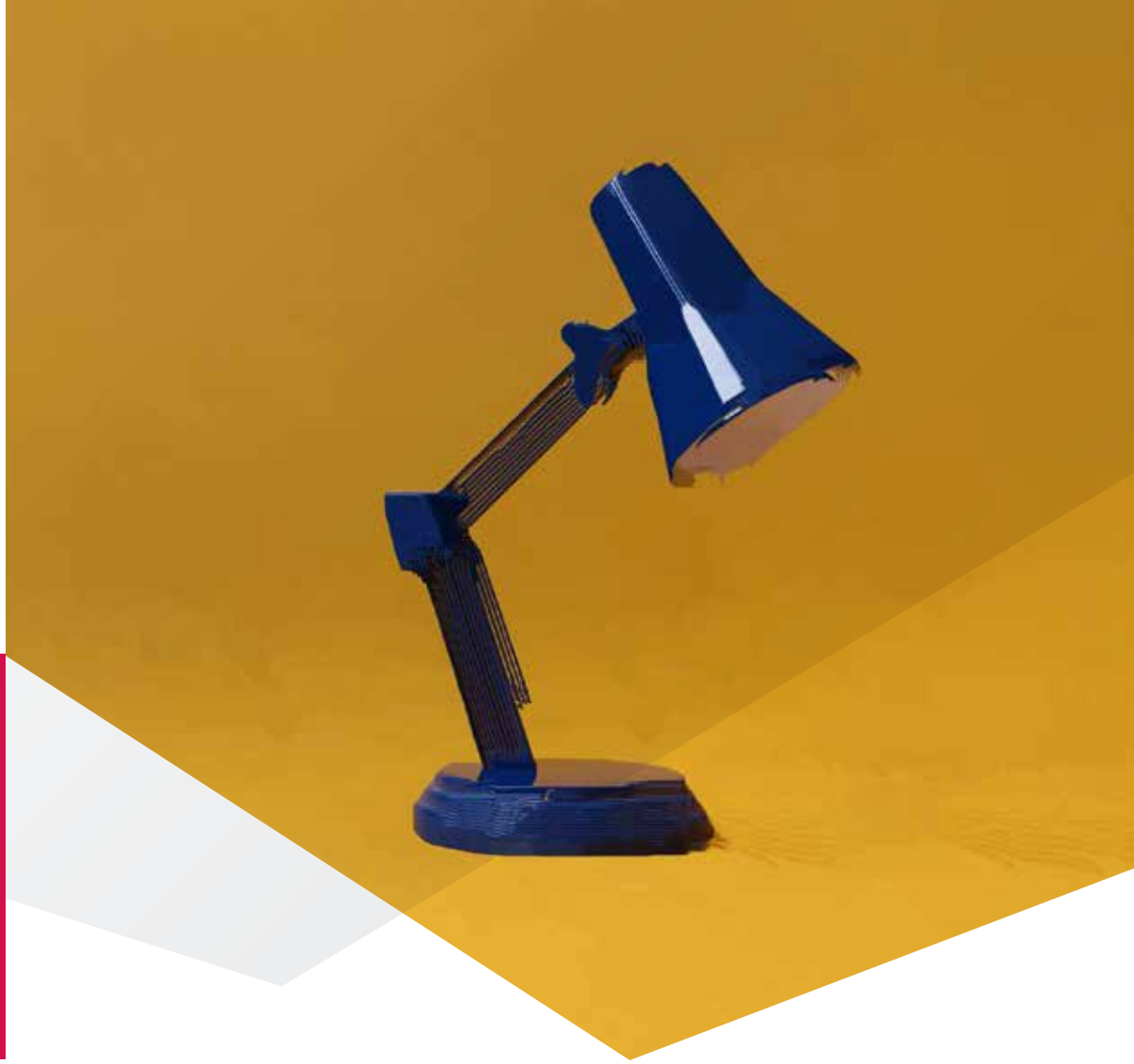
OPENBARE LES



Maaike Harbers

Verstand erbij

Verantwoord ontwerp van toepassingen met kunstmatige intelligentie

Verstand erbij

Verantwoord ontwerp van toepassingen met kunstmatige intelligentie

Hogeschool Rotterdam Uitgeverij

Colofon

ISBN: 9789493012035
1e druk, november 2018
© Maaike Harbers

Afbeeldingen: Christian A. Detweiler

Dit boek is een uitgave van Hogeschool Rotterdam Uitgeverij
Postbus 25035
3001 HA Rotterdam

Publicaties zijn te bestellen via
www.hr.nl/onderzoek/publicaties

Deze publicatie valt onder een Creative Commons
Naamsvermelding-NietCommercieel-GelijkDelen 4.0
Internationaal-licentie.

