechtszaak. Het jaar ervoor schoot een terrorist in San Bernardino veertien mensen dood. De FBI heeft de iPhone van de man destijds in beslag genomen, om zo beschikking te kunnen krijgen over belangrijke data. Zoals informatie over andere contactpersonen en mogelijke medeplichtigen. De iPhone is uiteraard beveiligd met een pincode. Wanneer er 10 keer een foute code wordt ingevoerd, wordt alle data van het toestel gewist. De FBI heeft daarom Apple gevraagd om te helpen deze beveiliging te omzeilen. Een dergelijke methode bestond nog niet en Apple wilde ook niet meewerken aan het ontwikkelen ervan. Nu lijkt het een relatief eenvoudig vraagstuk, als je ervan uit zou gaan dat het slechts om één telefoon gaat. Maar dat was niet het geval. Door het mogelijk te maken om de beveiliging te omzeilen, zou Apple namelijk de beveiliging van álle iPhones moeten verzwakken. Apple werkte al jaren aan

de gewoonte
ontwikkelen
om dat altijd
te doen
niet.

-- Aristoteles

Een
goede daad
verrichten
is makkelijk



een betere beveiliging en dat zou hun reputatie sterk kunnen schaden. Daarnaast wilde ze een voorbeeld stellen: wanneer ze de weg voor de FBI opende om persoonlijke gegevens van gebruikers in te zien was Apple bang dat het hek van de dam was en andere overheidsdiensten met vergelijkbare verzoeken zouden komen. Het recht op privacy kwam hierdoor in het geding. De stelling van Apple was daarom dat alleen de gebruiker zelf toegang heeft tot de gegevens van een vergrendelde iPhone en niemand anders.

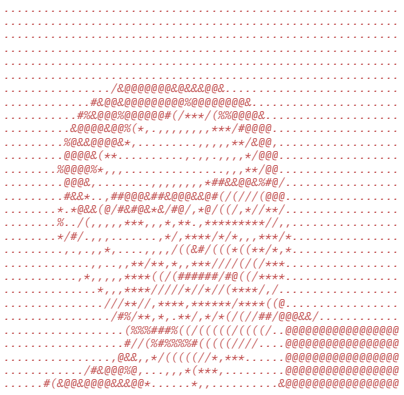
Om te kunnen bepalen of Apple hier een moreel verantwoord standpunt inneemt is lastig te bepalen. Het is namelijk niet eenduidig wat in dit geval de norm is en of privacy zwaarder moet wegen dan het bestrijden van terrorisme. Het is daarmee ook onduidelijk wat precies deugdelijk handelen is en welke gevolgen in het belang zijn van de grote groep. Een waarborging van de privacy of het aanpakken van terrorisme? Dit heeft ook te maken met de impact op de korte termijn versus de lange termijn. Het ging nu om een enkele casus, maar het leed van heel veel mensen zou in de toekomst bespaard kunnen worden wanneer terroristen effectiever aangepakt zouden kunnen worden. De rechter heeft uiteindelijk besloten dat Apple de FBI moest helpen om de iPhone van de terrorist te ontgrendelen. In tegenstelling tot wat eerder geëist werd, namelijk het creëren van een nieuwe ‘achteringang’ voor de FBI, hoefde Apple alleen de toegang tot de specifieke iPhone te garanderen.

Of er met de uitspraak recht is gedaan aan alle belanghebbenden is onduidelijk. Er bestaat in veel gevallen geen consensus over wat de beste oplossing is. Uiteindelijk gaat dit over de vraag wat de beste samenleving is en daar komen we vooralsnog niet zomaar uit. Wat is bijvoorbeeld beter? Een samenleving waarin een grote groep mensen marginaal gelukkig is of een samenleving waarin een kleinere groep heel erg gelukkig is? Er bestaat geen absolute filosofische theorie.

Uiteindelijk zou het aan de samenleving zelf moeten zijn om te bepalen wat de beste samenleving is. Het is een democratisch vraagstuk. En zelfs wanneer we als samenleving consensus vinden over wat we belangrijk vinden staat de vraag hoe je dit integreert in een AI-systeem nog open. Maar, first things first. Voordat we kunnen kijken op welke wijze we ethische richtlijnen kunnen integreren in het ontwerp van AI-toepassingen moeten we in kaart brengen welke ethische richtlijnen er zijn (beginselethiek) en in hoeverre deze standhouden in de praktijk (gevolgen- en deugdethiek).

» Mijn ethiek is niet per definitie ook jouw ethiek. «

-- Patrick van der Duin, STT



AI governance - the good, the bad and the ugly

Door Marijn Janssen,
Professor in ICT & Governance,
TU Delft

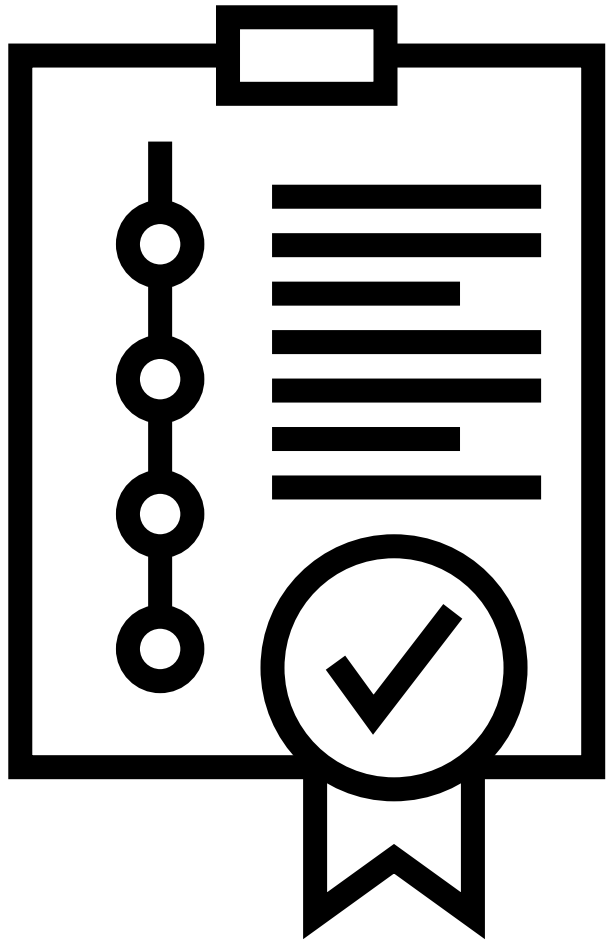
Artificiële intelligentie (AI) wordt op verschillende plekken ingezet om tot een betere maatschappij te komen. Zo helpt AI om ziekten in een vroeg stadium te detecteren en om sociale problemen aan te pakken (*'the good'*). Tegelijkertijd wordt AI gebruikt om te beïnvloeden op welke politieke partij je stemt en om op grote schaal te frauderen (*'the bad'*). AI kan daarnaast ongemerkt menselijke vooroordelen overnemen en ongezien het leven van alledag dicteren. AI leest je gedrag uit om je gezondheid en verzekeringspremies te bepalen en discrimineert bevolkingsgroepen zonder dat dit meteen duidelijk is (*'the ugly'*, oftewel de gemene). Het is met name bij dat laatste waar het venijn zit, omdat de onbedoelde effecten niet direct zichtbaar zijn en we er ons vaak niet bewust van zijn. Waar 'de goede' en 'de slechte' directer zichtbaar zijn, is 'de gemene' minder zichtbaar en moeilijker om te achterhalen.

De maatschappelijke roep om transparante, verantwoorde en eerlijke AI is dan ook meer dan terecht. Veelal wordt ingespeeld op het idee dat AI bijvoorbeeld transparant moet zijn, maar wat dan precies transparant moet zijn wordt niet uitgelegd. Impliciet wordt gezegd dat we het algoritme moeten begrijpen, maar is de situatie zonder AI dan wel transparant en begrijpen we wat er in de hersenen van mensen omgaat bij het nemen van beslissingen? Onbekend maakt onbemind. Kunnen we überhaupt verwachten dat een complex algoritme transparant is en de complexiteit door alle gebruikers doorgrond kan worden? De meeste mensen nemen al niet de moeite om de voorwaarden van een app of website door te lezen en klikken zonder nadenken op akkoord. Complexe algoritmen begrijpen gaat over hogere wiskunde. Volledige transparantie is voor de meeste mensen een illusie. De vraag is dan ook hoe onze samenleving hiermee om kan gaan.

In de sciencefictionserie *Star Trek* hebben de *'Vulcans'* een oplossing. *"Logic is the cement of our civilization, with which we ascend from chaos, using reason as our guide"* (T'Plana-Hath, Matron of Vulcan Philosophy, Star Trek IV: The Voyage Home, 1986). Logica zou moeten leiden tot het ter discussie stellen van vooroordelen, zorgen dat je alle feiten op een rijtje hebt en vermijden dat men alleen ziet wat men wil zien. De logica gevolgd door de Vulcans leidt tot een technocratische kijk en heeft ook nadelen, omdat er altijd mensen zijn die politiek bedrijven en het systeem bespelen. Informatie is bijna nooit compleet en niet alle addertjes onder het gras zijn duidelijk. Logica is nodig, maar moet bijgestaan worden door een geschikt bestuursmodel.

De gevoeligheid van algoritmen voor veranderingen en hun grillige natuur door de vele variabelen maken het moeilijk om AI algoritmen aan te sturen. Zo'n bestuursmodel moet de technologie niet alleen aansturen, maar ook bevatten. Hiervoor is technische kennis een voorwaarde, maar niet voldoende. Er is een bestuursmodel nodig waarbij er een goed gedefinieerd afsprakenstelsel is om met AI om te gaan. Voordat AI gebruikt mag worden zou het getoetst moeten worden, net zoals we pas medicijnen op de markt brengen na zorgvuldige toetsing en inzicht in mogelijke bijeffecten. AI-systemen moeten voldoen aan eisen om te voorkomen dat we in een *surveillance* staat eindigen, waar al ons handelen zonder dat we het weten of willen wordt beïnvloed door AI.

Daarnaast is organisatorische kennis nodig om de ethische consequenties te begrijpen. Bij automatisch rijden met AI is een stuur misschien niet meer nodig, maar dat neemt niet weg dat er geen besturing meer bestaat. Het model van besturing moet ervoor zorgen dat de slechte en gemene dingen van AI niet voorkomen. De aansturing van AI is nu nog niet volwassen, terwijl we AI wel op grote schaal gebruiken. Laten we met *AI governance* op weg gaan naar een AI-maatschappij, waar de mens aan het stuur zit en de computer ondersteunt.



2 Ethische richtlijnen voor AI

Pagina’s: 30
Woorden: 6643
Leestijd: ca. 50 minuten

Ethische richtlijnen voor AI

De tijd dat AI-systemen ↴ voornamelijk bekendheid verwierven door het verslaan van schaakgrootmeesters lijkt achter ons te liggen. Uiteraard rezen hierbij de nodige vragen op, want als een computer ons kan verslaan in een intellectueel spel als schaken of Go, op welke vlakken zouden computers ons dan nog meer kunnen verslaan? Wat doet dit met de relatie tussen mens en technologie? En wat betekent dit voor ons mens-zijn? Interessante vraagstukken, maar nog altijd vrij abstract en filosofisch van aard. Inmiddels worden AI-systemen toegepast in uiteenlopende domeinen en komen er steeds meer toegepaste ethische vraagstukken aan het licht. Dit gaat niet meer over *'what if…'*, maar over *'what now'*.

> Zo zijn de eerste doden door het gebruik van autonome voertuigen een feit, wordt ons recht op privacy ↴ ondermijnd door de inzet van locatieapps en worden groepen mensen benadeeld door fraudedetectiesystemen.

Nu de hypotheses zijn bevestigd lijken we wakker te zijn geschud. Van de private sector tot maatschappelijke organisaties en overheden, ze zijn allemaal in de pen geklommen om ethische richtlijnen te formuleren. Een mooie en belangrijke eerste stap. Zeker voor aanhangers van de beginselethiek. Het is hierbij interessant om te kijken wat de overeenkomsten en verschillen tussen deze richtlijnen zijn.

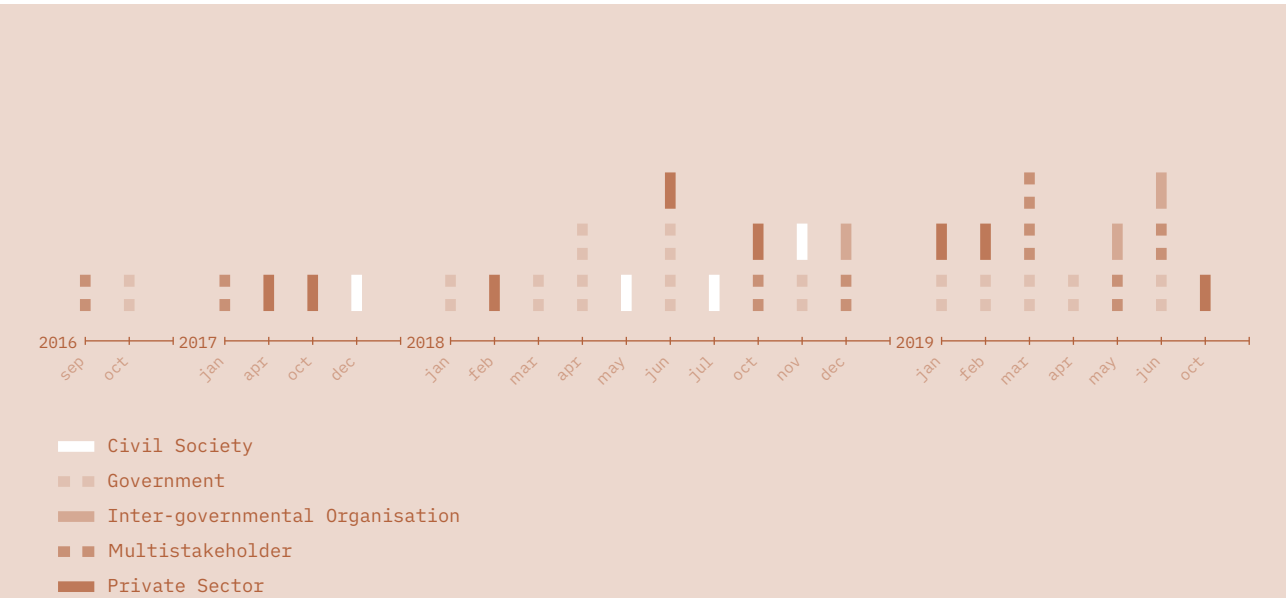
De vraag is echter of het ontwikkelen van richtlijnen voldoende is. Uiteindelijk moeten deze richtlijnen vertaald worden naar de praktijk. En zoals altijd lijkt de praktijk weerbarstiger te zijn dan de theorie. Hoe gaan we bijvoorbeeld om met conflicterende waarden? En welke uitdagingen liggen er nog voor ons? Het is van belang om ook vanuit de gevolgenethiek te redeneren. Gelukkig zijn er steeds meer instanties die aan verschillende *checklists*, *assessments* en *toolkits* hebben gewerkt. Op deze wijze komen we weer een stap dichterbij het in praktijk brengen van ethiek ↴. Het is echter nog wel van belang om te onderzoeken of aan de hand van deze tools ook echt de stap kan worden gemaakt naar het ontwerpproces. Kunnen we met deze tools daadwerkelijk deugdelijk handelen?

Van corporate tot government

In de afgelopen jaren hebben uiteenlopende bedrijven, onderzoeksinstellingen en overheidsorganisaties verschillende principes en richtlijnen opgesteld voor *ethical AI*. Zowel op nationaal en continentaal niveau, als op mondiaal niveau.

Opkomst van ethische richtlijnen

Om overzicht te scheppen in het gefragmenteerde gesprek over de ontwikkeling van ethisch verantwoorde AI-toepassingen hebben onderzoekers van het *Berkman Klein Center* in 2020 een analyse uitgevoerd van de 36 meest prominente AI-richtlijnen. Deze analyse is onder andere uitgewerkt in een overzichtelijke tijdslijn.



‘A map of Ethical and Rights-based Approaches to Principles for AI’ – Source: Berkman Klein Center (2020).

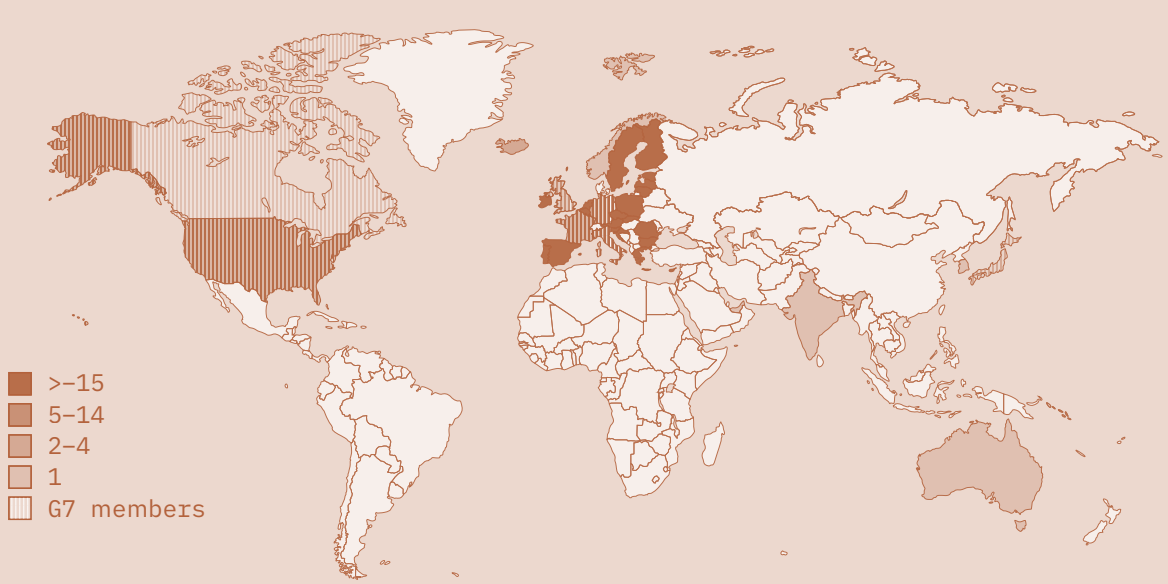
In de tijdslijn is duidelijk te zien dat de frequentie van publicaties in de afgelopen jaren sterk is toegenomen. Er wordt hierbij onderscheid gemaakt tussen principes en richtlijnen van:

> Maatschappelijke organisaties, zoals de *‘Top 10 Principles for ethical AI’* van de UNI Global Union (2017);
> Overheden, zoals de *‘AI Ethics Principles & Guidelines’* van Dubai (2019);
> Intergouvernementele organisaties, zoals de *‘Principles on AI’* van de OECD (2019);
> Multistakeholders, zoals de *‘Beijing AI Principles’* van de Beijing Academy of AI (2019);
> Particuliere sector, zoals de *‘Everyday Ethics for AI’* van IBM (2019).

Ethiek is overduidelijk niet langer alleen een Europese aangelegenheid. ‘Zelfs’ de Chinese overheid lanceerde in 2019 de *’Governance Principles for a New Generation of AI’*. China beseft immers ook wel dat als ze internationaal zaken willen blijven doen in AI, ze dan ook het gesprek over ethisch verantwoorde toepassingen moeten aangaan. Daarbij houden ze zelf graag de touwtjes in handen en zijn frauduleuze praktijken ook voor de Chinese overheid onwenselijk.

Spreiding van ethische richtlijnen

Er zijn ethische principes en richtlijnen in verschillende soorten en maten. Ondanks de grote hoeveelheid is er slechts een beperkte spreiding van richtlijnen. Onderzoekers van ETH Zürich analyseerden in 2019 maar liefst 84 ethische richtlijnen die de afgelopen jaren wereldwijd zijn gepubliceerd. Van de private sector tot maatschappelijke organisaties en overheden. Uit het onderzoek blijkt dat de meeste ethische richtlijnen uit de VS (21), Europa (19) en Japan (4) komen.



‘Geographic distribution of issuers of ethical AI guidelines by number of documents released’ – Source: ETH Zürich (2019).