Verstand erbij

Verantwoord ontwerp van toepassingen met kunstmatige intelligentie

Openbare Les

dr. Maaike Harbers
Lector Artificial Intelligence & Society

Rotterdam, 28 november 2018

INHOUDSOPGAVE

Voorwoord

Toen ik begon met studeren, ging ik ervan uit dat wetenschappelijk onderzoek objectief en waardevrij is, en als doel heeft om de waarheid boven water te krijgen. Het liefst wilde ik zoveel mogelijk leren en weten over die waarheid. Omdat ik ergens moest beginnen en nieuwsgierig was naar wat mensen denken en waarom mensen doen wat ze doen, besloot ik te beginnen ik met het bestuderen van intelligentie en gedrag. Aan de Rijksuniversiteit Groningen volgde ik bij psychologie vakken over hoe hersenen werken en hoe mensen denken, bij mijn studie kunstmatige intelligentie leerde ik hoe je menselijk denken en gedrag kunt nabootsen met computers, en bij mijn studie filosofie leerde ik denken over denken, bewustzijn en de vrije wil.

In de loop van de jaren is mijn beeld van waardevrije wetenschap veranderd. Een eerste barst in mijn geloof in objectief onderzoek ontstond tijdens mijn afstuderen bij kunstmatige intelligentie. Aan het *Instituto de Investigación en Inteligencia Artificial* in Barcelona deed ik onderzoek naar automatisch onderhandelen en inschatten welke onderhandelingspartners wel en niet te vertrouwen zijn. Op een gegeven moment vroeg ik mijn begeleider op basis waarvan ik moest bepalen welke factoren er in mijn model meegenomen moesten worden. Zijn antwoord was: "*What you think makes sense.*" De realisatie dat ík dit moest bepalen, ervoer ik als een enorme ontgoocheling. Ik wilde objectief onderzoek doen, maar het leek mij dat ikzelf verre van een objectieve maatstaf was. Dat mijn begeleider enthousiast was over het werk dat ik afleverde, hielp niet. Ik was toch niet degene die de waarheid in pacht had?

Ondanks deze teleurstelling besloot ik om na mijn studie toch verder te gaan met onderzoek. Vanuit een promotieplek bij TNO en de Universiteit Utrecht over *explainable AI* in virtuele trainingssystemen startte ik opnieuw met de ambitie zo objectief mogelijk onderzoek te doen. Dat leek mij nog steeds de enige juiste manier om wetenschap te bedrijven. Nu had ik jaren de tijd, dus dat zou toch moeten lukken. Toen ik me min of meer had ingelezen in het onderwerp, bleek het lastiger dan ik had gedacht om iets te gaan doen zonder het gevoel te hebben 'maar wat uit te proberen'. Was dat nou de manier om dichterbij de waarheid te komen? Naarmate de tijd vorderde, maakte ik me steeds minder druk over het vinden van de waarheid en steeds meer over het op tijd afkrijgen van mijn

proefschrift. Ik ging inderdaad 'dingen uitproberen' en deze meer pragmatische aanpak leverde gelukkig wel degelijk interessante resultaten op.

Ondertussen leerde ik de academische wereld beter kennen. Ik zag dat daar, naast inhoudelijke overwegingen, veel meer factoren meespelen bij het vormgeven van onderzoek. Het verzinnen van een theoretisch model is bijvoorbeeld een snellere manier om tot een publicatie te komen dan het uitvoeren van een experiment, voor exacte wetenschappen is meer onderzoeksgeld beschikbaar dan voor geesteswetenschappen, en sommige onderwerpen zijn 'hot' en daarmee is het gemakkelijker scoren. Dat riep vragen bij me op: Moet je hier als onderzoeker rekening mee houden? Ook als je denkt dat het interessanter is om ander onderzoek te doen? En zelfs als dat ten koste gaat van de kwaliteit van je onderzoek? Dat zijn bepaald geen objectieve keuzes, maar onvermijdelijk als wetenschapper in deze tijd.