De grootste ‘richtlijn-dichtheid’ is te vinden in het Verenigd Koningrijk. Daar werden maar liefst 13 ethische richtlijnen gepubliceerd. Lidstaten van de G7 brachten de meeste ethische richtlijnen voort. De G7 (Groep van 7) bestaat uit zeven belangrijke industriële staten, namelijk Canada, Duitsland, Frankrijk, Italië, Japan, het Verenigd Koninkrijk en de Verenigde Staten. In 1997 is ook de Europese Unie toegetreden, maar de naam is daar niet op aangepast. Dat juist deze landen de meeste richtlijnen publiceren is geen toeval. In 2018 hebben de ministers van de lidstaten die over innovatie gaan een G7-verklaring ondertekend aangaande *Human Centric AI*. Hierin zijn ze tot een gemeenschappelijke visie gekomen waarmee wordt ingezet op een balans tussen het stimuleren van economische groei door AI-innovatie, het vergroten van het vertrouwen ↴ in en de acceptatie van AI en het bevorderen van inclusiviteit bij de ontwikkeling en implementatie van AI.

Hoewel het overzicht een momentopname is (zo zijn er na de publicatie van het onderzoek in China twee nieuwe richtlijnen gepubliceerd) geeft het wel een duidelijke verdeling weer. Het zijn tot nu toe met name de rijkere landen die de wereldwijde discussie over de regulering van AI domineren. Hoewel sommige ontwikkelingslanden betrokken waren bij internationale organisaties die richtlijnen opstelden, hebben slechts enkelen hun eigen ethische principes daadwerkelijk gepubliceerd. Dit is volgens de onderzoekers echter wel van groot belang, aangezien verschillende culturen verschillende opvattingen hebben over AI. Er is mondiale samenwerking nodig om in de toekomst zorg te dragen voor ethische AI die bijdraagt aan het welzijn van individuen en samenlevingen.

Een eerste poging om mondiale samenwerking te realiseren is gedaan door de *Organization for Economic Co-operation and Development* (OECD). De OECD, een coalitie van landen die zich inzet voor het bevorderen van democratie en economische ontwikkeling, heeft in 2019 een reeks van vijf principes aangekondigd voor de ontwikkeling en inzet van AI. Maar bijvoorbeeld China valt buiten de OECD en is dus niet meegenomen in de totstandkoming van de richtlijnen. De uiteengezette principes lijken in contrast te staan met de manier waarop AI daar wordt ingezet. Met name wanneer het aankomt op gezichtsherkenning en het toezicht op etnische groepen die met politieke dissidentie zijn geassocieerd. Maar juist wanneer er tegenstrijdige opvattingen zijn is het van belang om elkaar op te zoeken en tot een vorm van consensus te komen.

Overeenkomsten en verschillen

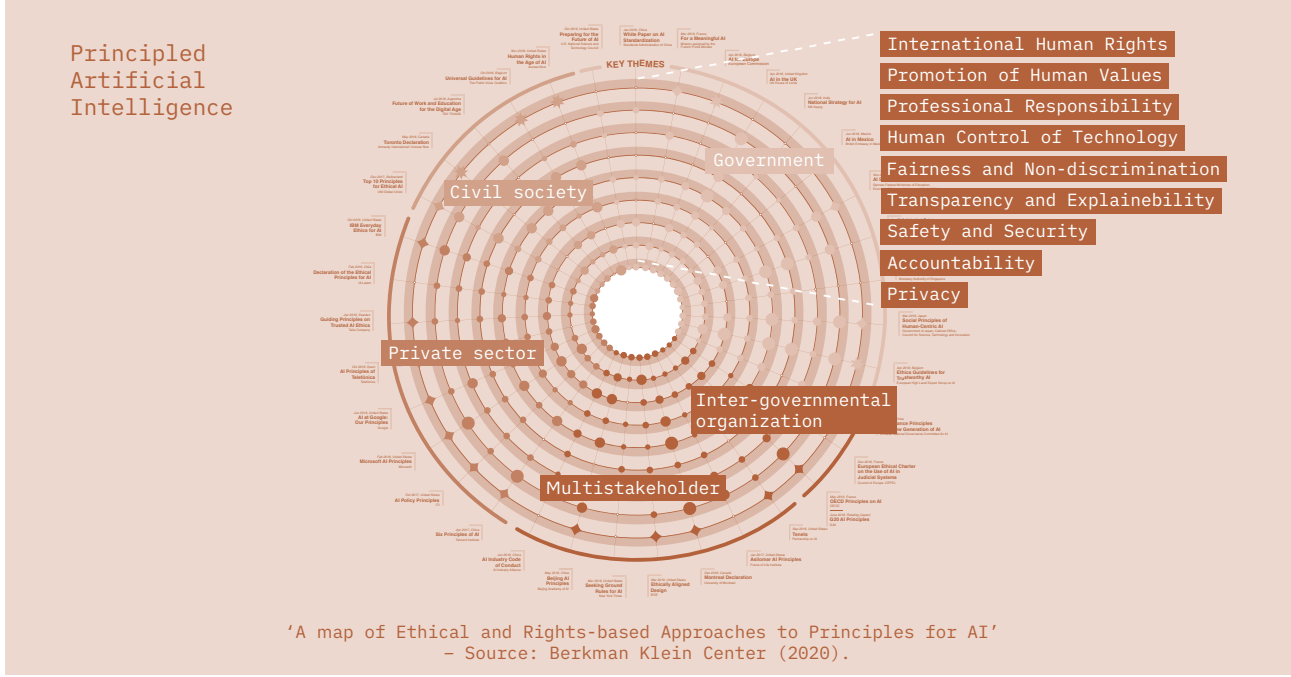
De onderzoekers van ETH Zürich keken niet alleen naar de geografische spreiding van de ethische principes en richtlijnen, maar ook naar de inhoudelijke overeenkomsten en verschillen tussen de principes. Uit het onderzoek blijkt dat hoewel geen enkel ethisch principe letterlijk overeenkomt er wel een duidelijke

convergentie zichtbaar is rondom de begrippen transparantie ↴, rechtvaardigheid ↴ & eerlijkheid ↴, betrouwbaarheid, verantwoordelijkheid ↴ en privacy. Deze principes worden in meer dan de helft van alle bronnen genoemd.

Toch zijn er significante verschillen in de manier waarop de ethische principes worden geïnterpreteerd. Met name de specifieke aanbevelingen en aandachts-gebieden die van elk principe zijn afgeleid blijken enorm van elkaar te verschillen. Zo moeten AI-systemen volgens sommige richtlijnen het beslissingsproces vooral verklaarbaar maken, terwijl het volgens andere richtlijnen noodzakelijk is dat de beslissingen ook volledig traceerbaar zijn. De onderzoekers benadrukken daarom het belang van het integreren van de verschillende richtlijnen om zo tot een mondiale overeenstemming te komen over adequate implementatiestrategieën.

Ethical principle	Number of documents	Ethical principle	Number of documents
Transparency	73/84	Beneficence	41/84
Justice & fairness	68/84	Freedom & autonomy	34/84
Non-maleficence	60/84	Trust	28/84
Responsibility	60/84	Sustainability	14/84
Privacy	47/84	Dignity	13/84
		Solidarity	6/84

‘Ethical principles identified in existing AI guidelines’
– Source: ETH Zürich (2019)



‘A map of Ethical and Rights-based Approaches to Principles for AI’
– Source: Berkman Klein Center (2020).

De onderzoekers van het *Berkman Klein Center* kwamen tot een vergelijkbaar inzicht. Op basis van een analyse van de gebruikte termen kwamen ze tot een overzicht van acht overkoepelende principes, die in grote lijnen overeenkomen met de begrippen van het onderzoek van ETH Zürich.

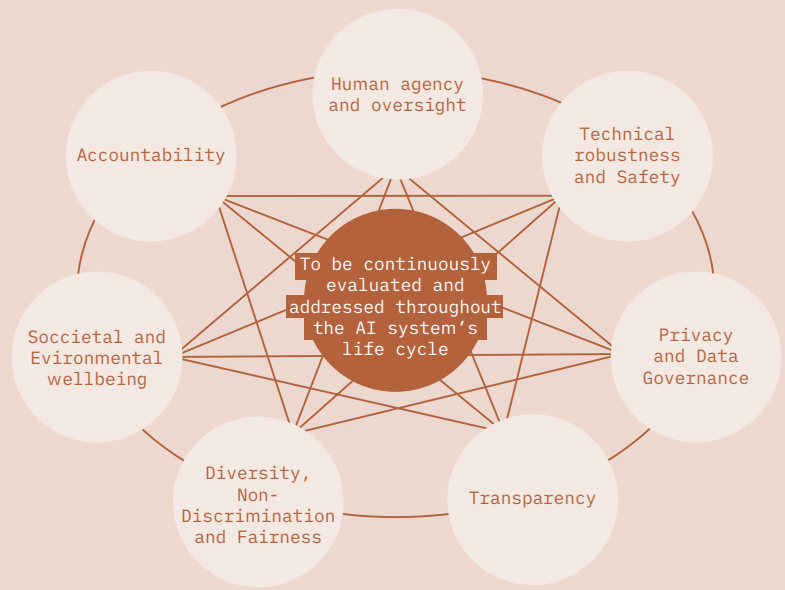
Ook hier wordt het duidelijk dat wanneer je naar de details en interpretaties kijkt er duidelijke verschillen tussen de ethische principes en richtlijnen bestaan. Niet alleen in de mate waarin bepaalde principes zijn uitgewerkt, maar vooral ook in de mate waarin verwezen wordt naar internationale mensenrechten. Zowel naar internationale mensenrechten als algemeen concept, als naar specifieke documenten, zoals de ‘*Universal Declaration of Human Rights*’ of de ‘*United Nations Sustainable Development Goals*’. In sommige ethische richtlijnen wordt zelfs expliciet gebruik gemaakt van een ‘*Human Rights Framework*’. Dit betekent dat mensenrechten de basis vormen voor de formulering van ethische richtlijnen voor de ontwikkeling van AI-toepassingen. Tegen de verwachting in zijn het vooral de richtlijnen van de private sector waarin naar mensenrechten wordt verwezen en in veel mindere mate de richtlijnen vanuit overheden.

Botsende waarden

Ethische ↴ principes en richtlijnen ↴ kunnen ons helpen om overeenstemming te vinden over wat we belangrijk vinden om zo AI-toepassingen ↴ te kunnen ontwikkelen die hieraan voldoen. De vraag is echter in hoeverre deze richtlijnen in de praktijk standhouden. Er wordt namelijk vooral gekeken naar de voorwaarden en in mindere mate naar de gevolgen. In de praktijk komen er echter dilemma's voor, waarbij waardenconflicten kunnen ontstaan. We zijn het er in de basis allemaal over eens dat je bijvoorbeeld niet mag doden. Maar wat doe je als een terrorist 100 mensen probeert neer te schieten? Mag een autonome wapendrone dan niet ingrijpen? Daarbij is er in de praktijk vaak sprake van verzachtende omstandigheden. Zo zijn we het er allemaal wel over eens dat je niet mag stelen. Maar wat als een alleen-staande moeder steelt omdat haar kind anders niets te eten heeft? Moet een robotrechter dan toch gewoon de regels volgen en de moeder volwaardige straffen opleggen? Ethische richtlijnen zouden niet alleen moeten gaan over wat we belangrijk vinden, maar ook over hoe belangrijk we verschillende waarden ten opzichte van elkaar vinden. En in welke omstandigheden we dat vinden.

> Wanneer we de '*Ethics Guidelines for Trustworthy AI*' van de Europese Commissie als uitgangspunt nemen, dan valt het op dat dergelijke *trade-offs* nauwelijks benoemd worden. Sterker nog, er wordt aangegeven dat alle vereisten even belangrijk zijn en elkaar ondersteunen.

Er is slechts een kleine alinea over trade-offs in het rapport opgenomen. Hierin wordt voornamelijk benoemd dát er trade-offs kunnen optreden en er hierbij afwegingen moeten worden gemaakt, maar welke afwegingen dat precies zijn en op elke wijze deze afwegingen gemaakt kunnen worden blijft onduidelijk. Er wordt alleen aangegeven dat de afwegingen moeten worden geëvalueerd en gedocumenteerd.



'Interrelationship of the seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle'
– Source: High Level Expert Group on AI (2019).

Wat vinden we écht belangrijk?

Volgens de richtlijnen van de Europese Commissie zijn de zeven '*key requirements*' even belangrijk. Wanneer je mensen echter vraagt een prioritering aan te geven, dan blijken er overduidelijke verschillen te bestaan tussen de waardering van verschillende principes. Uit eigen onderzoek blijkt dat de waarden autonomie ('*Human agency and Oversight*'), privacy ↴ ('*Privacy and Data Governance*') en gelijkwaardigheid ('*Diversity, Non-Discrimination and Fairness* ↴') veel hoger scoren dan onder andere uitlegbaarheid ↴ ('*Transparency*') en aansprakelijkheid ('*Accountability* ↴'). We legden respondenten (n=108) zeven principes voor met de vraag deze te prioriteren: 1 = een lage prioriteit en 7 = een hoge prioriteit.

Principe	Gemiddelde score
Menselijke controle behouden (autonomie)	4,81
Persoonsgegevens beschermen (privacy)	4,80
Ongelijkwaardigheid bestrijden (voorkomen biases)	4,67
Menselijke keuzes optimaliseren (efficiëntie)	3,88
Uitlegbaarheid van het systeem vergroten (oplossen blackbox)	3,77
Aansprakelijkheidsprobleem oplossen (wetgeving)	3,34
Internationale positie in AI verbeteren (geopolitiek)	2,73

Ook uit de overige vragen blijkt dat autonomie als enorm belangrijk wordt ervaren. Zo geeft de overgrote meerderheid van de respondenten aan dat strenge wetgeving en regulering nodig zijn om de controle over AI te behouden, ook als dit de ontwikkeling van de technologie vertraagt. Wanneer je echter inzoomt zie je dat er verschillen ontstaan, met name tussen de verschillende groepen respondenten. We legden dezelfde vragenlijst voor aan drie verschillende groepen, namelijk AI-experts, bestuurders en studenten.
Uit de analyse blijkt dat AI-experts geneigd zijn om meer risico te nemen met AI dan de andere groepen. Zo staan er in verhouding meer experts voor open dat AI-systemen autonome beslissingen maken en zijn ze bereid een hogere onzekerheidsmarge van de systemen te accepteren. Wanneer het aankomt op het verkrijgen van een internationale voorsprong in de ontwikkeling van AI ↴ zijn het juist bestuurders die bereid zijn om grotere risico's te nemen. Relatief meer bestuurders geven aan dat we er als land alles aan moeten doen om een voorsprong op de ontwikkeling van AI te krijgen, ook als dit tot internationale spanningen leidt. Studenten zijn in verhouding meer bezorgd over hun privacy.

Het wordt nog interessanter wanneer de verschillende principes tegen elkaar afgezet worden. In het algemeen krijgen zowel privacy als gelijkwaardigheid een zeer hoge score. Wanneer deze echter tegen elkaar afgezet worden blijkt dat het overgrote deel van de respondenten liever altijd over zijn/haar data kan beschikken, dan dat ze hun data afstaan om daarmee ongelijkwaardigheid te kunnen bestrijden. Op andere vlakken is veel meer verdeeldheid. Zo geeft een deel van de respondenten aan dat als we niet begrijpen hoe een AI-systeem aan zijn resultaten komt we het niet moeten gebruiken, terwijl het voor een evenredig deel van de respondenten niet uitmaakt hoe een AI-systeem aan zijn resultaten komt. Zij vinden het vooral belangrijk dat het systeem naar verwachting presteert. Mensen hebben dus verschillende opvattingen als het op dergelijke principes aankomt. Deze opvattingen zijn daarbij contextafhankelijk.
Het is daardoor lastig om algemene richtlijnen te vertalen naar de praktijk. In veel gevallen zullen er afwegingen gemaakt moeten worden.

» Normen en waarden zijn universeel. Het zijn de morele oordelen die verschillen tussen mensen. «

-- Wiegert van Dalen, Ethicus

Welke trade-offs zijn er?

Wanneer AI-systemen in de praktijk worden toegepast zijn er verschillende waardenconflicten denkbaar. Zowel tussen waarden, als binnen waarden. Wat we belangrijk vinden hangt onder andere af van de context waarin het systeem wordt toegepast. Het maakt nogal een verschil of het gaat om een aanbeveling voor een film, een diagnose op basis van longfoto's in het ziekenhuis of een advies ten aanzien van een bedrijfsovername. Er moeten daarbij ook technische afwegingen gemaakt worden. Wanneer we bijvoorbeeld vol inzetten op transparantie, dan zal dat de mate van privacy beïnvloeden.

Het wil niet zeggen dat we dan privacy volledig moeten opgeven. Het is geen '*zero-sum game*' waarbij we het een voor het ander moeten uitruilen, maar er zullen in het ontwerp wel keuzes gemaakt moeten worden. Om waarden ten opzichte van elkaar te kunnen maximaliseren, moeten de eventuele spanningen eerst geïdentificeerd worden.

Technische trade-offs

Een belangrijke trade-off in de praktijk is tussen *accuracy* en *explainability* ↴. Methodieken die momenteel bij de ontwikkeling van AI-systemen worden gebruikt, zoals *deep learning* ↴, zijn dermate complex dat de precieze beslissingsprocessen niet volledig herleidbaar zijn. De optimalisering van dergelijke systemen vindt nu nog vaak plaats aan de hand van *trial and error*: er wordt aan de input *getweaked* om te kijken wat dit met de output doet. Wanneer je de accuratesse van het systeem wilt optimaliseren, dan zal je een stukje van je uitlegbaarheid moeten inleveren. Aan de andere kant van het spectrum staat lineaire regressie. Deze methodiek is ten opzichte van deep learning totaal niet flexibel, maar wel goed uitlegbaar ↴. Soms kiezen mensen vanuit de behoefte aan uitlegbaarheid voor deze toepassingen, zelfs als ze weten dat de relatie tussen de onderliggende variabelen niet recht evenredig is.

> *» Als je grillige data hebt en je wil daarvan kunnen leren, dan betaal je daar een prijs voor. «*
>
> -- Maarten Stol, BrainCreators

Een vergelijkbare afweging doet zich voor op het gebied van *privacy* en *accuracy*. In het algemeen geldt dat hoe vollediger en omvangrijker de dataset is waarmee een AI-systeem getraind wordt, hoe nauwkeuriger het systeem is. Wanneer AI bijvoorbeeld wordt toegepast om toekomstige aankopen van consumenten te voorspellen op basis van hun aankoopgeschiedenis, dan zal het model nauwkeuriger worden naarmate de data verrijkt wordt met bijvoorbeeld demografische gegevens.

Het verzamelen van persoonsgegevens kan echter de privacy van de klanten schaden.

Wanneer de dataset niet volledig is kan dit leiden tot vertekende of discriminerende resultaten. Hierbij kan een trade-off ontstaan tussen *privacy* en *fairness*. Organisaties kunnen verschillende technische maatregelen nemen om dit risico te beperken, maar de meeste van deze technieken zorgen ervoor dat de nauwkeurigheid van het systeem lager wordt. Wanneer je bijvoorbeeld wilt voorkomen dat mensen bij een kredietbeoordeling geclassificeerd worden op basis van hun postcode als indicator voor hun ras, dan zou de postcode door het model buiten beschouwing moeten worden gelaten. Dit kan helpen om discriminerende uitkomsten te voorkomen, maar het kan ook leiden tot een minder nauwkeurige meting. Dit komt omdat de postcode mogelijk ook een indicator is voor een legitieme factor, zoals baanzekerheid, waardoor het de nauwkeurigheid van de uitkomsten verlaagt.

Deze afwegingen werken weer door op de veiligheid van het systeem. Als je model minder accuraat is, dan is de kans op fouten groter wat invloed heeft op de veiligheid. Als je *safety* optimaliseert en daardoor de uitlegbaarheid deels opgeeft, dan gaat dat weer bijten in je *accountability*. Aansprakelijkheid is namelijk lastiger te herleiden wanneer de uitlegbaarheid van het systeem beperkt is. Trade-offs vallen daarmee in een spectrum. Mensen moeten kiezen waar zij zich in het spectrum het meest comfortabel voelen. Daarvoor bestaat geen gulden snede, het is allemaal maatwerk.

» *An algorithm that performs exactly as intended and with perfect accuracy is not necessarily an ethical use of AI.* «

-- Kalev Leetaru, George Washington University

Alignment trade-offs

In 540 voor Christus wenste Koning Midas dat alles wat hij aanraakte in goud zou veranderen. Zijn eten en geliefden veranderden daardoor echter ook in goud. Hij werd hierdoor zo eenzaam en raakte zo uitgehongerd dat hij zijn krachten opgaf. Bij het formuleren van zijn doel dacht hij niet goed na over de gevolgen. Dit wordt ook wel het *Value Alignment Problem* (VAP) genoemd. Theoretisch gezien zou een intelligente machine die geprogrammeerd is om zoveel mogelijk paperclips te produceren álles in werking kunnen stellen om dat doel te bereiken. Nick Bostrom filosofeert in het boek *Superintelligence* ↴ dat de machine dan alles wat de productie belemmert uit de weg zal ruimen. Zelfs de mens, want die draagt immers niet bij aan de productie van paperclips. Een machine kan zo doelgedreven zijn, dat de resultaten niet overeenkomen met wat we willen.

Het is dus van belang om te bepalen wat een succesvolle uitkomst is. Het is daarbij de vraag of we op basis van wenselijkheid of op basis van werkelijkheid programmeren. Wanneer je een aantal jaar geleden de zoekterm ‘CEO’ googelde, dan kreeg je bij de afbeeldingen voornamelijk witte mannen van middelbare leeftijd te zien. Je moest een flink eindje scrollen voor de eerste afbeelding van een vrouw. Kijkend naar de statistieken is dit echter niet volledig inaccuraat. Vrouwen zijn nog steeds ondervertegenwoordigd in de hoogste management-posities. Uit de data van *Pew Research Center* blijkt dat in 2018 het percentage vrouwelijke CEO’s in de *Fortune 500* slechts 4,8% was. Wanneer je dezelfde zoekopdracht op Google Search in 2020 herhaalt dan zijn van de eerste 20 personen die worden afgebeeld, 3 personen van het vrouwelijke geslacht. Dat is 15%. Nog steeds een laag

percentage, maar wel een stuk hoger dan wat het in de werkelijkheid is. Daarbij is de vraag wie bepaalt wat een succesvolle uitkomst is. Wanneer je bijvoorbeeld een algoritme ↴ wilt inzetten die op basis van ingrediënten bepaalt wat je kunt koken, dan is het erg bepalend wie precies definieert wat succes is. Ouders willen dat hun kinderen gezond eten, maar de kinderen zelf zullen eerder voor iets lekkers kiezen. Nu ligt de vraag wat een succesvolle uitkomst is nog in de handen van een zeer kleine groep mensen.

Contextuele trade-offs

Wat we belangrijk vinden is sterk afhankelijk van de context. Denk bijvoorbeeld aan privacy. Als de dokter tijdens een spreekuur vraagt om je broek uit te doen is dat in de meeste gevallen geen probleem. Maar als de bakker hetzelfde vraagt is het een schending van je privacy. Ook voor bijvoorbeeld explainability is de context bepalend. Zo zal de mate van uitlegbaarheid die we van een systeem verwachten bij een chatbot lager liggen dan bij een zelfrijdende auto. Het belang is dus onder andere afhankelijk van de mogelijke risico’s. Zo is ook de geaccepteerde mate van subjectiviteit afhankelijk van de context. Zo zal een ‘foutieve’ aanbeveling van Netflix ons weinig schade berokkenen, maar wanneer een medische diagnose ernaast zit heeft dit veel grotere consequenties. Hoewel de subjectiviteit van het aanbevelingssysteem van Netflix veel hoger is dan de beeldherkenningssoftware ↴ van een ziekenhuis, zullen we aan die laatste veel hogere eisen stellen.

De vraag wat we belangrijk vinden hangt tevens af van het perspectief. Deze is vaak cultureel bepaald. De data voor beeldherkenningssystemen wordt in veel gevallen nog gelabeld door mensen. Een beeld van een man of vrouw met een glas bier wordt in sommige culturen gelabeld met gezelligheid, samenzijn, feest etc. Maar in andere culturen wordt hetzelfde beeld getagged met alcoholist, baldadigheid etc. Daarbij hangt het perspectief ook sterk af van leeftijd. Veel mensen oordelen bijvoorbeeld dat

het niet menselijk is om bejaarden te laten verzorgen door robots. Uit de Nationale Toekomstmonitor van 2019 blijkt dat de meerderheid van de Nederlanders negatief is over intieme relaties met een robot (79%).
Maar veel hulpbehoevende ouderen zelf vinden het een ideale uitkomst. Bijvoorbeeld ouderen die niet meer in staat zijn om zelfstandig te eten. Ze schamen zich vaak als ze gevoed worden door vreemden in een tehuis. Alleen hun kinderen mogen het en anders eten ze maar niet. Robots bieden uitkomst en geven juist autonomie. Wat we onder 'welzijn' verstaan is daarbij vaak een subjectieve aangelegenheid. Emotioneel gezien willen we het leven zo lang mogelijk oprekken, zeker wanneer het om de mensen in onze eigen omgeving gaat. De vraag is echter wat de waarde van het leven is voor terminaal zieke mensen zelf. Wat zijn in dergelijke situaties de optimalisatiedoelen voor AI-systemen? Zo efficiënt, sociaal, duurzaam of humaan mogelijk? En betekent humaan dan het leven verlengen, of juist verkorten om onnodig lijden te voorkomen?

Vergelijkbare vraagstukken worden ons door de coronacrisis in 2020 ineens geforceerd opgelegd waarbij duidelijk wordt dat verschillende belangen met elkaar verweven zijn en elkaar beïnvloeden. Vanuit gezondheids-overwegingen zou het bijvoorbeeld verstandig kunnen zijn om een volledige 'lockdown' in te voeren. Maar voor mensen in ontwikkelingslanden – die in sommige gevallen afhankelijk zijn van een dagloon – betekent binnenblijven dat hun volledige inkomsten wegvallen en ze mogelijk van de honger sterven. Wat is er erger? Sterven door een virus, of sterven van de honger? Een dergelijke crisis zet de verhoudingen op scherp en benadrukt de noodzaak van het stellen van fundamentele vragen. De waarde van het leven lijkt moeilijk in meetbare waardes uitgedrukt te kunnen worden, maar toch gebeurt dat in een pandemie wel. De volledige economie gaat op slot om het leven van kwetsbare groepen te beschermen. Hoe ver moet je hierin gaan? Het zijn geen populaire vraagstukken, maar ze kunnen niet ontweken worden. Zeker wanneer AI-systemen

een grotere rol gaan spelen in besluitvormingsprocessen, moeten we bepalen waar de balans ligt tussen ratio en emotie, objectiviteit en subjectiviteit, de lange termijn en de korte termijn. Willen we 'objectieve' besluiten van AI-systemen accepteren of moeten we emoties ↴ en morele intuïtie inbouwen?

Welke trade-offs zijn we bereid te maken?

In Nederland worden steeds meer camera's en *trackers* ingezet die bewegingen volgen en vastleggen. Wanneer het gaat om het verbeteren van de veiligheid en leefbaarheid, blijken mensen bereid te zijn om de inzet van sensoren en de verzameling van sensor data te accepteren.
Maar ze stellen daar wel specifieke voorwaarden aan. Uit onderzoek van het Rathenau Instituut in 2019 blijkt dat de acceptatie voornamelijk afhangt van de context. Mensen zijn op voorhand niet tegen bijvoorbeeld *bodycams* of *wifitrackers*, het ligt eraan wanneer en in welke situatie ze worden ingezet. Deze inzichten worden ondersteund door onderzoek van de Europese Commissie in 2020. Uit het onderzoek blijkt dat 59% van de respondenten bereid is om een deel van hun persoonlijke informatie veilig te delen om de openbare dienstverlening te verbeteren. Met name wanneer het gaat om het verbeteren van medisch onderzoek en medische zorg (42%), het verbeteren van de reactie op crisis (31%) of het verbeteren van het openbaar vervoer en het verminderen van luchtverontreiniging (26%).

Uit de analyses van Rathenau blijkt dat er twee factoren cruciaal zijn, namelijk de mate van veiligheid die burgers ervaren en het type leefomgeving waarbinnen sensortechnologie wordt toegepast. In situaties waarin burgers zich onveilig voelen blijken ze de inzet van sensoren eerder te accepteren dan in situaties waarin zij zich veilig voelen. Ook het type leefomgeving is hierbij van belang. De inzet van sensoren wordt in mindere mate geaccepteerd in de privéruimte, dan in openbare ruimte waar zich veel mensen bevinden. In onveilige situaties in drukke openbare ruimtes blijkt de inzet van sensoren

voor de verbetering van de veiligheid en leefbaarheid dus gewenst, maar in thuissituaties of in rustigere openbare ruimtes waar men zich veilig voelt niet. Opvallend is dat burgers hierbij niet alleen kijken naar een afweging tussen veiligheid en privacy. Er worden meerdere waarden in overweging genomen, zoals democratische rechten, transparantie, efficiëntie en menselijk contact.

Ethische principes en richtlijnen kunnen daarom niet een op een in de praktijk worden overgenomen. Elke specifieke situatie vraagt om specifieke afwegingen. Steeds moet in de context bepaald worden welke waarden met elkaar schuren en wat hierin acceptabel is. Een afweging die in bepaalde situaties wordt geaccepteerd, kan in andere situaties volledig onacceptabel zijn. We zullen samen met elkaar moeten bepalen wat we in welke situatie kunnen en willen accepteren.

» Sometimes we think that technology will inevitably erode privacy, but ultimately humans, not technology, make that choice. «

-- Hu Yong, Peking University

Uitdagingen in de praktijk

Het formuleren van ethische ↴ principes en richtlijnen vormt een belangrijke eerste stap bij de realisatie van ethisch verantwoorde AI-toepassingen ↴. Het is echter niet eenvoudig om deze richtlijnen te vertalen naar de praktijk. Wanneer het bijvoorbeeld over transparantie ↴ gaat, dan is de betekenis ervan onder andere afhankelijk van het domein en de omgeving waarin de AI-toepassing wordt ingezet. Wanneer bijvoorbeeld Spotify mij een liedje aanbeveelt dat me niet aanstaat kan ik daar zonder uitleg prima mee leven, maar wanneer ik door een algoritme ↴ word afgewezen tijdens een sollicitatieproces wil ik toch wel weten op basis van welke criteria ik precies ben afgewezen. Daarbij kunnen er in de praktijk allerlei waardenconflicten ontstaan. Denk bijvoorbeeld aan transparantie en privacy ↴. Hoe waarborgen we onze privacy wanneer we willen inzetten op transparantie? In veel gevallen is het niet precies duidelijke welke ethische kwesties er precies spelen bij de ontwikkeling van AI-toepassingen.

Gelukkig zijn er steeds meer instanties, zowel nationaal als internationaal, die verschillende tools hebben ontwikkeld die kunnen helpen bij het identificeren van ethische dilemma's in de praktijk. Aan de hand van deze tools kan in kaart worden gebracht wat de ethische implicaties zijn van het toepassen van AI in de praktijk en welke ethische vraagstukken hierbij een rol spelen. Denk hierbij bijvoorbeeld aan:

> The Algorithmic Impact Assessment (AIA)
van het *AI Now Institute*
> De Ethische Data Assistent (DEDA)
van de Utrecht Data School
> Artificial Intelligence Impact Assessment (AIIA)
van het ECP
> Toolbox voor Ethisch Verantwoorde Innovatie
van het ministerie van Binnenlandse Zaken & Koninkrijkrelaties

Aan de hand van verschillende vragen worden organisaties geholpen om scherper te krijgen welke ethische kwesties er bij hun AI-project spelen en hoe ze daarmee om willen gaan. Denk hierbij aan vragen zoals ‘Zijn er persoonsgegevens gebruikt in het project?’, ‘Zijn alle verschillende groepen burgers vertegenwoordigd in de dataset(s)?’ en ‘Wie missen er of zijn niet zichtbaar?’. Zo kan de mogelijke vooringenomenheid van de toepassing worden blootgelegd en kunnen *biases* ↴ voorkomen worden. Het helpt organisaties tevens om dergelijke afwegingen goed te documenteren, zodat het proces transparanter wordt en ze verantwoording kunnen afleggen aan hun stakeholders. Een uitdaging is echter dat dergelijke richtlijnen zich vaak moeilijk eenduidig laten uitdrukken en lastig te kwantificeren zijn.

» There’s no such thing as a single set of ethical principles that can be rationally justified in a way that every rational being will agree to. «
-- Tom Chatfield, Tech philosopher

De eenduidigheid van richtlijnen

We willen uiteraard allemaal dat bij de inzet van AI-systemen mensen eerlijk behandeld worden en niet worden benadeeld op basis van bijvoorbeeld hun geslacht of afkomst. *Fairness* ↴ is dan ook een veelgebruikt principe in ethische richtlijnen en assessment tools. Toch is het niet eenvoudig om te bepalen wat dan precies ‘*fair*’ is. Dit vraagstuk houdt filosofen al een paar honderd jaar bezig. Er bestaat geen eenduidig beeld van hoe de samenleving eruit zou zien wanneer er geen oneerlijkheid meer zou bestaan. Is een samenleving waarin iedereen exact hetzelfde behandeld wordt überhaupt wel eerlijk? Door de komst van AI krijgt dit vraagstuk een nieuwe dimensie. Het concept van eerlijkheid moet namelijk in wiskundige ↴ termen worden uitgedrukt. Denk bijvoorbeeld aan het gebruik van AI in het rechtssysteem.

Door de inzet van *predictive policing* kan crimineel gedrag worden voorspeld door middel van grootschalige monitoring en data analyses. Er bestaat echter altijd het risico dat mensen zonder de gestelde criteria toch positief scoren (*false-positives)* en dat mensen die wel aan de gestelde criteria voldoen toch negatief scoren (*false-negatives*). Wat is in dit geval *fair*? Zet je mensen mogelijk onterecht vast, of riskeer je dat ze een misdaad begaan?

In Amerika is veelvuldig gebruik gemaakt van software om te voorspellen hoe groot de kans is dat een veroordeelde opnieuw de fout in gaat. Uit onderzoek van *ProPublica* in 2016 blijkt dat deze software bevooroordeeld is in het nadeel van mensen met een donkere huidskleur. Dat dit oneerlijk is, is overduidelijk. Maar het is niet eenvoudig om te bepalen wat eerlijk is en waaraan je dit wilt toetsen. Wordt eerlijkheid gedefinieerd door het gebruik van dezelfde variabelen of door dezelfde statistieken na het gebruik van verschillende variabelen? Betekent eerlijkheid dat hetzelfde percentage zwarte en witte individuen hoge risicobeoordelingsscores behalen? Of dat hetzelfde risiconiveau zou moeten resulteren in dezelfde score, ongeacht het ras? De vraag is dus of eerlijkheid gaat over een gelijke behandeling (met het risico op een ongelijke uitslag) of over een gelijke uitslag (met een mogelijk ongelijke behandeling). Uit onderzoek blijkt dat verschillende wiskundige definities van eerlijkheid elkaar wederzijds uitsluiten (Selbst et al., 2019). Het is dus onmogelijk om beide definities tegelijkertijd te vervullen, waardoor er op een gegeven moment een keuze gemaakt moet worden.

Deze keuze kan echter niet zonder meer eenduidig worden gemaakt. Wat onder eerlijkheid wordt verstaan is afhankelijk van het toepassingsdomein. Je kunt het systeem dat voor een eerlijke rechtspraak wordt gebruikt, niet zomaar hanteren voor eerlijke sollicitatieprocedures. Toch is de opvatting vaak dat

een krachtig systeem in meerdere domeinen toegepast kan worden. Verschillende culturen en communities hebben daarbij verschillende opvattingen over eerlijkheid. Er gelden niet alleen andere normen, er zijn ook andere wetten van toepassing. Onze opvattingen over goed en fout kunnen daarbij in de loop van tijd veranderen. Dat maakt het lastig en misschien zelfs wel onwenselijk om vooraf te bepalen hoe een AI-systeem moet handelen.

» By fixing the answer, you're solving a problem that looks very different than how society tends to think about these issues. «

-- Andrew Selbst, Data & Society Research Institute

De meetbaarheid van richtlijnen

De verschillende ethische principes en richtlijnen kennen verschillende abstractieniveaus. Eindwaarden en instrumentele waarden lijken hierbij door elkaar te lopen. Zo zijn *Societal wellbeing* en *Safety* eindwaarden waar we als samenleving naar streven en zijn *Accountability* ↴ en *Transparency* instrumentele waarden die we kunnen inzetten om deze eindwaarden te bereiken. Sommige richtlijnen laten zich daarbij eenvoudiger kwantificeren dan andere. De mate van bijvoorbeeld accuratesse kun je meetbaar maken, maar voor transparantie is dat veel lastiger. Wanneer is iets bijvoorbeeld 'transparant genoeg'? Bij 70 procent transparantie? En wat houdt dat dan precies in?

Het is daarbij de vraag of 100 procent transparantie überhaupt het nastreven waard is. Uit onderzoek van Microsoft Research in 2018 blijkt dat te veel transparantie kan leiden tot een *overload* aan informatie. Transparante modellen blijken het juist moeilijker te maken om de fouten van het model te detecteren en te corrigeren. Er bestaat daarbij het risico dat transparante modellen onterecht vertrouwd worden. Uit een opvolgend onderzoek van Microsoft Research in samenwerking met de University of Michigan in 2020 blijkt dat door het gebruik van visualisaties over de trainingsresultaten van *machine learning-tools* ↴ er een misplaatst vertrouwen ontstaat over de toepassingsmogelijkheden van de modellen. Zelfs wanneer de gegevens waren gemanipuleerd en de uitleg niet overeenkwam met de werkelijkheid.

De motieven achter de richtlijnen

In 2019 introduceerde Google de *Advanced Technology External Advisory Council* (ATEAC). Een externe ethische raad die erop toe zou moeten zien dat het bedrijf zich zou houden aan de zelf opgestelde richtlijnen voor ethisch verantwoorde AI-toepassingen. De *ethical board* werd echter al na een week weer opgeheven. Direct na de bekendmaking van de samenstelling van de adviesraad ontstonden er hevige discussies, met name over de aanstelling van Kay Coles James. De president van de Heritage Foundation staat bekend om haar conservatieve opvattingen over onder andere de rechten voor de LHBTI-gemeenschap.

Het is daarbij überhaupt de vraag wat de motieven voor Google waren om een ethische adviesraad in het leven te roepen. Volgens experts zijn de ethische richtlijnen en adviesraden van Google en andere commerciële organisaties voornamelijk gericht op het omzeilen van overheidsregulaties. Dit wordt ook wel '*ethics washing*' genoemd. Het zou een manier zijn om kritiek af te wenden, niet om daadwerkelijk deugdelijk te handelen. Aangezien de adviesraden geen feitelijke macht hebben is het voor organisaties niet nodig om hun gedrag ook daadwerkelijk aan te passen. Het lijkt geen toeval dat Google de adviesraad juist introduceerde kort na een periode waarin ze flink onder druk stonden. Google werkte in die periode onder andere voor de Chinese overheid aan *Project Dragonfly*. Een zoekmachine die resultaten blokkeert die door de Chinese overheid als onwenselijk worden beschouwd. Volgens Amnesty International zou de gecensureerde zoekmachine het recht

op vrije meningsuiting en de privacy van miljoenen Chinezen bedreigen. Later kwamen ook de medewerkers van Google in opstand en schreven een open brief aan het management van Google. Google kondigde in 2018 aan de werkzaamheden stop te zetten, maar medewerkers betwijfelen of dit ook echt gebeurd is. De adviesraad lijkt voor Google vooral een manier te zijn geweest om te zeggen 'Kijk, wij doen er in ieder geval alles aan'.

Het lijkt geen toeval te zijn dat ook andere tech-organisaties ethische richtlijnen en adviesraden lanceerden in een periode waarin verschillende misstanden in de tech-branche aan het licht kwamen, zoals het Cambridge Analytica schandaal in 2018. Zo richtte Microsoft de *AI ethics committee* op en deed uitgebreid onderzoek naar de transparantie van AI-systemen. Amazon is sponsor van een onderzoeks-programma ten behoeve van '*fairness in artificial intelligence*' en Facebook heeft geïnvesteerd in een '*AI ethics research center*' in Duitsland.

» Ethics boards and charters aren't changing how companies operate. «

-- James Vincent, The Verge

De uitdaging is om te komen tot bindende richtlijnen. Volgens verschillende experts is wetgeving nodig om ervoor te zorgen dat ethische richtlijnen worden nageleefd. De eerste stap in deze richting wordt beschreven in de '*White Paper on AI*' die in 2020 door de Europese Commissie is gepresenteerd. In deze white paper zet de commissie voorstellen uiteen om de ontwikkeling van AI in Europa te bevorderen, met inachtneming van de Europese grondrechten. Een belangrijk onderdeel van dit rapport is het voorstel om een '*prior conformity assessment*' te ontwikkelen voor risicovolle AI-toepassingen, gebaseerd op de ethische richtlijnen van de *High Level Expert Group*. Dit wettelijk kader moet de risico's voor de grondrechten en veiligheid aanpakken.