Eenmaal gepromoveerd ging ik als postdoc onderzoeker aan de slag bij de Technische Universiteit Delft. Daar leerde ik een nieuw perspectief kennen, dat van ontwerpers. Hoewel ik in feite al langere tijd bezig was met het ontwerpen van intelligente systemen, leerde ik aan de TU Delft pas echt wat ontwerpen is. Ontwerpen doe je altijd om een bepaald doel te bereiken – dat doel kan zo groot zijn als het oplossen van het klimaatprobleem of zo klein als het handig inrichten van een besteklade. Als ontwerper kies je welk doel je wilt bereiken en daarmee geef je richting aan hoe de wereld eruit gaat zien, al is de bijdrage nog zo klein.

Het drong steeds meer tot me door dat, net als ontwerp, ook veel onderzoek iets aan de wereld verandert, toegepast onderzoek al helemaal. Als onderzoeker kies je welk onderzoek je wel en niet uitvoert en dus hoe je de wereld beïnvloedt. Daarmee is onderzoek niet waardevrij. Ik werkte bijvoorbeeld mee aan het project *Military Human Enhancement* waarin mijn collega's en ik onderzochten hoe intelligente technologie zodanig kan worden ontworpen en ingezet dat militairen altijd verantwoordelijk blijven voor wat die technologie doet. Voor dat onderzoek was het nodig te bepalen wanneer iemand verantwoordelijkheid kan nemen en op welke manier je dat meet. Uiteraard maakten we deze keuzes pas nadat we ons goed hadden geïnformeerd, maar ze waren niet objectief of waardevrij.

Dingen vielen voor mij nog meer op z'n plek toen ik kennismaakte met *Value Sensitive Design* (VSD), een ontwerpaanpak waarbij tijdens het ontwerpproces expliciet rekening wordt gehouden met menselijke waarden. VSD hielp me te zien dat als onderzoek en ontwerp niet waardevrij zijn, je maar beter zo goed mogelijk na kan denken over wat jouw waarden zijn en welke waarden je met je werk wilt nastreven. VSD bood ook technieken en methodes om hiermee aan de slag te gaan. Ik leerde over VSD, paste het toe en hielp het verder te ontwikkelen – met als

hoogtepunt een bezoek in Seattle, waar ik de kans kreeg om aan de *University of Washington* een aantal maanden met de grondleggers van VSD samen te werken.

Aan de TU Delft werkte ik aan een onderwerp waar ik meer dan ooit achter stond: onderzoek naar het meenemen van waarden bij het ontwerpen van toepassingen met kunstmatige intelligentie. Maar met mijn nieuwe overtuiging dat onderzoek naast het creëren van kennis ook dingen kan veranderen, vond ik dat mijn onderzoek niet snel genoeg tot veranderingen leidde. Ik kreeg de kans om bij Hogeschool Rotterdam aan hetzelfde onderwerp te werken, maar dan dichter bij de praktijk. Deze kans heb ik met beide handen aangegrepen. Ik ging lesgeven bij de opleiding *Creative Media and Game Technologies* en praktijkgericht onderzoek doen bij Kenniscentrum *Creating 010*. Dat deed ik eerst als docent-onderzoeker, toen als hoofddocent, en sinds september 2017 combineer ik het hoofddocentschap met de functie lector *Artificial Intelligence & Society*.

Op deze plek zijn nieuwe vragen naar boven gekomen. In mijn huidige positie gaat het niet alleen om mijn waarden als onderzoeker, maar ook over hoe zich dat verhoudt tot het onderwijs en de beroepspraktijk. Hoe kunnen ontwerpstudenten leren om verantwoorde ontwerpkeuzes te maken over toepassingen met kunstmatige intelligentie? Wat moeten docenten doen om studenten daarin te begeleiden? Hoe sluit dat aan bij de werkwijze van commerciële bedrijven? Hoe ver gaat de verantwoordelijkheid van studenten, docenten en professionals? Wat is daarin wenselijk? En wat is realistisch? Deze vragen zijn steeds actueler geworden doordat ontwikkelingen in kunstmatige intelligentie in de afgelopen jaren een vlucht hebben genomen en de impact van kunstmatige intelligentie op de samenleving enorm is gegroeid.