Betrouwbare AI-systemen kunnen hiermee een keurmerk verkrijgen, zodat het duidelijk is voor gebruikers welke systemen ze kunnen vertrouwen.

Hoewel het erg belangrijk is om ervoor te zorgen dat de ethische richtlijnen in de praktijk worden toegepast en AI-toepassingen de Europese wetten en grondrechten respecteren, biedt het onvoldoende handvatten om de richtlijnen daadwerkelijk in het ontwerpproces te integreren. De huidige beschikbare checklists en assessment tools zijn hiervoor onvoldoende kwantificeerbaar. Nu kan elk aspect van de lijst 'afgevinkt' worden zonder er volledig aan te voldoen. Ethische richtlijnen en assessments zijn dus voornamelijk instrumenten om te evalueren óf de AI-toepassing voldoet of zal voldoen, niet hóe de vereisten geïntegreerd moeten worden in het ontwerp zelf (en daarmee de evaluatie kan doorstaan). Hoewel veel richtlijnen en assessments verschillende checklists en vragenlijsten bieden, geven ze geen antwoord op de vraag hoe AI-systemen ethisch verantwoord kunnen handelen.

» Despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. «

-- Effy Vayena, ETH Zurich

Ethics boards and charters aren’t changing how companies operate.

-- James Vincent, The Verge

Compromissen rond betrouwbaarheid van AI

Door Maarten Stol, Principal Scientific Advisor, BrainCreators

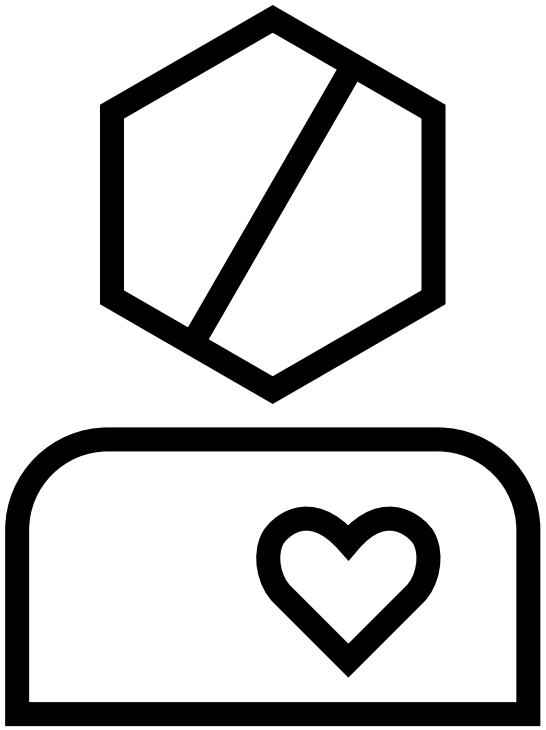
In 2019 publiceerde de Europese Commissie de *'Ethics Guidelines for Trustworthy AI'*. Hierin staat een lijst van zeven eisen waaraan een AI systeem moet voldoen om 'betrouwbaar' genoemd te mogen worden centraal. Tussen aanhalingstekens, want ook al heeft iedereen een gevoel over wat betrouwbaarheid is, in het geval van geavanceerde technische systemen is de definitie moeilijk om exact te maken. Er zijn vaker zulke lijsten gemaakt (denk aan de *Asilomar AI Principles*), maar de Europese richtlijnen proberen het begrip 'betrouwbaar' zo volledig mogelijk te dekken en tegelijk de lijst overzichtelijk te houden. Dit zijn ze:

1. *Human agency and oversight*
2. *Technical robustness and safety*
3. *Privacy and data governance*
4. *Transparency*
5. *Diversity, non-discrimination and fairness*
6. *Societal and environmental well-being*
7. *Accountability*

De *guidelines* zijn er echter ook duidelijk in dat er fundamentele spanningen kunnen bestaan tussen sommige van deze eisen. En juist hierin zit de complexiteit van het vraagstuk. Even voorbijgaand aan de vraag hoe elk van de eisen apart zou moeten worden bereikt (technisch makkelijker gezegd dan gedaan), wil ik hier graag ingaan op de mogelijke interacties en trade-offs.

Neem als eerste voorbeeld *Safety vs. Accountability.* Het gedrag van huidige AI systemen wordt grotendeels bepaald door data, meer dan door programmacode alleen. Het doel van *Machine Learning* is te pogen uit deze data automatisch een programma te leren dat beslissingen kan nemen in de algemene trend van de input data. Machine Learning is echter statistisch van aard en de beslissingen zullen daarom nooit echt 100% correct zijn. Lees: niet geheel veilig. Om toch zo dicht mogelijk bij 100% te komen worden steeds complexere systemen ontwikkeld. Door de aard van de techniek is het gevolg van deze toenemende complexiteit dat zulke systemen het karakter van een black-box krijgen. Dat wil zeggen dat, ook al zijn de beslissingen van goede kwaliteit, de herkomst van elke individuele beslissing minder toegankelijk wordt. Verklaarbaarheid wordt opgeofferd voor veiligheid. *Safety vs. Accountability* is een trade-off die met de huidige techniek nooit helemaal zal kunnen worden vermeden.

Er zijn meer van zulke trade-offs. *Privacy vs. Transparency:* hoeveel moet de burger aan privacy opgeven om data te ontsluiten die AI-systemen inzichtelijk moet houden voor andere burgers? *Human agency vs. Societal well-being:* welke belangen wegen zwaarder, die van het individu of het collectief? *Technical robustness vs. Environmental well-being:* hoeveel energie mag het continu trainen en onderhouden van AI-systemen gaan verbruiken? Et cetera. Het lijkt daarom redelijk om aan te nemen dat we in de toekomst niet altijd alle pijlers van de guidelines tegelijk zullen kunnen waarmaken. In plaats daarvan zal de praktijk vooral beheerst worden door de noodzaak van het sluiten van compromissen. Hierin is een taak weggelegd voor zowel de industrie als de overheid. De industrie zal een bepaalde mate van verantwoordelijkheid op zich moeten nemen om de guidelines na te streven bij het uitrollen van AI-systemen in de maatschappij. En de overheid zal zich verder moeten buigen over wetgeving die deze ontwikkelingen in banen moet gaan leiden. Een ding staat vast, op het gebied van een ethisch gebruik van AI zal de toekomst niet zonder compromissen zijn.



3 Ethiek in AI-ontwerp

Pagina's: 40
Woorden: 7802
Leestijd: ca. 1 uur

3. ----------

Ethiek in AI-ontwerp

Van overheden en maatschappelijke organisaties tot wetenschappelijke instanties en de private sector. Al deze partijen vinden het – om uiteenlopende redenen – van belang dat ethisch ↴ verantwoorde AI-toepassingen ↴ ontwikkeld worden. Ethische principes en richtlijnen ↴ schieten dan ook als paddenstoelen uit de grond. Verschillende richtlijnen schrijven onder andere voor dat AI-systemen veilig en robuust moeten zijn, onze privacy ↴ moeten waarborgen en ons rechtvaardig ↴ moeten behandelen. De vraag hierbij is in hoeverre dergelijke doelen in de praktijk tegen elkaar schuren. Kunnen we AI-systemen ontwikkelen die zowel accuraat als uitlegbaar zijn? Zijn uitlegbare ↴ systemen ook per definitie eerlijk? Kunnen we eerlijkheid ↴ vertalen naar wiskundige ↴ termen zonder mensen oneerlijk te behandelen? En zijn we het er überhaupt wel over eens hoe een samenleving eruit moet zien zonder oneerlijkheid?

Ethische richtlijnen – en dan met name de interpretatie ervan – zijn in hoge mate contextafhankelijk en subjectief. In dat opzicht beschrijven ethische richtlijnen eerder de moraal dan de ethiek. Ethiek moet gaan over de afstemming van de verschillende toepassingen met die moraal. Vanuit dit perspectief is ethiek dus veel meer een designvraagstuk, dan een verzameling van opvattingen over wat we belangrijk vinden. Als we in de toekomst ethisch verantwoorde AI-toepassingen willen gebruiken, dan zullen we ons nu moeten gaan bezighouden met de vraag hoe we toepassingen kunnen ontwikkelen die in overeenstemming zijn met de moraal. Het is hierbij niet voldoende om alleen te evalueren of een AI-toepassing aan de richtlijnen voldoet. Ethische principes en richtlijnen moeten daadwerkelijk geïntegreerd worden in het ontwerp. En dus ook in het ontwerpproces. Dit vraagt om een andere benadering van ethiek in de ontwikkeling van AI.

3.1 ----------

Een nieuwe kijk op ethiek

Voorheen ging ethiek ↴ voornamelijk over het menselijk handelen. Met de komst van AI ↴ is er echter een nieuwe speler in het spel, namelijk de zelflerende technologie. Naarmate de technologie autonomer wordt, meer beslissingen overneemt en de beslisregels lastiger te herleiden zijn, ontstaan er nieuwe ethische vraagstukken. De vraag hierbij is in hoeverre de huidige ethische begrippen nog voldoen. Hoe verhouden mens en technologie zich tot elkaar? En hoe gaat deze verhouding zich ontwikkelen?

Een verruiming van het begrip ethiek

Volgens Peter-Paul Verbeek, hoogleraar Filosofie van Mens en Techniek bij de Universiteit Twente, moeten de begrippen van ethiek worden verruimd. Om aan ethiek te kunnen doen is het volgens de gangbare overtuigingen nodig om intenties te hebben en in vrijheid te kunnen handelen. Volgens Verbeek (2011) bezitten technologieën die mede vormgeven aan morele beslissingen geen intentionaliteit ↴ en zijn mensen die in hun morele beslissingen gestuurd worden door technologieën niet vrij. Op de vraag of technologie intentionaliteit bezit lopen de opvattingen sterk uiteen, maar dat wij niet langer volledig autonoom zijn en ons leven wordt beïnvloed door technologie is overduidelijk.

Achter de huidige ethiek zit dus vaak onterecht het beeld van een compleet autonome mens. Het is niet de vraag óf we volledig uitlegbare systemen willen (dat station lijkt immers al gepasseerd), maar de vraag hóe we systemen kunnen toepassen die niet meer volledig uitlegbaar zijn. We zouden ons dus meer moeten bezighouden met het in kaart brengen van de impact van deze systemen en een constructief plan moeten maken over hoe we daarmee kunnen en willen omgaan. In plaats van ons enkel te focussen op de vraag wat goed of fout is, moeten we een concreet ethisch raamwerk ontwikkelen waarmee we kunnen omgaan met fouten van AI-systemen. Dit kunnen we niet alleen theoretisch benaderen, maar

moeten we in de praktijk ondervinden. We moeten hier in gecontroleerde omgevingen mee experimenteren. Van *human in the loop* (HITL) naar *human on the loop* (HOTL). Met AI kunnen we juist meer controle krijgen over de ongelijkheid op het gebied van bijvoorbeeld geslacht of ras. De condities die voor *biases* ↴ zorgen, kunnen we bij AI veel beter sturen dan bij mensen. Zo kunnen we zorgen voor voldoende diversiteit in de datasets én in de achtergrond van de programmeurs die de algoritmen ↴ ontwikkelen en de wetgevers die hierop toezien. We moeten streven naar systemen die oneerlijkheid eerlijker spreiden. AI kan onze fouten misschien niet helemaal oplossen, maar wel beter verdelen.

» De ergste vorm van ongelijkheid is proberen ongelijke dingen gelijk te maken. «

-- Aristoteles

Ethiek als leidmotief

In veel ethische discussies gaat het over de vraag wanneer het wel en niet acceptabel is dat AI wordt ingezet. Er wordt hierbij geprobeerd om harde grenzen en regels te formuleren waarmee de technologie ‘in toom gehouden kan worden’. Hierbij wordt echter gesuggereerd dat de samenleving en de technologie volledig van elkaar te scheiden zijn. Maar het is in de praktijk niet zo zwart-wit. Mens en technologie beïnvloeden elkaar wederzijds; wij geven de technologie vorm en de technologie geeft ons vorm. Naast een verruiming van het begrip, vraagt dit om een andere benadering van ethiek.

Begeleidingsethiek

ECP, het Platform voor de InformatieSamenleving, heeft in 2019 in samenwerking met Peter-Paul Verbeek een rapport over ‘begeleidingsethiek’ gepubliceerd. In plaats van de vraag hoe we AI kunnen beoordelen, zou de focus volgens deze aanpak moeten liggen op de vraag hoe we de implementatie van AI het beste kunnen begeleiden in de samenleving en hoe we er op een verantwoorde manier mee om kunnen gaan. In deze aanpak ligt de focus op de ontwikkeling van technologie met handelingsperspectief.

» In plaats van ethiek te zien als ‘beoordelen’ zou ze ook gezien kunnen worden als het normatief ‘begeleiden’ van technologie in de samenleving. «

-- Peter-Paul Verbeek & Daniël Tijink, ECP

Er wordt hierbij onderscheid gemaakt tussen drie soorten handelingen. Bij de eerste handeling (*ethics by design*) gaat het voornamelijk om het ontwerp van de technologie zelf. Waarden als privacy ↴ moeten meegenomen worden in het ontwerp. Zo moeten bijvoorbeeld medische diagnoses met behulp van AI gemaakt kunnen worden, zonder dat de privacy van de patiënt in het gedrang komt. AI-systemen zouden bijvoorbeeld getraind kunnen worden met gegevens van verschillende ziekenhuizen, zonder dat die gegevens ooit het pand van een ziekenhuis verlaten of de servers van een technologiebedrijf aanraken. Bij de tweede handeling (*ethics in context*) gaat het over context-specifieke afspraken. De introductie van een nieuwe technologie gaat vaak gepaard met veranderingen in de omgeving. Dat is echter niet altijd zichtbaar, denk bijvoorbeeld aan de inzet van gezichtsherkenningssoftware. Daarom moet het voor iedereen duidelijk zijn dat er gebruik wordt gemaakt van AI, welke gegevens worden verzameld, wie er toegang hebben tot de data, wat de mogelijkheden zijn om de beslissingen van een AI-systeem aan te vechten, et cetera. Bij de laatste handeling (*ethics by user*) staat het gebruik van de technologie centraal. Hierbij is het van belang dat iedereen voldoende kennis heeft om kritisch en verantwoord om te kunnen gaan met de technologie. Zowel de mensen die het ontwikkelen als de eindgebruikers. Zo zou er bij het gebruik van zelfrijdende auto’s nog steeds een rijbewijs kunnen bestaan, zodat nieuwe skills worden aangeleerd die bijvoorbeeld nodig zijn voor de communicatie met autonome systemen of wanneer er moet worden ingegrepen als het systeem daar om vraagt.

Deze aanpak zorgt voor een meer holistische benadering van ethiek. Zowel het gebruik en de gebruikers als de technologie worden in deze aanpak meegenomen. Het doet daarmee recht aan de dynamische praktijk van AI: de afstemming tussen mens en technologie is een continu proces. Het is niet óf de mens óf de technologie die alle beslissingen neemt. Ze geven elkaar wederzijds vorm. De beperking van deze aanpak is echter dat er wordt gesuggereerd dat de technologie zich hoe dan ook zal ontwikkelen en dat we ons er dus maar beter aan kunnen toevertrouwen. Maar de vraag óf we AI in bepaalde contexten moeten gaan inzetten blijft onverminderd relevant. We moeten AI niet zien als een doel op zichzelf, maar als een middel om een bepaald doel te bereiken. De vraag of AI dan wel het beste middel is om dat doel te bereiken mag niet worden overgeslagen. We moeten onszelf dus blijven afvragen wat voor samenleving we willen zijn, gegeven alle technologische ontwikkelingen.

» AI is an Ideology, Not a Technology. «

-- Jaron Lanier & Glen Weyl, Wired

Lonkend perspectief

Wanneer het over de toekomst van AI gaat wordt vaak gesuggereerd dat ethiek een rem is op innovatie en dat bijvoorbeeld Europa een keuze moet maken: of strenge regulering waardoor innovatie wordt belemmerd óf innoveren en aansluiting vinden bij Amerika en China. Volgens een rapport van PwC Nederland uit 2020 bestaat er een trade-off tussen strakke regelgeving en snelle innovatie. Met de '*White Paper on AI*', waarin de Europese Commissie de ambitie uitspreekt om de ontwikkeling van AI te versnellen en tegelijkertijd strengere regulering aankondigt, zou Europa zowel op het gaspedaal, als op de rem drukken.

» Wie strenge regelgeving invoert, accepteert daarmee dat de ontwikkeling van AI aanvankelijk langzamer zal verlopen. «

-- Mona de Boer, PwC Nederland

Er is echter ook een middenweg. Denk bijvoorbeeld aan de tuinbouw. Juist door regulaties zijn Nederlandse kassenbouwers enorm innovatief geworden. Zo innovatief zelfs, dat hun kassen als de meest duurzame van de wereld worden gezien. Nederlandse kassenbouwers zijn tegenwoordig dan ook in toenemende mate in het buitenland actief en dragen bij aan de realisatie van duurzame tuinbouwprojecten over de hele wereld. Ethiek hoeft dus geen anker te zijn dat je meesleept, maar kan juist een lonkend perspectief bieden voor innovatie.

Waarden verenigen

Trade-offs worden nog te vaak afgespiegeld als geforceerde binaire keuzes. Wil je dat de overheid alles van je weet, of wil je onveilig leven? Ook bij de inzet van *tracking-apps* tijdens de coronacrisis in 2020 leek dat het devies. Overal doken krantenkoppen en artikelen op over de trade-off tussen privacy en de volksgezondheid. Hierbij lijkt de focus gericht te zijn op het individu versus het collectief. Je wordt geacht je persoonlijke privacy op te geven voor het grote geheel. Er wordt hierbij echter ten onrechte vanuit gegaan dat dergelijke applicaties alleen werken als privacy volledig wordt opgegeven. Dat is niet het geval. De volksgezondheid kan worden gewaarborgd mét inachtneming van de privacy. Het is dan ook niet voor niets dat de Nederlandse overheid tijdens de nationale '*appathon*' in 2020 heeft besloten de zeven ingebrachte apps niet te gebruiken, omdat ze niet voldeden aan de privacyrichtlijnen. Ook Apple en Google kondigden aan dat apps die gebruikmaken van locatiegegevens van gebruikers geen toegang krijgen tot de besturingssystemen van Apple en Google.

-- Jaron Lanier & Glen Weyl, Wired

AI is an Ideology

not a
Technology

» Those who would give up essential Liberty, to purchase a little temporary Safety, deserve neither Liberty nor Safety. «

-- Benjamin Franklin

Value Sensitive Design

Waarden kunnen dus verenigd worden in het ontwerp. Dit vormt het uitgangspunt voor *Value Sensitive Design* (VSD). Het 'ontwerpen voor waarden' biedt een alternatief denkpatroon om met innovatie om te gaan. Volgens Jeroen van den Hoven, hoogleraar Ethiek & Technologie aan de TU Delft, moeten we innovatie inzetten om waarden te bedienen en waardenconflicten op te heffen. Op deze wijze kunnen we verantwoord innoveren met AI. Je praat pas over trade-offs als ze in de praktijk daadwerkelijk ervaren worden. De uitdaging is dus om deze situaties te voorkomen door middel van design. Hiervoor moeten we omgevingen creëren waarbij we niet hoeven te kiezen tussen verschillende waarden, maar waarbij we waarden kunnen maximaliseren in relatie tot elkaar. Ja er zijn keuzes, maar met een goed ontwerp kun je ervoor zorgen dat de keuzes niet schaden.

» Ethiek is in belangrijke mate een ontwerpdiscipline en gaat over het verantwoord vormgeven van onze samenleving en leefwereld. «

-- Jeroen van den Hoven, TU Delft

Ethics by Design

Als we in de toekomst ethisch ↴ verantwoorde AI-systemen ↴ willen gebruiken, dan redden we het niet met alleen richtlijnen ↴ en assessments. Er zijn handvatten nodig om dergelijke principes daadwerkelijk te integreren in het ontwerp. We moeten daarom de stap maken van evalueren naar integreren. Kortom, er is behoefte aan '*Ethics by Design*'. Ondanks het feit dat steeds meer ethici dit inzicht delen blijft het vaak bij de constatering dát het noodzakelijk is om waarden te integreren in AI-toepassingen. De vraag hóe dit precies in de praktijk moet gebeuren blijft grotendeels onbeantwoord. Zo worden voorbeelden aangehaald van AI-toepassingen die ethisch verantwoord tot stand zijn gekomen, maar dat zegt niet zo veel over het AI-systeem zelf. Denk bijvoorbeeld aan de Fairphone: een smartphone die zowel eerlijk is voor het milieu, als voor de mensen in het productieproces. Deze telefoon laat weliswaar zien dat waarden verenigd kunnen worden, maar geeft geen antwoord op de vraag of het besturingssysteem ervan ook in staat is om ethisch verantwoorde beslissingen te nemen.

Hetzelfde geldt voor de bestaande impact- en assessmenttools die gebruikt worden bij de beoordeling van AI. Deze gaan meer over de totstandkoming van het AI-systeem, dan over de handeling van het AI-systeem. De meeste assessmenttools zijn checklists waarbij voornamelijk vragen worden gesteld over het gebruik van de datasets. Is de data geanonimiseerd? Is de data divers genoeg? En is het proces transparant ↴? Ze geven vaak geen antwoord op de vraag op welke wijze een AI-systeem tot ethisch verantwoorde beslissingen kan komen. Om ethiek daadwerkelijk in de AI-praktijk te kunnen brengen moet worden gekeken naar de verschillende manieren waarop een AI-systeem kan leren wat wel en niet ethisch verantwoord is. De vraag is dus op welke wijze we systemen kunnen bouwen die in verschillende situaties ethisch verantwoord kunnen handelen.

Kun je ethische regels in het systeem programmeren? Moeten we het systeem ethische doelfuncties meegeven? Of kan het systeem zelf morele afwegingen maken? Om hier antwoord op te geven onderscheiden we drie verschillende systeembenaderingen, namelijk statisch leren, adaptief leren en intuïtief leren.

Statisch leren

Bij statisch leren worden ethische principes en regels in het intelligente systeem geprogrammeerd. Het doel van het AI-systeem is daarmee impliciet onderdeel van het algoritme ↴ en moet uiteindelijk door een programmeur worden ingevuld. Als we willen dat een autonoom voertuig ons zo snel mogelijk van A naar B vervoert, dan moeten ook de uitzonderingen in het algoritme zijn ingebed. We willen natuurlijk niet dat een voertuig de verkeersregels negeert en met 300 km per uur dwars door het weiland scheurt. Zo snel mogelijk betekent objectief gezien namelijk letterlijk zo snel mogelijk. Het algoritme moet tevens rekening houden met waarden als veiligheid. Deze benadering lijkt ook gehanteerd te worden door ethici en ontwikkelaars binnen de *Value Sensitive Design* (VSD) community. Het uitgangspunt hierbij is dat waarden, zoals veiligheid, in een zo vroeg mogelijk stadium van het ontwerpproces expliciet gemaakt moeten worden. Deze waarden kunnen vervolgens geformaliseerd worden en op deze wijze worden ingebed in het AI-systeem.

» Ethiek moet bij technologie al in het ontwerp zitten. «

-- Jeroen van den Hoven, TU Delft

Voordelen

Het grote voordeel van deze benadering is dat ethische principes vrij transparant zijn en goed te interpreteren zijn door mensen. We kunnen daardoor samen nadenken over wat we als samenleving belangrijk vinden en dit vastleggen in AI-systemen. Dit geeft een zekere mate van '*human control*'. Nog voordat AI-systemen op de markt komen kunnen we de ontwikkeling ervan controleren en kunnen we keurmerken uitgeven aan de toepassingen die voldoen aan de opgestelde ethische richtlijnen. Het is daarbij vrij helder wanneer bepaalde richtlijnen worden overtreden. En organisaties die de regels overtreden kunnen dan ter verantwoording worden geroepen. Dit biedt de mogelijkheid om toezichthoudende organen in positie te brengen en de ontwikkeling van AI te monitoren.

» Niet de ethici maar de ingenieurs staan aan de frontlinie van de ethiek. «

-- Peter-Paul Verbeek, Universiteit Twente

Nadelen

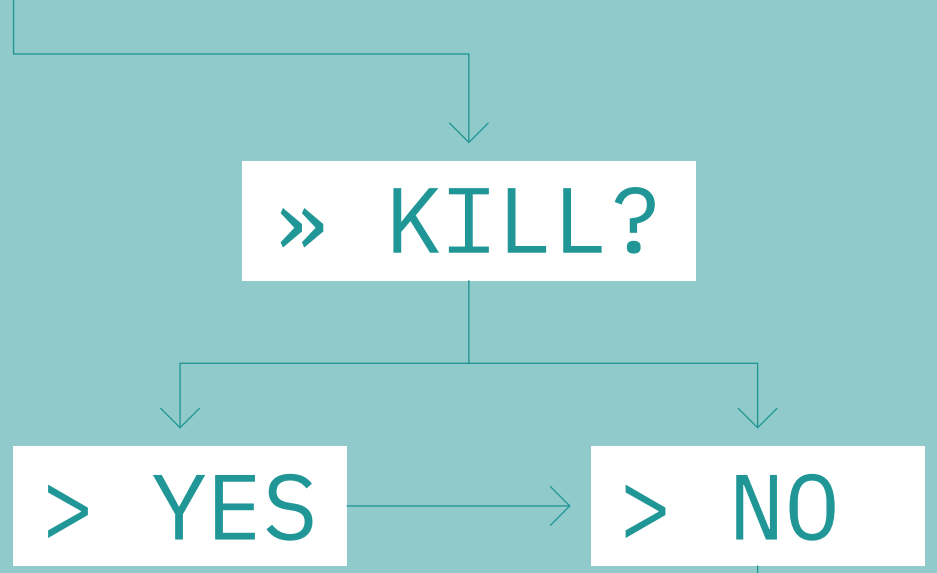
Er wordt binnen deze benadering echter onvoldoende rekening gehouden met uitzonderingen die zich in de praktijk kunnen voordoen. Deze aanpak vereist dat er regels zijn voor élke situatie die kan voorkomen. Dat is in praktijk vrijwel onmogelijk om te realiseren. Bovendien zijn er situaties waarin tegenstrijdige regels gelden. Zo mag je bijvoorbeeld niet door rood rijden. Maar wanneer een groep mensen ontweken kan worden moet een autonoom voertuig wel degelijk door rood rijden. Het is bijna onmogelijk om ál deze uitzonderingen vast te leggen. Daarnaast zijn niet alle gevolgen vooraf te overzien. Toen GPS-functionaliteiten voor de ruimtevaart werden ontwikkeld kon niemand voorzien dat deze functionaliteit uiteindelijk via onze telefoons in onze broekzak terecht zou komen. Met alle gevolgen van dien, zowel positief als negatief. Daarnaast krijgen AI-systemen updates: moet er dan per update opnieuw een keurmerk worden aangevraagd? Het is vrijwel onmogelijk om dit allemaal vooraf vast te leggen in ethische richtlijnen.

» Je kunt een bewakingsrobot meegeven dat hij geen mensen mag verwonden. Maar dat is een belemmering als die robot een aanslag moet voorkomen. «

-- Leon Kester, TNO

Je kunt een bewakingsrobot meegeven dat hij geen mensen mag verwonden.

-- Leon Kester, TNO



Maar dat is een belemmering als die robot een aanslag moet voorkomen.

Deze benadering legt daarnaast te veel verantwoordelijkheid bij de programmeur. De richtlijnen vertellen namelijk niet op welke wijze de waarden geformaliseerd moeten worden in wiskundige termen. De vraag wat precies *fair* is, hangt af van de context en de specifieke gebruikerstoepassing. Er bestaat daarbij altijd het risico dat bijvoorbeeld bij een medische test de uitslag van patiënten onterecht als positief of negatief wordt geclassificeerd. Dit zou ertoe leiden dat gezonde patiënten onterecht medicijnen krijgen toegediend en noodlijdende patiënten onterecht geen medicijnen krijgen. Het is niet eerlijk en daarnaast onverantwoord om dergelijke configuraties enkel aan de programmeur over te laten.

Adaptief leren

Bij adaptief leren worden de regels niet voor-geprogrammeerd, maar leert het systeem wat 'goed en slecht' is van menselijk gedrag. Hiervoor wordt het algoritme uitgerust met een doelfunctie, waarmee gespecificeerd kan worden waarop er door het algoritme geoptimaliseerd moet worden. Hierbij wordt een duidelijk onderscheid gemaakt tussen het probleem-oplossend vermogen van het intelligente systeem en de doelfunctie. Op deze manier kunnen specifieke toepassingsdoelen gecombineerd worden met ethische doelen. Zo kan een autonoom voertuig ons zo snel mogelijk van A naar B vervoeren (toepassingsdoel) én rekening houden met onze veiligheid (ethisch doel). Wanneer er in de praktijk tegenstrijdige regels gelden dan moet het AI-systeem in staat zijn om afwegingen te maken. Deze opvatting is populair bij onder andere *Open AI* en de *Future of Life Institute*. Volgens AI-pionier Stuart Russell kunnen AI-systemen alleen dergelijke afwegingen maken als de systemen in de praktijk leren wat menselijke waarden inhouden. In zijn TED-Talk uit 2017 zet hij hiervoor drie pijlers uiteen om veiligere AI-toepassingen te kunnen ontwikkelen:

1. Het enige doel van het AI-systeem is om de realisatie van menselijke waarden te maximaliseren;
2. Het is voor het AI-systeem aanvankelijk niet zeker wat die waarden precies zijn;
3. Menselijk gedrag biedt voor het AI-systeem informatie over menselijke waarden.

Oftewel, *learning on the job*. Hiervoor is het van belang dat machines alles leren over menselijke waarden, om te achterhalen wat echt belangrijk voor ons is.

Voordelen

Het voordeel van deze benadering is dat het AI-systeem in de context leert wat de juiste handeling is en daardoor om kan gaan met tegenstrijdige regels. Wanneer de regels vooraf geprogrammeerd worden is dat vrijwel onmogelijk. AI-systemen worden hierdoor veel flexibeler en beter inzetbaar. Deze benadering maakt het daarbij mogelijk dat het systeem leert van het menselijk gedrag, zonder slechte eigenschappen te kopiëren. Het systeem leert namelijk niet alleen van individuen (we doen allemaal weleens iets 'slechts'), maar van de samenleving als geheel ('slecht' gedrag wordt dan in perspectief geplaatst). Zo kan het systeem leren dat mensen bijvoorbeeld iets stelen, wanneer ze te weinig geld hebben om hun kinderen te laten studeren. In plaats van te leren dat stelen in zo'n situatie is toegestaan, zal het systeem proberen te helpen om de kinderen naar school te krijgen. Systemen zijn niet 'belast' met menselijke driften en emoties, zoals status en macht, die voortkomen uit een biologische evolutie.

» The robot does not have any objective of its own. It's purely altruistic. «

-- Stuart Russell, University of Berkeley

Nadelen

De uitdaging van deze benadering is dat een AI-systeem moet handelen naar menselijke waarden van de gehele samenleving, niet alleen van de gebruiker. Als een machine jouw welzijn vooropstelt, dan kan dat ten koste gaan van een ander. Het systeem moet op de een of andere manier de voorkeuren van veel verschillende mensen afwegen. In zijn Talk geeft Russell een aantal voorbeelden, waarbij deze benadering verkeerd kan uitpakken. Stel: je bent de verjaardag van je vrouw vergeten en je hebt een belangrijke afspraak gepland die je niet kunt afzeggen. Een AI-systeem kan dan voor jouw geluk besluiten het vliegtuig van je afspraak te laten vertragen, zodat je toch met je vrouw uit eten kunt gaan. Maar dat ontregelt het leven van anderen.
De redenatie kan ook andersom gelden. Stel dat je honger hebt en je vraagt een 'robotkok' om een broodje ham te bereiden, dan kan deze weigeren omdat er elders op de wereld mensen leven waar de noodzaak aan eten groter is. Het kan volgens Russell ook zo zijn dat er afwegingen gemaakt moeten worden tussen menselijke waarden.
Stel dat de robotkok toch besluit je broodje ham te gaan maken, dan kan het voorkomen dat er geen vlees in de koelkast ligt. Maar er loopt wel een kat rond in huis. Welke waarde weegt dan zwaarder: de behoefte aan eten of de sentimentele waarde aan een huisdier? Volgens de Piramide van Maslov wint eten dan.
Doordat we niet vooraf meegeven wat goed en slecht is geven we een groot deel van de controle over het systeem af. Wanneer een AI-toepassing onbedoelde gevolgen veroorzaakt is het erg lastig om in te grijpen.

» We had better be quite sure that the purpose we put into the machine is the purpose which we really desire. «

-- Norbert Wiener, 1960

Een andere uitdaging is dat veel van onze morele opvattingen impliciet ↴ zijn. We spreken ze niet letterlijk uit. Dat maakt het lastig voor een AI-systeem om te leren. Daarbij is het lastig voor computers om emoties goed in te schatten. Wanneer mensen lachen is het moeilijk om te bepalen of dit oprecht is, of dat er een andere emotie ↴ of een ander motief achter schuilgaat.

» Computers can't tell if you're happy when you smile. «

-- Angela Chen, MIT Technology Review

Intuïtief leren

Bij intuïtief leren worden elementen van statisch en adaptief leren gecombineerd. Het algoritme krijgt ook bij deze benadering een doelfunctie, maar het systeem leert hierbij niet van het gedrag van mensen. In plaats daarvan bepalen mensen hoeveel waarde ze aan bepaalde doelen hechten door middel van het toekennen van gewichtsfactoren. Het gewicht wordt bepaald aan de hand van het nut dat het doel heeft voor de samenleving. Er wordt daarom ook wel gesproken over een '*utility function*'. Dit stelt het systeem in staat om beredeneerde afwegingen te maken op basis van deze vooraf gewogen factoren. Wanneer een autonoom voertuig ons van A naar B moet vervoeren zijn er verschillende doelen relevant, zoals reistijd, comfort, verkeersveiligheid en duurzaamheid. Aan deze verschillende doelen worden verschillende gewichtsfactoren toegekend. Het autonome voertuig zal dan de doelfunctie gebruiken om te beslissen welke route en welk rijgedrag het beste rekening houdt met zowel de wensen van de passagier (comfort en aankomsttijd) als van de samenleving (veiligheid en milieu). Afhankelijk van het gewicht en de huidige staat van de omgeving (zoals de hoeveelheid verkeer op de weg) zijn verschillende uitkomsten mogelijk.

Computers can't tell if you're happy when you smile.

OK :)

-- Angela Chen, MIT Technology Review

Deze vorm van leren lijkt sterk op het menselijk besluitvormingsproces ↴; het is intuïtief. Het is voor mensen mogelijk om op basis van wetten en regels auto te rijden, omdat ze deze naar de specifieke context vertalen. Deze laatste stap kun je niet voorprogrammeren. Zo is de wettelijke maximumsnelheid op veel plekken in Nederland verlaagd naar 100 km per uur. Beter voor de verkeersveiligheid én beter voor het milieu is de gedachte. Hoewel mensen op deze plekken 100 km per uur mógen rijden, doen ze dat niet overal. De snelheid wordt continu afgestemd op de omgeving.

Voordelen

Het voordeel van deze benadering is dat de meerwaarde van statisch en adaptief leren wordt gecombineerd: het biedt de controle van de statische benadering en de flexibiliteit van de adaptieve benadering. Een AI-systeem krijgt op deze wijze de waarden mee die de maatschappij van belang vindt om mee te wegen in een advies of beslissing. Zo kan de rekenkracht van de computer benut worden om de best mogelijke uitkomst in elke situatie te berekenen. Op deze wijze blijven ethisch verantwoorde uitkomsten gewaarborgd, zonder dat we de controle volledig uit handen geven. Het is immers nog steeds aan de mens om gewicht te geven aan de factoren die meegenomen worden in de weging. Hiermee wordt ook afgestapt van de scherpe trade-off tussen goed óf slecht. In de realiteit doen er zich voortdurend situaties voor waarbij we moeten kiezen tussen twee ‘kwaden’. Welk kwaad zwaarder weegt hangt dan af van de context. Met deze benadering kan het systeem bepalen welke uitkomst dan het best is voor het individu en de samenleving.

» Een autonoom voertuig kan dan zo goed mogelijk kiezen tussen twee in principe ongewenste alternatieven. «

-- Leon Kester, TNO

Nadelen

Het nadeel van deze benadering is dat ervan uit wordt gegaan dat algoritmen in staat zijn om genuanceerde afwegingen te kunnen maken. Volgens Peter Eckersley, onderzoeksdirecteur van *The Partnership on AI*, zijn algoritmen ontworpen om een enkel wiskundig doel na te streven, zoals het minimaliseren van de kosten of het maximaliseren van het aantal gepakte fraudeurs. Wanneer er wordt geprobeerd om meerdere doelen gelijktijdig na te streven – waarvan sommige doelen met elkaar concurreren – stuit de ontwikkeling van AI-systemen op praktische en conceptuele problemen. Dit wordt ook wel de ‘*impossibility theorem*’ genoemd. Met name wanneer immateriële waarden, zoals vrijheid en welzijn, gemaximaliseerd moeten worden, bestaat er volgens Eckersley soms gewoonweg geen wiskundige oplossing. Ethiek lijkt meer te zijn dan een kosten-baten afweging. Het gaat ook om de minder tastbare begrippen, zoals empathie ↴, mededogen en respect. Eckersley beschrijft in een inmiddels ‘berucht’ artikel dat het onmogelijk is om formeel te specificeren wat een goede uitkomst is voor een samenleving, zonder menselijke ethische intuïties te schenden.

» Such systems should not use objective functions in the strict mathematical sense. «

-- Peter Eckersley, The Partnership on AI

Ondanks de kritiek lijkt intuïtief leren de enige manier om te kunnen omgaan met de complexiteit van ethische vraagstukken. Er is een systeem nodig dat het vermogen heeft om te kunnen redeneren met onzekerheid en een idee van 'zelf' heeft om met 'trolleyachtige' problemen om te kunnen gaan. Er moet hierbij zowel gekeken worden naar de waarde die de maatschappij hecht aan de handeling ↴ en de gevolgen van de handeling ↴, als naar de persoon (of het object) die de handeling uitvoert ↴. 'De ethiek' zou daarom niet letterlijk in het ontwerp moeten zitten, maar het systeem zelf zou in staat moeten zijn om ethische afwegingen te kunnen maken. We zouden daarom niet moeten spreken over '*Ethics by Design*' maar over '*Designing for Ethics*'. Momenteel is de technologie nog niet zo ver, maar er wordt hard gewerkt aan nieuwe invalshoeken. Denk bijvoorbeeld aan ontwikkelingen op het gebied van *hybrid AI*. Hierbinnen wordt aan systemen gewerkt die kunnen kijken (met behulp van onder andere neurale netwerken ↴) én kunnen redeneren (met behulp van onder andere formele logica ↴). Deze stroming wordt ook wel *Deep Reasoning* ↴ genoemd, een combinatie van *deep learning* ↴ en *symbolic reasoning*. Op deze wijze zijn systemen steeds beter in staat om intuïtief te leren.

> *» Deep Reasoning is the field of enabling machines to understand implicit relationships between different things. «*
>
> -- Adar Kahiri, Towards Data Science

De Ethische Scrum

Er zijn meerdere benaderingen om AI-systemen ↴ te laten leren over wat wel en niet ethisch ↴ verantwoord is. Zo kunnen ethische principes en richtlijnen ↴ worden geprogrammeerd in het systeem (statisch leren), kan het systeem leren van menselijk gedrag en menselijke waarden optimaliseren (adaptief leren) en kan het systeem afwegingen maken tussen meerdere doelen op basis van vooraf bepaalde gewichtsfactoren (intuïtief leren). In veel ethische discussies worden deze benaderingen onvoldoende meegenomen. De systeem-benadering bepaalt namelijk op welke wijze geformuleerd moet worden wat we belangrijk vinden en wat we wel en niet aan het systeem meegeven. Bij statisch leren gaat het om een volledige set aan regels en uitzonderingen, terwijl het bij adaptief en intuïtief leren gaat om het bepalen van de doelfuncties. Wanneer er voor intuïtief leren gekozen wordt moet daarbij niet alleen bepaald worden welke doelen we belangrijk vinden, maar ook hoe zwaar deze doelen in verschillende situaties ten opzichte van elkaar wegen.

Welke benadering er ook wordt gekozen, het is vrijwel onmogelijk om vooraf in kaart te brengen wat de mogelijke implicaties van verschillende ontwerpkeuzes zijn. De optimalisering van AI-systemen gebeurt voor-namelijk aan de hand van *trial and error*. Dit betekent dat gedurende het ontwikkelingsproces nieuwe uitdagingen zichtbaar worden. Denk bijvoorbeeld aan veiligheid. Door keuzes in het proces worden de aanvankelijke veiligheidsprincipes beïnvloed. Om kwetsbaarheden in het systeem te voorkomen moeten gedurende het proces afwegingen worden gemaakt en ontwerpcriteria worden bijgesteld. Het is daarom van belang om niet alleen naar het ontwerp te kijken, maar ook naar het ontwerpproces.