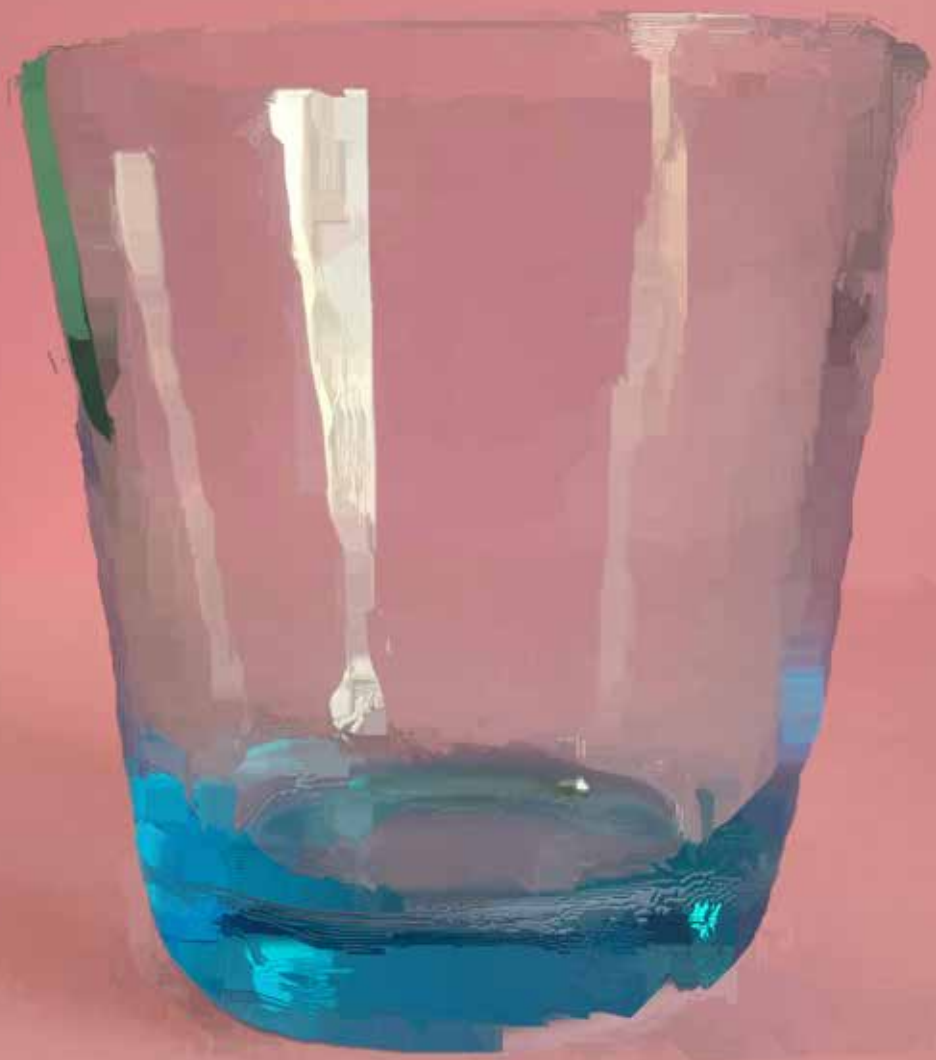
Bovenstaande vragen hebben niet altijd eenduidige antwoorden en standpunten erover kunnen verschillen. Dat maakt de vragen niet minder relevant en ik ben ervan overtuigd dat het belangrijk is om ze wel te stellen. Het doel daarvan is niet zozeer om tot antwoorden te komen waar iedereen het over eens is, maar om het gesprek erover aan te gaan. Dit kan studenten, docenten en andere professionals inzicht geven in de consequenties van verschillende keuzes, waardoor ze die keuzes bewuster kunnen maken. Deze openbare les biedt achtergrond om dat gesprek over verantwoord ontwerpen van toepassingen met kunstmatige intelligentie te voeren en het geeft een eerste aanzet voor het ontwikkelen van handvatten die ontwerpers ondersteunen bij het maken van bewuste keuzes.

Ik prijs me gelukkig te werken op een plek waar ik volop de ruimte krijg om me met deze onderwerpen en vragen bezig te houden, samen met studenten, docenten, onderzoekers, bedrijven en overheden. Voor dit alles wil ik verschillende mensen bedanken. Allereerst wil ik het College van Bestuur bedanken voor het vertrouwen

en het bieden van deze kans. Collega's van het CMGT-team, bedankt voor alle collegialiteit en gezelligheid, door jullie heb ik me vanaf de eerste dag thuis gevoeld in Rotterdam. Collega's bij Creating 010, dank voor alle bevlogenheid en nieuwsgierigheid, jullie maken het kenniscentrum een inspirerende plek om onderzoek te doen. Ook wil ik alle studenten bedanken voor de vele eerlijke reacties en creatieve ideeën.

De ideeën voor deze openbare les zijn medegevormd door de gesprekken en samenwerking met verschillende collega's. Bij CMI en Creating 010 wil ik daarvoor in het bijzonder bedanken Komala, Nathalie, Peter van W., Isabella, Raul, Liane, Ruben, Rob Z., Emiel, Mortaza, Sunil, Peter T. en Paul. Daarbuiten wil speciaal bedanken Marieke, Mark, Joachim, Dave, Batya, Susan en Jasper. Heleen, Paul, Elin en Jan, veel dank voor het meelezen en de nuttige feedback op (delen van) eerdere versies van deze openbare les.

Tot slot, lieve Chris, wil ik jou bedanken. Ik ben erg blij met de mooie afbeeldingen die je hebt gemaakt. Ze laten zien dat slim gebruik van kunstmatige intelligentie tot bijzondere resultaten kan leiden. Dank ook voor je scherpe blik en creatieve ideeën bij het lezen van de tekst en je warme aanmoedigingen tijdens het schrijven. Je inspireert en steunt me bij de dingen die ik doe. Ik ben dankbaar voor ons samenzijn.



HOOFDSTUK 1

Inleiding

Kunstmatige intelligentie of *artificial intelligence* (AI) is in korte tijd ons dagelijks leven binnengekomen. Zoekmachines helpen ons met zoeken naar relevante informatie, spamfilters houden ongewenste email buiten ons zicht, we krijgen gepersonaliseerde aanbevelingen voor nieuwsartikelen (Facebook), muziek (Spotify), filmpjes (Youtube) en series (Netflix), onze digitale foto's worden automatisch voor ons geclassificeerd, de thermostaat leert zelf hoe het ons huis tot een aangename temperatuur verwarmt en bij sommigen van ons wordt het huis schoongehouden door een stofzuigerrobot. Marktonderzoek voorspelt dat kunstmatige intelligentie de komende jaren wereldwijd miljarden dollars aan bedrijfswaarde zal genereren (Lovelock, Tan, Hare, Woodward & Priestley, 2018). In het nieuws komen voortdurend berichten met nieuwe doorbraken voorbij; zo is er zorgtechnologie met kunstmatige intelligentie die betere diagnoses stelt dan menselijke artsen (Van Noort, 2018) en lijkt de zelfrijdende auto steeds dichterbij te komen (Van de Weijer, 2018).

Naast dat kunstmatige intelligentie soms indrukwekkende prestaties levert, roept het fenomeen ook lastige vragen op. Welke taken willen we wel en niet aan kunstmatige intelligentie overlaten? Wie is er verantwoordelijk als er iets misgaat? Is het oké om ons emotioneel te hechten aan een robot? Welke data mogen AI-toepassingen allemaal over ons verzamelen en wat betekent dat voor onze privacy? Mag kunstmatige intelligentie mensen profileren en op basis daarvan verschillend behandelen? Ethische bezwaren rondom kunstmatige intelligentie krijgen steeds meer aandacht in het nieuws, de politiek en het bedrijfsleven (Kool, Dujso & Est, 2018). Een aantal jaar geleden is in de VS bijvoorbeeld het *Future of Life Institute* opgericht door onder anderen Elon Musk (CEO van SpaceX en Tesla), Stephen Hawking (natuurkundige) en Jaan Tallinn (oprichter van Skype) met als doel om risico's van kunstmatige intelligentie voor de mensheid te beperken (futureoflife.org).

Dit zijn nogal wat ontwikkelingen. De vele, soms tegenstijdige informatie over een nogal complexe technologie maakt het lastig om een goede inschatting te maken van wat er nou eigenlijk speelt. Is kunstmatige intelligentie een hype die wel weer overwaait of moeten we ons echt zorgen gaan maken? Wat kan kunstmatige

intelligentie eigenlijk en hoe zal dat zijn over een paar jaar? Wat vormt de grootste bedreiging van kunstmatige intelligentie? En hoe kan die dreiging buiten de deur gehouden worden? In deze openbare les zal ik ingaan op deze vragen en om dat te doen, begin ik met een kijkje in de toekomst.