» Problemen zitten bij ethiek niet in de perspectieven, maar in de processen. «

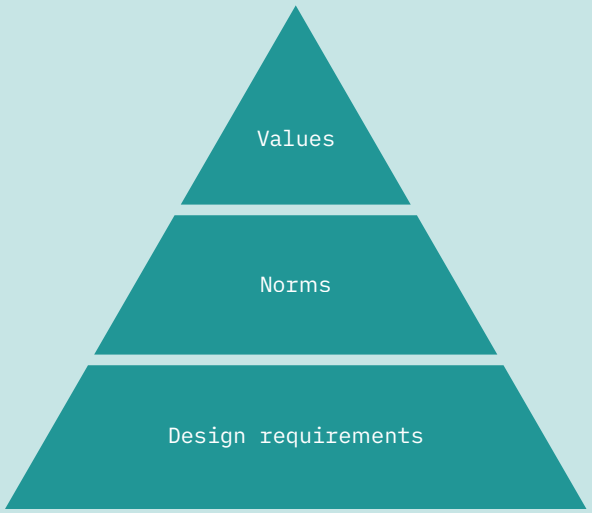
-- Robert de Snoo, Human & Tech Institute

Ethici vs. Technici

Wanneer het over AI en ethiek gaat dan zijn het vaak ethici die het over technologie hebben, of technici die het over ethiek hebben. In beide gevallen is het niet hun eigen vakgebied. Daarbij bestaat er vaak een groot onderscheid tussen de verschillende benaderwijzen. Technici benaderen het vraagstuk vooral vanuit optimalisatie. Hoe kan ik waarden als privacy ↴ en transparantie ↴ wiskundig ↴ benaderen en formaliseren? Ethiek is vanuit dat opzicht een probleem dat moet worden opgelost; hiervoor zijn concrete antwoorden nodig. Ethici daarentegen houden zich veel meer bezig met het bevragen van het vraagstuk zelf. Bestaande vraagstukken mogen nieuwe vragen oproepen. Dergelijke abstracte inzichten laten zich echter moeilijk vertalen naar de concrete gebruikerspraktijk.

Om tot ethisch verantwoorde AI-systemen te komen zijn beide zienswijzen nodig. Het is daarom van belang om een meer holistische aanpak te kiezen. Momenteel concurreren wetenschappers van verschillende disciplines nog te veel met elkaar, terwijl ze elkaar juist kunnen aanvullen. De ontwikkeling van AI gaat daarbij verder dan technologie en filosofie. De inzet van AI heeft impact op de gehele samenleving; op de manier waarop we samenwerken en samenleven. Ethici en technici zouden daarom meer moeten gaan samenwerken met onder andere sociologen en economen. Maar ook met biologen en psychologen. Het besluitvormingsproces van systemen lijkt in steeds hogere mate op menselijke besluitvormingsprocessen ↴. Hiervoor moeten we beter begrijpen hoe dergelijke processen in het menselijk brein werken en tot uiting komen in menselijk gedrag. Er is daarom een meer transdisciplinaire aanpak nodig.

Onderzoekers van verschillende Amerikaanse universiteiten en bedrijven hebben in 2019 een artikel gepubliceerd in *Nature* waarin ze oproepen voor een transdisciplinaire wetenschappelijke onderzoeksagenda. Het doel is om meer inzicht te krijgen in het gedrag van AI-systemen. Ontwikkelingen op het gebied van AI brengen machines met een vorm van '*agency*' steeds dichterbij. Oftewel, machines die zelfstandig acties ondernemen en autonome besluiten nemen. Machines vormen hiermee een nieuwe klasse van actoren in onze samenleving met hun eigen gedragingen en ecosystemen. Dit vraagt volgens de experts om de ontwikkeling van een nieuw wetenschappelijk onderzoeksgebied, namelijk die van *Machine Behavior*. Het uitgangspunt hierbij is dat we AI-systemen op dezelfde manier moeten bestuderen als de manier waarop we dieren en mensen bestuderen, namelijk door middel van empirische observatie en experimenten.



'Values hierarchy'
– Source: Van de Poel (2013)

Operationalisatie in de praktijk

De verschillende benaderwijzen van onder andere ethici en technici laten zich goed duiden aan de hand van de '*values hierarchy*'.

Bovenin de piramide staan de meer abstracte waarden waar veel ethici zich mee bezigen. Onderaan staan de concrete *design requirements* waar veel technici mee werken. Uiteindelijk moeten waarden geoperationaliseerd worden in de praktijk. Design requirements moeten namelijk meetbaar gemaakt worden om te kunnen hanteren voor en door AI-systemen. Normalisatie kan hierbij helpen om de kloof te dichten. Hiervoor moeten grofweg drie stappen in het ontwerpproces doorlopen worden:

1. *Conceptualisatie:* waarden moeten allereerst gedefinieerd worden. De betekenis moet duidelijk zijn en universeel toepasbaar zijn. Dit is wat veel ethische principes en richtlijnen doen. Ondanks de beperkingen die veel richtlijnen hebben, is het een belangrijke eerste stap die gezet moet worden.

2. *Specificatie:* de gedefinieerde waarden moeten vertaald worden naar de specifieke context. Waarden hebben in verschillende situaties namelijk andere betekenissen. Hierdoor ontstaan meer concrete normen die sturing kunnen geven aan het ontwerpproces.

3. *Operationalisatie:* de gespecificeerde normen moeten vervolgens vertaald worden naar meetbare design requirements. Op deze wijze kunnen verschillende ontwerpkeuzes ten opzichte van elkaar gewogen worden.

Deze meetbare requirements kunnen daarbij vertaald worden naar technische standaarden, waarop keurmerken en certificaten gebaseerd kunnen worden. De uitdaging hierbij is dat dergelijke standaarden internationaal geharmoniseerd worden. Op deze wijze kunnen technici concreter aan de slag met ethisch verantwoorde AI-toepassingen.

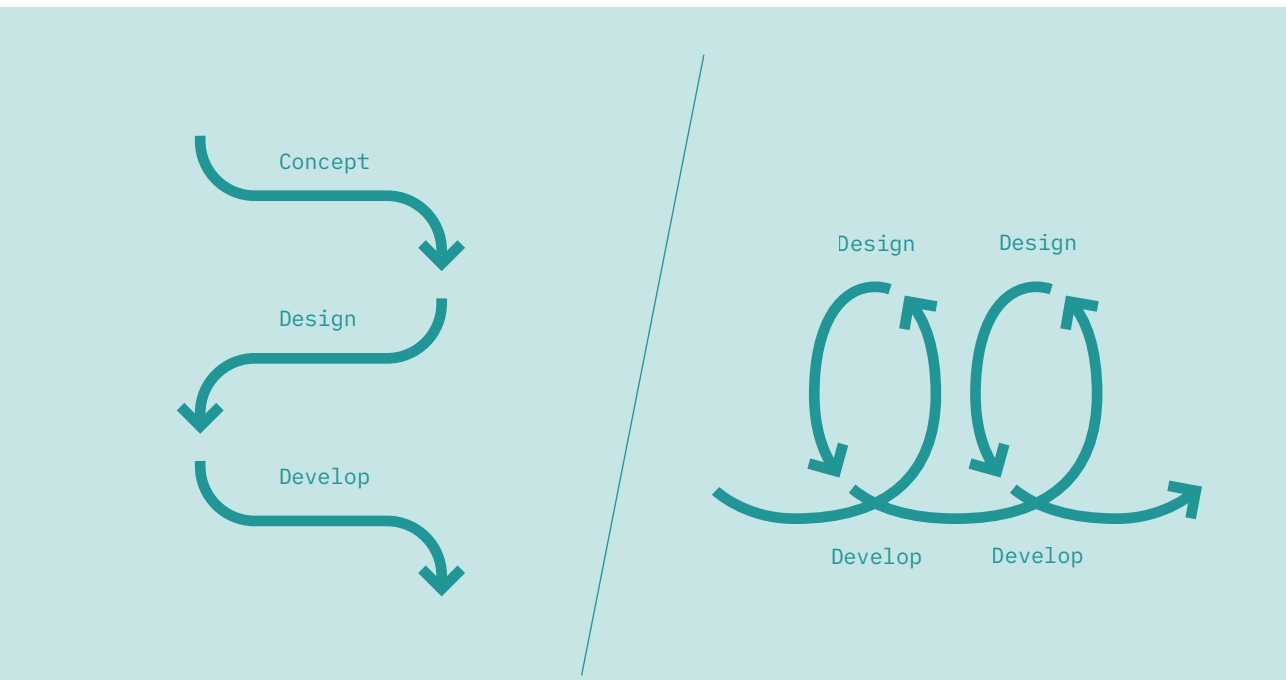
» Standaarden zijn breed gedragen afspraken over ethiek, governance en techniek van AI, waardoor AI overal aan dezelfde eisen kan voldoen. «

-- Yvette Mulder, NEN

Er mist echter nog een belangrijke stap in dit proces. Namelijk het kwantificeren van wat ‘goed’ en ‘niet goed’ is. Zonder dergelijke wegingen kan een systeem in de context niet bepalen wat de juiste handeling is. Machines hebben een vorm van morele intuïtie nodig die moet meegroeien met de samenleving. Universele waarden veranderen wellicht niet zo snel, maar de weging van verschillende specificaties ervan in verschillende contexten wel. Dit vraagt om een andere benadering van het ontwerpproces.

Het ethisch ontwerpproces

Er zijn verschillende benaderingen om software te ontwikkelen. Deze variëren van zogenaamde ‘*Waterfall*’ modellen tot ‘*Agile*’ benaderingen (Leijnen et al., 2020). Bij waterval-modellen wordt het ontwerp bij de aanvang van het proces bepaald en bij agile benaderingen is er sprake van een meer iteratieve aanpak.



Volgens de iteratieve aanpak ga je niet lineair van A naar B, maar kom je gedurende het proces nieuwe uitdagingen tegen. Hierdoor moet je soms een stap terugzetten om vooruit te komen of kom je tot het inzicht dat je niet naar B, maar naar C moet toewerken.

Het waterval-model komt in grote lijnen overeen met de huidige door waarden gedreven ethische ontwerp-disciplines, zoals *Value Sensitive Design* (VSD). Hierbij is het uitgangspunt dat waarden in een zo vroeg mogelijk stadium van het ontwerpproces gedefinieerd moeten worden, zodat zoveel mogelijk waarden ten opzichte van elkaar gemaximaliseerd kunnen worden. Innovatie staat daarbij in dienst van het optimaliseren van deze waarden. Hiervoor is het nodig om al in een vroeg stadium het ontwerp te bepalen en de waarden expliciet te maken. Een belemmering van deze waterval-aanpak is echter dat de focus voornamelijk ligt op waardenconflicten die zich op een abstract niveau voordoen. Er wordt hierbij onvoldoende rekening gehouden met spanningen die tijdens het ontwerpproces kunnen optreden. Met name wanneer het om AI-toepassingen gaat, waarbij factoren als veiligheid heel erg belangrijk zijn, is het lastig om vooraf te bepalen welke design requirements ingebed moeten worden. Dergelijke requirements zijn namelijk gevoelig voor verandering. Wanneer je dit te statisch benaderd kunnen er kwetsbaarheden in het systeem ontstaan. Er wordt hierbij ten onrechte vanuit gegaan dat requirements die vooraf opgesteld worden als het ware overvloeien in de volgende stap van het ontwerpproces. In de praktijk bestaat echter het risico dat delen hiervan verloren gaan en buiten het proces vallen. Het is daardoor mogelijk dat er nieuwe requirements nodig zijn om een ethische toepassing van AI te kunnen waarborgen. Een agile benadering maakt het beter mogelijk om onvoorziene omstandigheden op te vangen en het ontwerp gedurende het proces af te stemmen op nieuwe uitdagingen. Hierdoor is er ook gedurende het ontwerpproces voldoende aandacht voor ethische overwegingen.

Huidige agile aanpakken focussen zich echter te veel op de functionele systeemeisen. Niet de mens, maar het systeem lijkt hierbij centraal te staan. Voor de ontwikkeling van AI-systemen die ethisch verantwoorde beslissingen kunnen maken is daarom een benadering nodig waarbij ook gekeken wordt naar minder tastbare systeemeisen, zoals waarden. Bij de ontwikkeling van AI-systemen ligt de nadruk nu nog vaak op de vraag hoe we de betrouwbaarheid van AI-systemen kunnen vergroten, in plaats van op de vraag hoe we deugdelijke AI-systemen kunnen ontwikkelen die de mens op waarde schatten.
Het is daarom van belang om uitspraken te doen over hoe zwaar bepaalde waarden ten opzichte van elkaar wegen in de specifieke context. Om dit te kunnen doen is het noodzakelijk om te beseffen dat deze weging van verschillende factoren afhankelijk is.
Er zijn tenminste vier factoren van invloed:

> Betrokken stakeholders
> Maatschappelijke doelen (waarden)
> Specifieke belangen
> Context

Om ethiek in praktijk te kunnen brengen moeten al deze factoren vertaald worden naar het ontwerpproces. Voor elke toepassing is de context namelijk anders en gelden er andere belangen. Hierbij geldt dat hoe specifieker de context is, hoe sterker het design wordt. Op deze wijze krijg je maatwerk. En dat is precies wat er in de huidige ethische discussies omtrent AI mist. Er is behoefte aan een proces waarbij ethiek in de context wordt bepaald.

Scrum

Binnen de agile methodiek lijkt met name het scrumproces inspiratie te kunnen bieden om ethiek te borgen in het ontwerpproces. Zo wordt binnen scrum onder andere gebruik gemaakt van zogenaamde '*user stories*'.
Het voordeel van user stories is dat die de mens centraal zetten en de verschillende factoren die van belang zijn voor de weging van waarden aan bod laten komen.
Een user story is als volgt opgebouwd:

Als……(*stakeholder*) wil ik……(*waarden*)
om zo……(*belangen*) bij……(*context*).

Wanneer we dit vertalen naar een specifiek toepassingsdomein – zoals de zelfrijdende auto – en dit in een specifieke context plaatsen – zoals een botsing tussen twee autonome voertuigen – dan ontstaan er bijvoorbeeld voor de waarde 'transparantie' de volgende *user stories*:

> Als fabrikant wil ik de traceerbaarheid vergroten om zo de systeemfout te kunnen opsporen en verhelpen bij een botsing
> Als gebruiker wil ik de communicatie vergroten om zo geïnformeerd te zijn over de te nemen stappen en verdere afhandeling bij een botsing
> Als wetgever wil ik de verklaarbaarheid vergroten om zo waar nodig strengere eisen te kunnen stellen aan het systeem bij een botsing
> Als verzekeraar wil ik de verklaarbaarheid vergroten om zo de schuldvraag te kunnen herleiden bij een botsing

Door de universele waarde 'transparantie' in de context te plaatsen wordt deze gespecificeerd. Op deze wijze wordt er binnen transparantie onderscheid gemaakt tussen traceerbaarheid (de gegevenssets en de processen waaruit de beslissing van het AI-systeem voortkomt), verklaarbaarheid (een geschikte verklaring van het besluitvormingsproces van het AI-systeem) en communicatie (de communicatie over de mate van nauwkeurigheid ↴ en

de beperkingen van het systeem). Hierdoor wordt het duidelijk dat verschillende stakeholders verschillende belangen hebben binnen dezelfde waarde. Dit geldt ook voor andere waarden, zoals privacy:

> Als fabrikant wil ik de kwaliteit en integriteit van de data vergroten, om zo onnauwkeurigheden, fouten en vergissingen te kunnen verhelpen bij een botsing
> Als bestuurder wil ik privacy en databescherming vergroten, om zo de bescherming van mijn persoonsgegevens te garanderen bij een botsing
> Als wetgever wil ik de toegang tot data beheersen, om zo protocollen te kunnen instellen en de toegang tot gegevens te kunnen beheren bij een botsing
> Als verzekeraar wil ik de toegang tot data beheersen, om zo duidelijkheid te krijgen over wie onder welke omstandigheden gegevens kan inzien bij een botsing

Op deze wijze ontstaan requirements op gebruikersniveau en kan er dus maatwerk worden toegepast. Het voordeel van deze benadering is daarbij dat programmeurs gewend zijn om met dergelijke processen te werken. Dit versoepelt de implementatie ervan in het ontwerpproces.

Weging

Om als fabrikant een zelfrijdende auto te kunnen ontwikkelen is het niet alleen van belang om te weten wélke verschillende belangen er zijn, maar ook op welke wijze deze verschillende belangen ten opzichte van elkaar wegen. Hierbij kan tevens inspiratie gehaald worden uit het scrumproces, namelijk door te kijken naar '*planning poker*'. Normaliter wordt deze methodiek ingezet om te bepalen welke activiteiten er in het ontwerpproces voorrang krijgen en als eerste moeten worden opgepakt. Hierbij worden er meetbare waardes toegekend, zoals 0, 1, 40 en 100. In het geval van het wegen van ethische belangen werkt het echter beter om het zogenaamde '*T-shirt sizing*' toe te passen. Hierbij worden de verschillende belangen gewogen door gebruik te maken

van S, M, L en XL. Op deze wijze kan er uitdrukking gegeven worden aan de zwaarte van het belang, zonder dat de illusie wordt gewekt dat een bepaald belang bijvoorbeeld honderd keer zoveel waard is dan een ander belang. Een slechte daad is daarbij niet in te ruilen voor een goede daad. Uiteindelijk worden alle belangen meegenomen in het ontwerpproces, maar kunnen er op deze wijze wel keuzes worden gemaakt in het ontwerpproces.

Deze methodiek kan ook gehanteerd worden om mogelijke waardenconflicten op te sporen en ten opzichte van elkaar te wegen. Want wat weegt zwaarder wanneer er een botsing tussen twee autonome voertuigen heeft plaatsgevonden? De traceerbaarheid van het beslissingsproces voor de fabrikant (transparantie)? Of het beschermen van de persoonsgegevens van de gebruiker (privacy)? Ook deze belangen kunnen ten opzichte van elkaar gewogen worden. Op deze wijze wordt het inzichtelijk dat de meeste trade-offs plaatsvinden tussen gebruikersgroepen en in mindere mate tussen mens en systeem. En ook hier geldt dat het niet gaat om een uitruil, maar om een ordening van prioriteiten. Het doel is om deze waarden uiteindelijk ten opzichte van elkaar te maximaliseren door middel van het design.

Ethisch ontwerpspel

Een boek over ethiek in het ontwerpproces van AI, waarin wordt geconstateerd dat er vooral veel wordt geschreven over ethiek en een handelingsperspectief veelal ontbreekt, kan niet bij het geschreven woord blijven. We hebben daarom, geïnspireerd op het scrumproces, een ethisch ontwerpspel ontwikkeld dat kan worden ingezet om het gesprek over ethiek beter te stroomlijnen. Het spel is ontwikkeld in samenwerking met de normcommissie AI van de NEN (Stichting Koninklijk Nederlands Normalisatie Instituut) en het lectoraat Artificial Intelligence van de Hogeschool Utrecht. Het biedt beleidsmakers, ontwikkelaars, filosofen en eigenlijk iedereen met een interesse in ethiek, de mogelijkheid om met elkaar in gesprek te

gaan over wat we als samenleving belangrijk vinden bij de ontwikkeling van AI. Het helpt om meer inzicht te verkrijgen in de verschillende stakeholderperspectieven en belangen en moet bijdragen aan een constructiever gesprek over AI en ethiek. Doordat belangen ten opzichte van elkaar worden afgewogen kunnen er concrete keuzes gemaakt worden in het ontwerpproces.

A final piece of advice

Om ethisch verantwoorde AI-toepassingen in de toekomst te kunnen waarborgen moeten uiteindelijk alle stromingen en benaderingen verenigd worden. Er zijn ethische richtlijnen, assessments en standaarden nodig (beginselethiek) om waardenconflicten in een vroegtijdig stadium in kaart te kunnen brengen en waarden te kunnen verenigen (gevolgenethiek), om vervolgens AI-toepassingen te kunnen ontwikkelen die ethisch verantwoorde beslissingen kunnen nemen (deugdethiek). Hierbij zullen de beste elementen uit statisch leren en adaptief leren gecombineerd moeten worden om AI-systemen te kunnen ontwerpen die onze menselijke intuïtie kunnen nastreven en in de specifieke context de juiste beslissingen kunnen nemen. Volledige controle over deze technologieën is wellicht een illusie, maar we kunnen wel degelijk AI-systemen ontwerpen die in ons belang zullen handelen. Nu en in de toekomst.