1.1 Een toekomst met kunstmatige intelligentie

Niemand weet precies hoe de toekomst eruit gaat zien, maar veel ontwikkelingen wijzen erop dat de toepassing van kunstmatige intelligentie in de samenleving voorlopig zal blijven toenemen. Kunstmatige intelligentie wordt nu al op veel plekken succesvol toegepast, wat grote besparingen en nieuwe mogelijkheden oplevert. Door deze successen investeren veel bedrijven in het toepassen kunstmatige intelligentie, waardoor bestaande AI-toepassingen worden verbeterd en nieuwe toepassingen worden ontwikkeld. Naar alle verwachting zal kunstmatige intelligentie dus steeds meer worden ingezet en zullen AI-toepassingen steeds meer taken zelfstandig kunnen uitvoeren.

Minder duidelijk is op welke manier wij in de toekomst met deze mogelijkheden van kunstmatige intelligentie zullen omgaan. Kunstmatige intelligentie kan voor veel verschillende doeleinden en op veel verschillende manieren worden ingezet. Keuzes die hierin gemaakt worden, bepalen hoe de toekomst eruit gaat zien. Er zijn veel verschillende toekomstscenario's denkbaar. Hieronder volgen er twee, waarvan het eerste scenario een prettig en het tweede een minder prettig beeld schetst van een toekomst met kunstmatige intelligentie.

In het eerste scenario wordt kunstmatige intelligentie ingezet om het welzijn van mensen, dieren en de planeet te verbeteren. In de gezondheidszorg zou kunstmatige intelligentie zorgen voor betere diagnoses, secuurdere operaties, de ontdekking van nieuwe medicijnen en het ontlasten van verzorgend personeel van zwaar en eentonig werk. Steden zouden duurzamer worden dankzij een efficiëntere inrichting van logistiek, het ophalen van vuilnis en het verdelen van water en energie. Transport zou veiliger worden en toegankelijker voor ouderen en mensen met een beperking. Doordat voertuigen efficiënter ingezet zouden worden ingezet, zouden er minder voertuigen nodig zijn en ruimte ontstaan voor meer groen in de stad. Mensen zouden gezonder oud worden, minder zwaar, gevaarlijk of saai werk hoeven uit te voeren en meer tijd overhouden voor zaken die er echt toe doen.

Om er in dit scenario voor te zorgen dat kunstmatige intelligentie in dienst blijft staan van mensen, ook bij veranderende omstandigheden, zouden gebruikers van intelligente toepassingen zelf de controle hebben over wat AI-toepassingen doen. Dat vereist dat mensen en machines goed met elkaar kunnen communiceren. Robots en andere intelligente toepassingen zouden in dit scenario altijd begrijpen wat mensen bedoelen en alleen die dingen doen waarvan mensen willen dat ze het

doen. Mensen zouden ook altijd aan deze toepassingen kunnen vragen waar ze mee bezig zijn en waarom, en waar nodig, halverwege kunnen bijsturen. Hierdoor zouden mensen inzicht en controle houden op wat technologie doet en waarom. Beslissingen met grote consequenties blijven mensen in dit scenario zelf nemen.

In het tweede toekomstscenario wordt kunstmatige intelligentie met name ingezet om kosten te besparen. Daarbij zou ernaar worden gestreefd dat kunstmatige intelligentie zoveel mogelijk taken volledig van mensen overneemt. Een paar grote technologiebedrijven zouden intelligente toepassingen en robots ontwikkelen voor alle sectoren, zoals gezondheidszorg, energie, transport, onderwijs, rechtspraak en journalistiek. De meeste mensen zouden zelf steeds minder hoeven te doen. Er zou een scala aan intelligente virtuele spellen, werelden en attracties zijn om mensen te vermaken.

In dit scenario is het niet meer nodig om de intelligente toepassingen aan te sturen of te begrijpen, want deze toepassingen zouden zelf bedenken wat er moet gebeuren en ze zouden zelfstandig belangrijke beslissingen nemen. Mensen zouden in hun dagelijks leven steeds afhankelijker worden van die toepassingen. Kennis over de werking van de kunstmatige intelligentie in de toepassingen zou in handen zijn van een paar grote bedrijven, maar ook daar zouden zelfs experts op een gegeven moment niet meer goed begrijpen hoe en waarom AI-toepassingen tot bepaalde uitkomsten komen. Wanneer deze intelligente technologie onze privacy zou schaden, zou discrimineren of gevaarlijke situaties zou creëren, zouden mensen er weinig meer aan kunnen doen. In dit scenario zouden mensen op den duur volledig overgeleverd raken aan de technologie.