Hiervoor moeten onder andere technici, ethici, sociologen en economen de krachten bundelen om aan AI-systemen te werken die ethisch verantwoorde beslissingen kunnen maken. Hierbij is het van belang om te concretiseren wat precies onder ‘goed’ en ‘slecht’ gedrag wordt verstaan en hoeveel waarde we aan verschillende elementen hechten. Er ligt hierbij nog een belangrijke taak voor de wetgever weggelegd. Een programmeur zou alleen verantwoordelijk moeten zijn voor het zo intelligent mogelijk maken van het systeem en het optimaliseren op de doelfuncties. Het is aan de wetgever om, geïnformeerd door de samenleving, deze doelfuncties vast te leggen. Zonder duidelijke wetgeving weet een ontwikkelaar of gebruiker niet waar hij of zij op afgerekend wordt. We zijn er nog niet. Maar ik hoop met deze uitgave en de tool de ontwikkeling van ethisch verantwoorde AI-toepassingen weer een stapje dichterbij geholpen te hebben.

» Echte ethiek begint daar, waar het gebruik van woorden ophoudt. «

-- Albert Schweitzer, 1923

Echte ethiek begint daar,
waar het gebruik van woorden

ophoudt.

-- Albert Schweitzer, 1923

De ethiek van AI in de praktijk

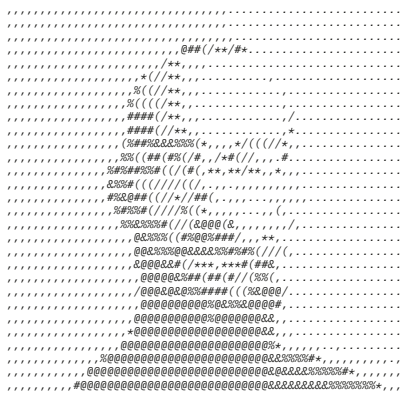
Door Bernard ter Haar, Buitengewoon adviseur, Ministerie BZK

Er wordt veel nagedacht, geschreven en gesproken over het belang van het borgen van ethische waarden voor artificiële intelligentie (AI), of zelfs voor het hele digitale ecosysteem. Er is veel theorie, en maar weinig praktijk. Het heeft ook iets verlammends, al die theorie. Alsof ethiek eigenlijk een te verheven of te moeilijk onderwerp is om concreet handen en voeten te geven. Dat is in elk geval een enorm misverstand.

> Ik zie ethiek maar even simpel gezegd als het nadenken over goed en kwaad. Daar zijn we elke dag mee bezig. We beoordelen voortdurend ontwikkelingen in deze wereld op hun goed en hun kwaad. Menigeen vindt het ethisch bijvoorbeeld onverantwoord dat we vluchtelingen laten creperen in kampen op Griekse eilanden. Anderen vinden het ethisch enorm van belang om de authentieke Nederlandse cultuur te beschermen. We oordelen niet alleen iedere dag, die oordelen verschuiven ook in de tijd. Tot in het begin van de jaren zestig van de vorige eeuw moesten vrouwen in het onderwijs ontslag nemen als ze in het huwelijk traden. Dat was vanuit de ethiek rondom de waarde van het gezin. Tegenwoordig beoordelen we dat als onderdrukking van de vrouw. Althans, de meesten van ons beoordelen dat zo. Honderd procent consensus over ethische waarden bestaat niet.
> In China hebben ze goed nagedacht over goed en kwaad en de mogelijkheden van AI. Er is een sociale krediet score systeem opgezet, waarin mensen gescoord worden of ze zich maatschappelijk goed of slecht gedragen. Een in Chinese ogen mooie manier om het ethisch gehalte van het maatschappelijk gedrag te verhogen. In Nederland gruwen we ervan, omdat het lijnrecht ingaat tegen onze normen en waarden rond individuele vrijheid en expressie. Dat een ethisch positief doel als onderlinge sociale verbondenheid als ‘sausje’ wordt gebruikt voor puur commerciële belangen, zoals bij Facebook, roept ook steeds meer wantrouwen op.

Het borgen van ethische waarden is dus dagelijkse kost, en een wezenlijk en heel dynamisch onderdeel van onze democratische rechtsstaat. Wij borgen onze ethiek in de wet- en regelgeving die we democratisch tot stand laten komen. En toch, zoals ik hierboven schreef, als het over ethiek en AI gaat zien we veel theorie en weinig praktijk. Wat gaat er mis? We zien veel handelings-verlegenheid van de wetgever, zowel departementaal als parlementair. Er is een groot gebrek aan kennis en inzicht. Er is de vraag op welk niveau de verantwoordelijkheid ligt, landelijk, Europees of mondiaal? En er is een wijdverbreide fictie over de waarde van totale vrijheid/blijheid van het internet. Bescherming van eigendom is de basis in bijna elk ethisch systeem, maar hebben digitale data ook een eigenaar? Natuurlijk bestaat die ook, maar dat moet wel netjes vastgelegd worden.

> Om ethiek in AI een goede invulling te geven zijn dus een paar dingen van belang. De wetgever moet weten wat er gebeurt in de digitale wereld. Uiteraard, maar het gaat verder. We moeten het met zijn allen weten. De democratie kan immers alleen maar goed functioneren bij een breed maatschappelijk debat. De transparantie van elk digitaal handelen moet enorm worden vergroot. Dat is ingewikkeld, daar moeten we dus heel hard over na gaan denken. Als we meer zicht hebben op alles wat er digitaal gebeurt, ook in de zich ontwikkelende wereld van AI, kunnen we van dag tot dag bedenken wat wel door de beugel kan, en wat niet door de beugel kan. En die laatste categorie gaan we dan op de één of andere manier begrenzen.
> De ervaring moet ons wijzer maken, we kunnen het echt niet allemaal van tevoren reguleren. Dat is in de fysieke wereld niet anders. Het vraagt wel een wetgever, die het tempo van de digitale ontwikkeling kan bijbenen. Een uitdaging, maar niet één waar je voor weg kunt lopen!



Leg niet alle verantwoordelijkheid bij de programmeur

Door Rudy van Belkom

Er was veel kritiek op de ‘appathon’ die het ministerie van Volksgezondheid organiseerde om slimme technologieën te kunnen inzetten tegen de verspreiding van het coronavirus. Het proces zou te gehaast en te chaotisch zijn. In een *pressure cooker* moesten ontwikkelaars de laatste verbeterslagen in de tracking-apps doorvoeren. De resultaten waren volgens de betrokken experts onbevredigend. Geen van de ingebrachte apps voldeed aan de gestelde privacyrichtlijnen. Ethisch onverantwoord, zo klonk het eindoordeel.

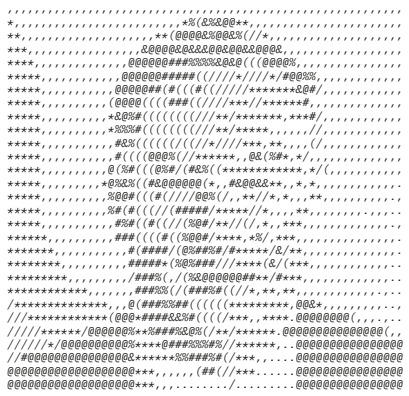
> Toch zit de echte kritiek wat mij betreft niet in de appathon zelf, maar in de wijze waarop we met z’n allen ethiek bedrijven. Nu ligt de eindverantwoordelijkheid onterecht (en wellicht onbedoeld) bij de programmeurs. We denken dat we alles met de verschillende ethische principes en richtlijnen hebben dichtgetimmerd, maar niets is minder waar. Dergelijke voorschriften zeggen namelijk niets over de wijze waarop waarden als privacy in wiskundige termen uitgedrukt moeten worden. Bij de ontwikkeling van artificiële intelligentie (AI) draait alles namelijk om statistiek. Slimme systemen moeten in staat zijn om zelfstandig patronen uit grote hoeveelheden data te halen en daarvan te kunnen leren. Naast het feit dat hierbij persoonsgegevens op straat kunnen komen te liggen, kan het systeem bepaalde groepen gebruikers onbedoeld benadelen door verkeerde interpretaties te maken op basis van deze data.

Het gaat in dat opzicht dus eigenlijk niet zozeer over privacy, maar over *fairness.* En de vraag is wat in dit geval eerlijk is vanuit een statistisch oogpunt. Willen we geen coronagevallen over het hoofd zien, of willen we geen mensen onterecht in quarantaine zetten? En wat als blijkt dat mensen in bepaalde postcodegebieden een hogere kans op

besmetting hebben? Moeten deze variabelen dan meegewogen worden, of willen we toch dat iedereen gelijkwaardig wordt behandeld? De vraag is daarbij wat ‘eerlijk genoeg’ is. Welk percentage onnodige quarantainegevallen kunnen en willen we accepteren? De verschillende wiskundige definities van eerlijkheid blijken elkaar in de praktijk uit te sluiten. Dus als we dergelijke voorschriften niet kwantificeren zal een programmeur volgens de huidige benadering zélf een keuze moeten maken.

> Wat we onder eerlijkheid verstaan is daarbij context-afhankelijk. Een aantal maanden geleden konden we het ons überhaupt niet voorstellen dat er mensen in quarantaine geplaatst zouden moeten worden. Daarbij zullen we waarschijnlijk anders oordelen als het over een onterechte gevangenisstraf gaat, in plaats van een relatief luxe quarantaine in eigen huis. Universele waarden zijn niet zo veranderlijk, maar onze normen wel. Dat maakt het niet alleen lastig, maar ook onverantwoord om principes en richtlijnen in AI-systemen te programmeren. Uiteindelijk moet het systeem zelf in staat kunnen zijn om morele afwegingen te maken. Dat klinkt eng, maar zonder een moreel intuïtief systeem is het bijna onmogelijk om AI op een ethisch verantwoorde manier in de praktijk toe te passen. De enige reden dat mensen met regels kunnen omgaan, is omdat we deze in de context kunnen vertalen naar gedrag. We moeten momenteel 1,5 meter afstand houden van elkaar. Maar als iemand zijn evenwicht verliest en van de trap dreigt te vallen, is het toch wel wenselijk dat ik die persoon zal opvangen als ik daartoe in staat ben.

We zouden daarom minder energie moeten steken in het opstellen van ethische richtlijnen en ons meer moeten bezighouden met het bouwen van systemen die in de context ethisch verantwoorde beslissingen kunnen nemen. Ethiek is bij uitstek een ontwerpvraagstuk. Naast programmeurs en ethici, moeten ook sociologen, psychologen, biologen en economen hieraan meewerken. Lukt ons dat niet? Dan moeten we wellicht in de toekomst geen AI-systemen meer willen inzetten. Of accepteren dat onze ethiek onethisch is.



Ethiek in actie

Nawoord door Patrick van der Duin
Directeur, Stichting Toekomstbeeld der Techniek

Ere wie ere toekomt. Het was voormalig minister-president Jan Peter Balkenende die in 2002 zei dat we het weer moeten gaan hebben over normen en waarden. Ook ik behoorde tot de velen die daar toen smakelijk om moesten lachen met het argument dat dit wel een erg oude discussie was. En dat zijn normen en waarden niet de mijne zijn (en dat het eigenlijk waarden en normen zijn). Maar zie hier, anno 2020, zijn normen en waarden belangrijker dan ooit tevoren. Balkenende is een ware profeet gebleken die daarvoor terecht beloond is door de vernoeming van een norm naar hem: de Balkenende-norm, die aangeeft dat leidinggevenden in de publieke en semipublieke sector jaarlijks niet meer mogen ontvangen dan een minister.

De huidige *ethical turn* (misschien wel vergelijkbaar met de *linguistic turn*) maakt ook andere nostalgische gevoelens bij mij los. Rond 2002 werkte ik namelijk op de faculteit Techniek, Bestuur en Management van de TU Delft. Er was een sectie Filosofie van de Techniek, maar dat mocht geen grote naam hebben als we af zouden gaan op het aantal medewerkers: een handjevol. Maar anno 2020 is deze sectie de grootste van de hele faculteit. Onder de paraplu van ‘responsible innovation’ zijn de dames en heren ethici en filosofen uit hun leunstoel gekomen om te onderzoeken hoe ze ethiek in actie kunnen brengen. Geen vrijblijvend gefilosofeer meer, maar onderzoek doen naar hoe ontwerpprocessen ethisch gestuurd kunnen worden en daarvoor praktische methodieken ontwikkelen die de wereld echt een beetje humaner moeten maken.

De STT-studie naar AI in de toekomst uitgevoerd door projectleider Rudy van Belkom moet zeer in dit licht van het toenemende belang van ethiek in onze maatschappij, economie en technologie gezien worden. Dit laatste deel is een logische afsluiting van een drieluik. Eerst is onderzocht wat AI is, hoe divers het is en ook wat het niet is. Niet om de studie af te bakenen maar vooral om helderheid te verschaffen over wat AI betekent zodat de hype rondom AI beter op waarde(n) kan worden geschat. In de tweede studie werd bewust weer enige verwarring gecreëerd door met behulp van een aantal scenario’s te laten zien welke verschillende mogelijke toekomsten voor AI voorstelbaar zijn. Met de bedoeling om het dominante discourse van AI als iets dat heel slecht is voor de mensheid te doorbreken en te laten zien dat er alternatieven zijn. Alternatieven die niet uit de lucht komen vallen maar het gevolg zijn van wat wij als maatschappij willen en wensen. Velen stellen dat AI misschien wel de meeste ingrijpende technologie is die de mensheid ooit heeft en zal ontwikkelen. Juist daarom is ‘human agency’ belangrijk zodat we AI vorm geven zoals wij dat willen.

In dit derde en laatste deel heeft Rudy van Belkom laten zien hoe je woorden over ethiek in daden kunt omzetten. Niet alleen door uit te leggen dat ethiek (net als AI) een veelkoppig, goedbedoeld monster is, maar ook door te stellen dat deze daden alleen uitgevoerd kunnen worden door er een goed gesprek over te voeren met eenieder die zich interesseert in en bekommert om AI. Over ethiek moet je praten, maar ethiek moet je ook doen. Deze STT-studie legt als een van de eerste studies een directe koppeling tussen nadenken over de toekomst en ernaar handelen. Wat betreft de ontwikkeling van AI is dat overigens geen luxe meer maar pure noodzaak. Niet alleen om te voorkomen dat het de ‘verkeerde’ kant opgaat met AI, maar vooral ook door de mogelijkheden om AI in te zetten ten behoeve van de normen en waarden van onze maatschappij.

Begrippenlijst

Accuracy

In het algemeen geldt dat hoe vollediger en omvangrijker de dataset is waarmee een AI-systeem getraind wordt, hoe nauwkeuriger het systeem is. III.2.2 Botsende waarden ↴

Accountability

Verantwoordelijkheden die door de wet worden voorgeschreven noemen we aansprakelijkheid. III.1.2 Prangende ethische vraagstukken ↴

Affective Computing

Affective Computing gaat over systemen die emoties kunnen opsporen en herkennen. I.1.1 Hoe werkt AI?

Afhankelijkheid van data

Een van de grootste beperkingen van AI is dat het afhankelijk is van grote hoeveelheden data. Bij deep learning wordt daarom ook wel gesproken over data-hungry neural networks. Dit maakt dat de technologie niet goed is in randgevallen, waarbij weinig data beschikbaar is. I.2.3 Voordelen en beperkingen

AI

De meest dominante associatie met AI is machine learning. Uit onderzoek van de World Intellectual Property Organisation (WIPO) uit 2019 blijkt tevens dat machine learning de meest dominante AI-technologie is die in patentaanvragen is opgenomen. Binnen deze toekomstverkenning richten we ons daarom hoofdzakelijk op machine learning en aanverwante methodieken. I.1.1 Hoe werkt AI?

Algoritmen

Een algoritme is een wiskundige formule. Het is een eindige reeks instructies die vanuit een gegeven begintoestand naar een vooraf bepaald doel leidt. I.1.1 Hoe werkt AI?

Beeldherkenningssystemen

Bij toepassingen op het gebied van beeldherkenning wordt vaak gebruik gemaakt van een zogenaamd Convolutional Neural Network (CNN). CNN werkt als een filter die over een afbeelding beweegt en op zoek gaat naar de aanwezigheid van bepaalde eigenschappen.
I.1.1 Hoe werkt AI?

Beginselethiek

Bij beginselethiek wordt steeds een beginsel als uitgangspunt genomen. Denk bijvoorbeeld aan eerbied voor het leven en menselijke waardigheid. Bij de oplossing van een ethisch probleem moet recht gedaan worden aan een of meer van deze beginselen of principes. Het beginsel dient hierbij te allen tijde toegepast te worden, onafhankelijk van de gevolgen. III.1.3 Een kwestie van ethisch perspectief ↴

Besluitvorming

We willen graag geloven dat wij mensen rationele wezens zijn. Maar het besluitvormingsproces is grillig. Beeldvorming en ambities spelen naast feitelijke kennis een belangrijke rol. I.2.2 Geautomatiseerde besluitvorming

Biases

Aangezien we evolutionair zijn geprogrammeerd om zoveel mogelijk energie te besparen sluipen er wel wat ‘denkfouten’ in onze informatieverwerking. Deze denkfouten worden ook wel cognitive biases genoemd. I.2.2 Geautomatiseerde besluitvorming

Clusters

Het algoritme gaat bij het maken van clusters zelfstandig op zoek naar overeenkomsten in de data en probeert op deze wijze patronen te herkennen. I.1.1 Hoe werkt AI?

Common Sense

Common sense bestaat uit alle kennis over de wereld; van fysieke en zichtbare aspecten, tot culturele en dus meer impliciete regels, zoals hoe je met elkaar omgaat. I.2.1 Perspectieven op intelligentie

Deep learning

Deep learning is een machine learning-methode die gebruik maakt van verschillende gelaagde artificial neural networks.
I.1.1 Hoe werkt AI?

Deep reasoning

Deze nieuwe benadering pakt problemen van de oude benaderingen aan door ze te combineren. Deep Reasoning verhelpt het schaalbaarheidsprobleem van het symbolisme (het is onmogelijk om alle opties efficiënt te programmeren) en het pakt tegelijkertijd het dataprobleem van neurale netwerken aan (grote datasets zijn vaak niet beschikbaar of onvolledig). I.3.1 Expert voorspellingen

Deugdethiek

Bij de deugdethiek staan niet de regels of bepaalde principes centraal binnen het moreel oordelen, maar het karakter van de persoon die handelt. Ook hier staat het handelen los van de expliciete gevolgen die het heeft. Om goed te kunnen handelen zijn bepaalde karaktereigenschappen nodig, oftewel deugden.
I.1.3 Een kwestie van ethisch perspectief ↴

Empathie

Empathie is de vaardigheid om je in te kunnen leven in de situatie en gevoelens van anderen. Door empathie zijn we in staat om ook de non-verbale communicatie van anderen te lezen en begrijpen.
I.2.1 Perspectieven op intelligentie

Ethiek

Ethiek is een tak binnen de filosofie die zich bezighoudt met de systematische reflectie van wat als goed of juist handelen bestempeld wordt. III.1 AI en Ethiek ↴

Ethische richtlijnen

In de afgelopen jaren hebben uiteenlopende bedrijven, onderzoeksinstellingen en overheidsorganisaties verschillende principes en richtlijnen opgesteld voor *ethical AI*. Zowel op nationaal en continentaal niveau, als op mondiaal niveau.
III.2.1 Van corporate tot government ↴

Explainibility

Het wordt steeds lastiger voor mensen om te achterhalen op basis van welke data de uitkomsten van AI gebaseerd zijn. AI wordt daarom nu nog vaak vergeleken met een black box. I.2.2 Geautomatiseerde besluitvorming

Fairness

Fairness is een veelgebruikt principe in ethische richtlijnen en assessment tools. Het vraagstuk wat precies ‘fair’ is houdt filosofen al een paar honderd jaar bezig. Door de komst van AI krijgt dit vraagstuk een nieuwe dimensie. Het concept van eerlijkheid moet namelijk in wiskundige termen worden uitgedrukt.
III.2.3 Uitdagingen in de praktijk ↴

Formele logica

In de eerste fase van AI, van 1957 tot eind jaren ‘90, werd voornamelijk gebruikt gemaakt van formal logic, oftewel als-dan regels. Deze vorm van AI was vooral gericht op high level cognition, zoals redenering en probleemoplossing. I.1.1 Hoe werkt AI?

General AI

Artificial General Intelligence zou in staat moeten zijn om alle intellectuele taken uit te voeren die een mens ook kan uitvoeren.
I.1.1 Hoe werkt AI?