Bovenstaande scenario's maken duidelijk dat het voor onze toekomst nogal uitmaakt hoe we kunstmatige intelligentie gaan gebruiken: Welke taken en beslissingen worden er wel en niet aan AI-toepassingen overgelaten? Hoe inzichtelijk zijn toepassingen met kunstmatige intelligentie? Wie heeft de controle over AI-toepassingen? Juist omdat de toepassing van kunstmatige intelligentie nu een enorme groei doormaakt, is dit hét moment om na te denken over hoe we kunstmatige intelligentie willen inzetten en om invloed uit te oefenen op hoe de samenleving met kunstmatige intelligentie er in de toekomst uit gaat zien.

1.2 Makers van AI-toepassingen hebben verantwoordelijkheid

Kunstmatige intelligentie op zichzelf 'doet' niks. Kunstmatige intelligentie krijgt pas betekenis wanneer het wordt gebruikt om een bepaald doel te dienen: in een AI-toepassing. AI-toepassingen zijn digitale apparaten, systemen, producten en diensten die zelfstandig taken kunnen uitvoeren, zoals auto's, robots, smart objects, artificial agents, virtuele assistenten, web- of mobiele applicaties (apps) en aanbevelingssystemen. Welke invloed kunstmatige intelligentie op de samenleving

zal hebben, hangt af van de toepassingen die ermee worden gecreëerd en hoe die worden ingezet. Dat betekent dat degenen die deze toepassingen creëren dus medebepalen welke effecten kunstmatige intelligentie heeft op de samenleving. De makers van AI-toepassingen spelen daarmee een hele belangrijke rol in het richting geven aan hoe de toekomst met kunstmatige intelligentie eruit gaat zien.

Tot voor kort konden alleen experts toepassingen met kunstmatige intelligentie ontwikkelen, maar tegenwoordig kunnen ook mensen met beperktere kennis van kunstmatige intelligentie AI-toepassingen ontwikkelen, doordat er steeds meer softwareprogramma's met kunstmatige intelligentie beschikbaar zijn. Veelgebruikte programma's zijn bijvoorbeeld TensorFlow, Watson, SageMaker en SparkML. Deze programma's kunnen zelfstandig bepaalde taken uitvoeren, zoals het herkennen van objecten in foto's, het omzetten van gesproken taal in tekst of het zelfstandig genereren van kunst. Zo kan een ontwikkelaar bijvoorbeeld een mobiele applicatie maken die spraakcommando's kan opvolgen zonder zelf volledig spraakherkenning te hoeven ontwikkelen. Dergelijke softwareprogramma's maken het creëren van AI-toepassingen steeds toegankelijker voor (kleinere) bedrijven, zelfstandigen en hobbyisten. Er komen dus steeds meer potentiële makers van AI-toepassingen bij, die invloed hebben op de rol van kunstmatige intelligentie in de samenleving.

De invloedrijke rol van makers van AI-toepassingen brengt een verantwoordelijkheid met zich mee (Friedman, 1996; Friedman & Nissenbaum, 1996; Nissenbaum, 2001; Van den Hoven, 2007; Verbeek, 2006; Wallach, 2015). Keuzes van makers van AI-toepassingen tijdens het ontwerp- en ontwikkelproces, kunnen zowel positieve als negatieve ethische gevolgen hebben voor de samenleving. Door tijdens het ontwerp- en ontwikkelproces aandacht te besteden aan de mogelijke ethische implicaties van een toepassing, kunnen makers keuzes maken waarmee ze nadelige gevolgen van kunstmatige intelligentie voorkomen en positieve gevolgen ervan juist bevorderen. Dit is geen gemakkelijke taak. Het is niet altijd even duidelijk of een gevolg positief of negatief is en wat een juiste keuze is, maar dat neemt niet weg dat makers een verantwoordelijkheid hebben om bewuste keuzes te maken (Verbeek, 2006).