Gevolgenethiek

De gevolgenethiek stelt dat de gevolgen van een bepaalde handeling bepalen of er sprake is van ‘juist handelen’. Het gedrag moet dus positieve gevolgen hebben, zelfs wanneer hierbij bepaalde beginselen ondermijnd worden. De handeling zelf wordt dus niet ter discussie gesteld, alleen de gevolgen ervan. III.1.3 Een kwestie van ethisch perspectief ↴

Intelligentie

Intelligentie kan omschreven worden als een aaneenschakeling van verstandelijke vermogens, processen en vaardigheden, zoals kunnen redeneren en je flexibel kunnen aanpassen in nieuwe situaties. I.2.1 Perspectieven op intelligentie

Intentionaliteit

Zelfs wanneer je alle kennis over de wereld in een computer programmeert, dan blijft de vraag of een computer daadwerkelijk begrip heeft van zijn handelen. Dit begrip wordt ook wel intentionaliteit genoemd. I.2.1 Perspectieven op intelligentie

Machine biases

Niet alleen menselijke intelligentie, maar ook AI heeft last van vooroordelen. De output van algoritmen kan gender en race biases bevatten. De verklaring ervan is simpel. Wanneer de input niet zuiver is, dan is de output dat ook niet. Biases van algoritmen worden dus veroorzaakt door cognitieve biases van mensen. I.2.2 Geautomatiseerde besluitvorming

Machine Learning

Het gaat bij machine learning over een revolutie waarin mensen niet meer programmeren (als dit, dan dat), maar waarin machines zelf regels afleiden uit data. I.1.1 Hoe werkt AI?

Moraliteit

In de discussies lopen ethiek en moraliteit vaak door elkaar heen. Toch is er een duidelijk verschil. Moraliteit is het geheel van meningen, beslissingen en acties waarmee mensen (individueel of collectief) uitdrukking geven aan wat zij denken dat goed of juist is. Ethiek is vervolgens de systematische reflectie op wat moreel is. III.1.3 Een kwestie van ethisch perspectief ↴

Motorkap

We kunnen de complexe concepten die wij aan het brein willen koppelen, zoals bewustzijn en vrije wil, nog niet modelleren en we kunnen ze niet een-op-een linken aan gebieden in het brein. I.2.1 Perspectieven op intelligentie

Narrow AI

Artificial Narrow Intelligence is een vorm van AI die zeer goed is in het doen van specifieke taken. Denk hierbij aan schaken, aanbevelingen doen en het geven van kwantificeerbare voorspellingen. I.1.1 Hoe werkt AI?

Neurale netwerken

Artificiële neurale netwerken worden gebruikt binnen deep learning en zijn oorspronkelijk gebaseerd op het menselijk brein, waarbij neuronen laagsgewijs met elkaar verbonden zijn. I.1.1 Hoe werkt AI?

Privacy

In het Europees Verdrag voor de Rechten voor de Mens is privacy opgenomen als het recht op respect voor het privéleven. Deze bepaling vereist een eerlijke balans tussen het maatschappelijk belang dat een technologie dient en de inbreuk op het privéleven die de technologie maakt. III.1.2 Prangende ethische vraagstukken ↴

Rechtvaardigheid

Wanneer het over rechtvaardigheid gaat, dan gaat het in de basis over de gelijkwaardigheid van mensen. Mensen dienen gelijk behandeld te worden én gelijke kansen te krijgen. III.1.2 Prangende ethische vraagstukken ↴

Schrikbeeld

Scenario's over robots die in opstand komen vormen al bijna 100 jaar een populaire verhaallijn (ervan uitgaande dat 'Metropolis' uit 1927 van regisseur Fritz Lang de eerste echte sciencefiction film is waarbij een robot slechte intenties heeft). III.1.1 Ethiek in de spotlight ↴

Superintelligence

Artificial Super Intelligence (ASI) kan bereikt worden wanneer AI het kunnen van het menselijk brein op alle mogelijke domeinen overstijgt. I.1.1 Hoe werkt AI?

Transparency

Transparantie binnen het besluitvormingsproces van AI gaat voornamelijk over het proces en de opgestelde criteria vooraf. III.1.2 Prangende ethische vraagstukken ↴

Uitleg en uitlegbaarheid

Binnen het besluitvormingsproces van AI gaan de uitleg en uitlegbaarheid over de toelichting op en herleidbaarheid van het besluit achteraf. III.1.2 Prangende ethische vraagstukken ↴

Verantwoordelijkheid

Wanneer het over de ontwikkeling van AI gaat wordt het begrip 'verantwoordelijkheid' vaak genoemd. Wie is er bijvoorbeeld verantwoordelijk bij een ongeluk met een zelfrijdende auto? Maar om te bepalen wie er verantwoordelijk is moeten we eerst bepalen waar zij dan precies verantwoordelijk voor zijn en welk gedrag wel en niet te verantwoorden is. Daarbij is het de vraag op welke wijze we de mate van verantwoordelijkheid kunnen herleiden. III.1.2 Prangende ethische vraagstukken ↴

Vertrouwen

Uit onderzoek van onder andere de University of Pennsylvania uit 2014 blijkt dat wanneer mensen een algoritme een kleine en betekenisloze fout zien maken, dat de kans dan groot is dat het vertrouwen volledig verloren gaat. Dit wordt door onderzoekers ook wel algorithm aversion genoemd. I.2.2 Geautomatiseerde besluitvorming

Vrijheden en rechten

Vrijheid gaat in de basis over de vrijheid die mensen hebben om zelf te kunnen bepalen hoe zij hun leven inrichten. Dit is zelfs een recht, het zelfbeschikkingsrecht. Dit recht wordt vervolgens ingeperkt door het verbod om bij anderen schade aan te richten. III.1.2 Prangende ethische vraagstukken ↴

Wedstrijd

Momenteel is de interesse in AI dermate groot dat wereldmachten de strijd met elkaar aangaan. Uit onderzoek van PwC uit 2017 blijkt dat het wereldwijde Gross Domestic Product (GDP) in 2030 door ontwikkelingen op het gebied van AI maar liefst 14% hoger zal liggen. Volgens de Russische president Poetin zal het land met de beste AI over de wereld heersen (2017). I.3.2 Economische en politieke belangen

Wiskunde

AI is in de basis 'gewoon' wiskunde. Weliswaar een enorm geavanceerde vorm van wiskunde, maar het blijft wiskunde. Het is boven alles een middel om een optimalisatiedoel te bereiken. I.1.1 Hoe werkt AI?

Bronnen

AI Now Institute (2018). AI in 2018: a year in review. Geraadpleegd van https://medium.com/@AINowInstitute/ai-in-2018-a-year-in-review-8b161ead2b4e

AI Now Institute (2019). AI in 2019: a year in review. Geraadpleegd van https://medium.com/@AINowInstitute/ai-in-2019-a-year-in-review-c1eba5107127

Araujo, T. et al. (2018). Automated Decision-Making Fairness in an AI-driven World: Public Perceptions, Hopes and Concerns. Geraadpleegd van http://www.digicomlab.eu/wp-content/uploads/2018/09/20180925_ADMbyAI.pdf

Beijing Academy of Artificial Intelligence (2019). Beijing AI Principles. Geraadpleegd van https://www.baai.ac.cn/blog/beijing-ai-principles

Belkom, R. van (2019). Duikboten zwemmen niet; de zoektocht naar intelligente machines. Den Haag: STT

Belkom, R. van (2019). De computer zegt nee; scenario's over onze toekomst met AI. Den Haag: STT

Blauw, S. (2018). Algoritmes zijn even bevooroordeeld als de mensen die ze maken. Geraadpleegd van https://decorrespondent.nl/8802/algoritmes-zijn-even-bevooroordeeld-als-de-mensen-die-ze-maken/4676003192124-d89e66a9

Boer, M. de (2020). The many futures of Artificial Intelligence: Scenarios of what AI could look like in the EU by 2025. Geraadpleegd van https://www.pwc.nl/nl/actueel-publicaties/assets/pdfs/the-many-futures-of-artificial-intelligence.pdf

Bostrom, N. (2016). Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press

Dalen, W. van (2012). Ethiek de basis; Morele competenties voor professionals. Groningen/Houten: Noordhoff Uitgevers

Delvaux, M. (2016). DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics. Geraadpleegd van https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf?redirect

Duin, P. van der, Snijders, D. & Lodder, P. (2019). De Nationale Toekomstmonitor: Hoe kijken Nederlanders naar technologie en de toekomst? Den Haag: STT

Eckersley, P. (2018). Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function) Geraadpleegd van https://arxiv.org/abs/1901.00064

Elsevier (2018). ArtificiaI Intelligence: How knowledge is created, transferred, and used. Geraadpleegd van https://www.elsevier.com/research-intelligence/resource-library/ai-report

Est, R. van & Gerritsen, J. (2017). Human rights in the robot age Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality. Geraadpleegd van https://www.rathenau.nl/nl/digitale-samenleving/mensenrechten-het-robottijdperk

Eubanks, V. (2018). Automating Inequality; How High-Tech Tools Profile, Police, and Punish the Poor. New York: St. Martin's Press

Europese Commissie (2020). Attitudes towards the impact of digitalisation on daily lives. Geraadpleegd van https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/survey/getsurveydetail/instruments/special/surveyky/2228

Europese Commissie (2020). White Paper on Artificial Intelligence: a European approach to excellence and trust. Geraadpleegd van https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center Research Publication No. 2020-1. http://dx.doi.org/10.2139/ssrn.3518482

Future of Life Institute (2017). Asilomar AI Principles. Geraadpleegd van https://futureoflife.org/ai-principles/

G7 Innovation Ministers (2018). G7 Innovation Ministers' Statement on Artificial Intelligence. Geraadpleegd van http://www.g8.utoronto.ca/employment/2018-labour-annex-b-en.html

Gartner (2019). Top 10 Strategic Technology Trends for 2019: Digital Ethics and Privacy. Geraadpleegd van https://www.gartner.com/en/documents/3904420/top-10-strategic-technology-trends-for-2019-digital-ethi

Genesys (2019). New Workplace Survey Finds Nearly 80% of Employers Aren't Worried About Unethical Use of AI — But Maybe They Should Be. Geraadpleegd van https://www.genesys.com/en-gb/company/newsroom/announcements/new-workplace-survey-finds-nearly-80-of-employers-arent-worried-about-unethical-use-of-ai-but-maybe-they-should-be

Google (2019). Perspectives on Issues in AI Governance. Geraadpleegd van https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf

High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI. Geraadpleegd van https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Hill, K. (2020). The Secretive Company That Might End Privacy as We Know It. Geraadpleegd van https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

Hoven, J. van den, Miller, S. & Pogge, T. (2017). Designing in Ethics. Cambridge: Cambridge University Press

IBM (2019). Everyday Ethics on AI. Geraadpleegd van https://www.ibm.com/design/ai/ethics/everyday-ethics

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat Mach Intell* 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kant, I. (1788). Kritik der praktischen Vernunft. Keulen: Anaconda Verlag

Kaur, H. et al. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. Geraadpleegd van http://www.jennwv.com/papers/interp-ds.pdf

Leijnen, S. et al. (2020). An Agile Framework for Trustworthy AI. ECAI 2020 position paper.

Maurits, M. & Blauw, S. (2019). In de stad van de toekomst praten lantaarnpalen mee en burgers niet. Geraadpleegd van https://decorrespondent.nl/9148/in-de-stad-van-de-toekomst-praten-lantaarnpalen-mee-en-burgers-niet/4859813360776-3fcc1087

Microsoft Research (2018). Manipulating and Measuring Model Interpretability. Geraadpleegd van https://arxiv.org/pdf/1802.07810.pdf

Ministerie van Economische Zaken en Klimaat (2019). Strategisch Actieplan voor Artificiële Intelligentie. Geraadpleegd van https://www.rijksoverheid.nl/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie

NeurIPS. (z.d.). Code of Conduct. Geraadpleegd van https://nips.cc/public/CodeOfConduct

OECD (2019). Principles on AI. Geraadpleegd van http://www.oecd.org/going-digital/ai/principles/

Pew Research Center (2018). The Data on Women Leaders. Geraadpleegd van https://www.pewsocialtrends.org/fact-sheet/the-data-on-women-leaders/#ceos

Poel, I. van de & Royakkers, L. (2011). Ethics, Technology and Engineering; An Introduction. Hoboken: Wiley-Blackwell

Poel, I. van de. (2013). Translating Values into Design Requirements. In: MichelFelder D., McCarthy N., Goldberg D. (eds), Philosophy and Engineering: Reflections on Practice, Principles and Process. Vol.

15 of Philosophy of Engineering and Technology, Springer, Dordrecht (pp.253-266)

ProPublica (2016). Machine Bias. Geraadpleegd van https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Rahwan, I. (2019). Machine behavior. Geraadpleegd van https://www.nature.com/articles/s41586-019-1138-y

Russel, S. (2017). 3 principles for creating safer AI. Geraadpleegd van https://www.ted.com/talks/stuart_russell_3_principles_for_creating_safer_ai

Selbst, A.D. et al. (2019). Fairness and Abstraction in Sociotechnical Systems. Geraadpleegd van https://dl.acm.org/doi/pdf/10.1145/3287560.3287598

Snijders, D., Biesiot, M., Munnichs, G. & Est, R. van (2019). Burgers en sensoren – Acht spelregels voor de inzet van sensoren voor veiligheid en leefbaarheid. Den Haag: Rathenau Instituut

Smart Dubai (2019). AI Ethics Principles & Guidelines. Geraadpleegd van https://www.smartdubai.ae/pdfviewer/web/viewer.html?file=https://www.smartdubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf?sfvrsn=d4184f8d_6

The Pontifical Academy for Life (2020). Rome Call for AI Ethics. Geraadpleegd van http://www.academyforlife.va/content/dam/pav/documenti%20pdf/2020/CALL%2028%20febbraio/AI%20Rome%20Call%20x%20firma_DEF_DEF_.pdf

UNI Global Union (2017). Top 10 Principles for ethical AI. Geraadpleegd van http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

U.S. Department of Defense (2020). AI Principles: Recommendations on the Ethical Use of Artificial Intelligence. Geraadpleegd van https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

Utrecht Data School (2017). De Ethische Data Assistent (DEDA). Geraadpleegd van: https://dataschool.nl/wp-content/uploads/sites/272/2019/10/DEDA_3_0.pdf

Verbeek, P.-P (2011). De grens van de mens: Over ethiek, techniek en de menselijke natuur. Rotterdam: Lemniscaat

Verbeek, P.-P. & Tijink, D. (2019). Aanpak begeleidingsethiek: Een dialoog over technologie met handelingsperspectief. Geraadpleegd van https://ecp.nl/wp-content/uploads/2019/11/060-001-Boek-Aanpak-begeleidingsethiek-240165-binnenwerk-digitaal.pdf

Wees, K.A.P.C. van (2018). Voertuigautomatisering en productaansprakelijkheid. Geraadpleegd van https://www.verkeersrecht.nl/vr-kort/voertuigautomatisering-en-productaansprakelijkheid

Weijer, B. van de (2019). Binnenkort op de weg: zelfrijdende robottaxi's zonder reservemens achter het stuur. Geraadpleegd van https://www.volkskrant.nl/economie/binnenkort-op-de-weg-zelfrijden-de-robottaxi-s-zonder-reservemens-achter-het-stuur~b671f376/

Wilson, B., Hoffman, J. & Morgenstern, J. (2019). Predictive Inequity in Object Detection. Geraadpleegd van https://arxiv.org/abs/1902.11097

Bijlagen

Over de auteur

Rudy van Belkom MA studeerde Brand, Design & Reputation Management aan het European Institute for Brandmanagement (EURIB). Voor zijn masterthesis deed hij onderzoek naar de factoren die doorslaggevend zijn bij de acceptatie van radicale technologische innovaties. Vertrouwen in de technologie blijkt hierin een cruciale rol te spelen. Tot zijn indiensttreding bij STT deed hij onder andere onderzoek naar de rol van sciencefiction bij toekomstonderzoek in het Fontys lectoraat Futures Research & Trendwatching en gaf hij jarenlang les in conceptontwikkeling en trendonderzoek in het hoger onderwijs.

Deelnemers

Voor dit deel van de toekomstverkenning is een online vragenlijst uitgezet onder drie groepen, namelijk experts, bestuurders en studenten. Deze vragenlijst is door meer dan 100 respondenten anoniem ingevuld. Namens STT en mijzelf enorm bedankt voor het invullen van deze vragenlijst. Daarnaast een specifiek dankwoord aan de experts die mee hebben gedacht, gelezen en geschreven aan dit deel van de toekomstverkenning.

Roland Bijvank	HU	Lecturer Informatics
Marc Burger	Capgemini	CEO
Patrick van der Duin	STT	Directeur
Arjen Goedegebure	OGD ict-diensten	Data Scientist
Bernard ter Haar	Ministerie BZK	Buitengewoon adviseur
Fred Herrebout	T-Mobile	Senior Strategy Manager

Jeroen van den Hoven	TU Delft	Professor of Ethics and Technology
Marijn Janssen	TU Delft	Professor ICT & Governance
Leon Kester	TNO	Senior Research Scientist
Maria de Kleijn-Lloyd	Kearney	Senior Principal
Stefan Leijnen	HU	Lector Artificial Intelligence
Leendert van Maanen	Utrecht University	Assistant professor in Human-centred AI
Michel Meulpolder	BigData Republic	Lead Data Architect
Evelien Mols	STT	Afstudeerder
Gerard Nijboer	Gemeente Rotterdam	Procesmanager Innovatie
Yvette Mulder	NEN	Committee manager AI
Eefje Op den Buysch	STT	Toekomstverkenner
Marieke van Putten	Ministerie BZK	Senior Innovation Manager
Edgar Reehuis	Hudl	Engineer
Jelmer de Ronde	SURF	Projectmanger SURFnet
Martijn Scheltema	EUR	Professor in Civil Law
Klamer Schutte	TNO	Lead Scientist Intelligent Imaging
Robert de Snoo	Human & Tech	Founder
Marc Steen	TNO	Senior Research Scientist
Maarten Stol	Brain-Creators	Principal Scientific Advisor

Stichting Toekomstbeeld der Techniek

De Stichting Toekomstbeeld der Techniek (STT) is in 1968 opgericht door het Koninklijk Instituut Van Ingenieurs (KIVI). STT is een onafhankelijke stichting die gefinancierd wordt uit bijdragen van overheid en bedrijfsleven. STT voert brede toekomstverkenningen uit op het snijvlak van technologie en samenleving die domeinoverstijgend en interdisciplinair zijn. Het Algemeen Bestuur (AB) van STT bestaat uit topmensen uit de overheid, het bedrijfsleven, de onderzoekswereld en uit maatschappelijke organisaties. Het AB denkt mee over de STT-programmering, is betrokken bij verkenningen en vormt een belangrijke denktank waarbinnen de bestuursleden praten over toekomstige technologische ontwikkelingen en innovatie.

Daarnaast ontplooit de STT Academy diverse activiteiten, zoals het co-financieren van bijzondere leerstoelen, methodiekontwikkeling en het beheer van het Netwerk Toekomstverkenningen en Young STT. Deze laatste bestaat uit young high potentials uit de deelnemende organisaties.

Informatie over STT, haar activiteiten en haar producten is te vinden op www.stt.nl.

Eerder verschenen publicaties STT

STT90 *Bioprinters, grondstof-rotondes en brainternet? Hoe wij produceren, consumeren en herverdelen in 2050.* Silke den Hartog-de Wilde, 2019

STT89 *Veiligheid in de toekomst.* Carlijn Naber, 2019

STT88 *Het eeuwige leren. Over leren, technologie en de toekomst.* Dhoya Snijders, 2018

STT87 *En toen ging het licht aan…; Transities naar een emissievrij energiesysteem.* Soledad van Eijk, 2017

STT86 *Data is macht. Over Big Data en de toekomst.* Dhoya Snijders, 2017

Colofon

STT92 deel III: AI heeft geen stekker meer

Onderzoek en projectleiding: Rudy van Belkom, STT
Tekst- en taalredactie: Japke Schreuders, STT
Grafisch ontwerp: YURR.Studio

NUR-nr. 950

ISBN 978-94-91397-23-3

Trefwoorden
AI, artificiële intelligentie, ethiek, toekomst, maatschappij, ontwerp, backcasting

In deze reeks over de toekomst van AI verscheen eerder deel I: 'Duikboten zwemmen niet' over de techniek van AI en deel II: 'De computer zegt nee' over de maatschappelijke impact van AI.

Stichting Toekomstbeeld der Techniek
Koninginnegracht 19
2514 AB Den Haag
070-302 98 30
info@stt.nl