Een concrete ontwerpkeuze met ethische gevolgen speelt bijvoorbeeld bij het creëren van een onlinedienst waar met behulp van met kunstmatige intelligentie gepersonaliseerde advertenties worden getoond. Makers van zo'n dienst moeten nadenken over welke persoonsgegevens worden verzameld, opgeslagen en gebruikt om die gepersonaliseerde advertenties aan te kunnen bieden. Dat zou kunnen gaan om naam, leeftijd, geslacht of locatie, maar ook om seksuele geaardheid en geloof. Makers bepalen welke factoren wel en niet mede bepalen welke advertenties iemand te zien krijgt. Ook kiezen zij of de gebruiker de

mogelijkheid heeft om haar profiel in te zien en of ze haar profiel zelf kan aanpassen, bijvoorbeeld wanneer zij een reis maakt naar een land waar homoseksualiteit of een bepaald geloof agressie oproept, of überhaupt, omdat ze dat een prettiger idee vindt.

Een ander voorbeeld van een ontwerpkeuze die invloed heeft op de ethische implicaties van een AI-toepassing is of er een noodknop op een robot moet komen. Zo'n knop zorgt ervoor dat de robot te allen tijde stopt met waar hij mee bezig is, als de gebruiker deze knop indrukt. Aan de ene kant kan dit de veiligheid bevorderen, maar misschien gaat een grote rode knop op het hoofd van de robot ten koste van een vriendelijke uitstraling. Ook zou de knop wantrouwen kunnen oproepen bij gebruikers. Ze zouden kunnen denken: als er zo'n knop is, dan zal dat wel niet voor niks zijn. Dit zou ervoor kunnen pleiten om de knop wat minder prominent aanwezig te laten zijn, of misschien wel helemaal weg te laten. Deze beslissing, die makers van de robot moeten maken, heeft dus gevolgen voor onder andere de veiligheid en het vertrouwen van de gebruiker.

Een laatste voorbeeld betreft de privacy-instellingen van een slimme thermostaat. Zelfs als de gebruikers de mogelijkheid hebben om zelf hun instellingen aan te passen, dan nog kunnen ontwerpers van deze toepassing het gedrag van gebruikers sturen. De meeste gebruikers wijzigen de standaardinstellingen van een systeem niet. Dus wanneer de instellingen van het platform standaard zijn ingesteld op 'informatie delen', dan zullen de meeste gebruikers daarin meegaan. Ook al hebben gebruikers dus de keuze, het ontwerp van een systeem is van sterke invloed op hun gedrag. Het beïnvloeden van gebruikersgedrag in een bepaalde richting heet ook wel *nudgen* (Thaler & Sunstein, 2009). Wanneer de standaard privacy-instellingen van de slimme thermostaat op 'niet delen' staan, worden gebruikers genudged tot privacybeschermend gedrag.

De verantwoordelijkheid van makers om rekening te houden met de gevolgen van de dingen die ze maken geldt voor technologie in het algemeen. Maar omdat de toepassing van kunstmatige intelligentie hard aan het groeien is en de impact ervan groot is, is het juist nu belangrijk om rekening te houden met de implicaties van technologie met kunstmatige intelligentie tijdens het ontwerp- en ontwikkelproces.

1.3 Handvatten voor makers

Makers van AI-toepassingen kunnen worden ondersteund bij het invullen van hun verantwoordelijkheid tijdens het ontwerp- en ontwikkelproces met behulp van handvatten, zoals richtlijnen, methodes en aanpakken. Er bestaan verschillende richtlijnen die makers helpen om rekening te houden met de ethische gevolgen van de toepassingen die ze creëren en om verantwoorde keuzes te maken.

Een van de eerste visies op hoe toepassingen met kunstmatige intelligentie eruit zouden moeten zien en hoe ze zich zouden moeten gedragen, komt van de sciencefiction schrijver Isaac Asimov. Hij dacht in 1942 al na over de gevolgen van intelligente robots, oftewel, machines met kunstmatige intelligentie. Om narigheden te voorkomen, moeten alle robots volgens hem aan de volgende drie wetten, de wetten van Asimov, voldoen (Asimov, 1950):

1. Een robot mag een mens geen letsel toebrengen of door niet te handelen toestaan dat een mens letsel oploopt.
2. Een robot moet de bevelen van een mens uitvoeren, behalve als die opdrachten in strijd zijn met de eerste wet.
3. Een robot moet zijn eigen bestaan beschermen, voor zover die bescherming niet in strijd is met de eerste of tweede wet.

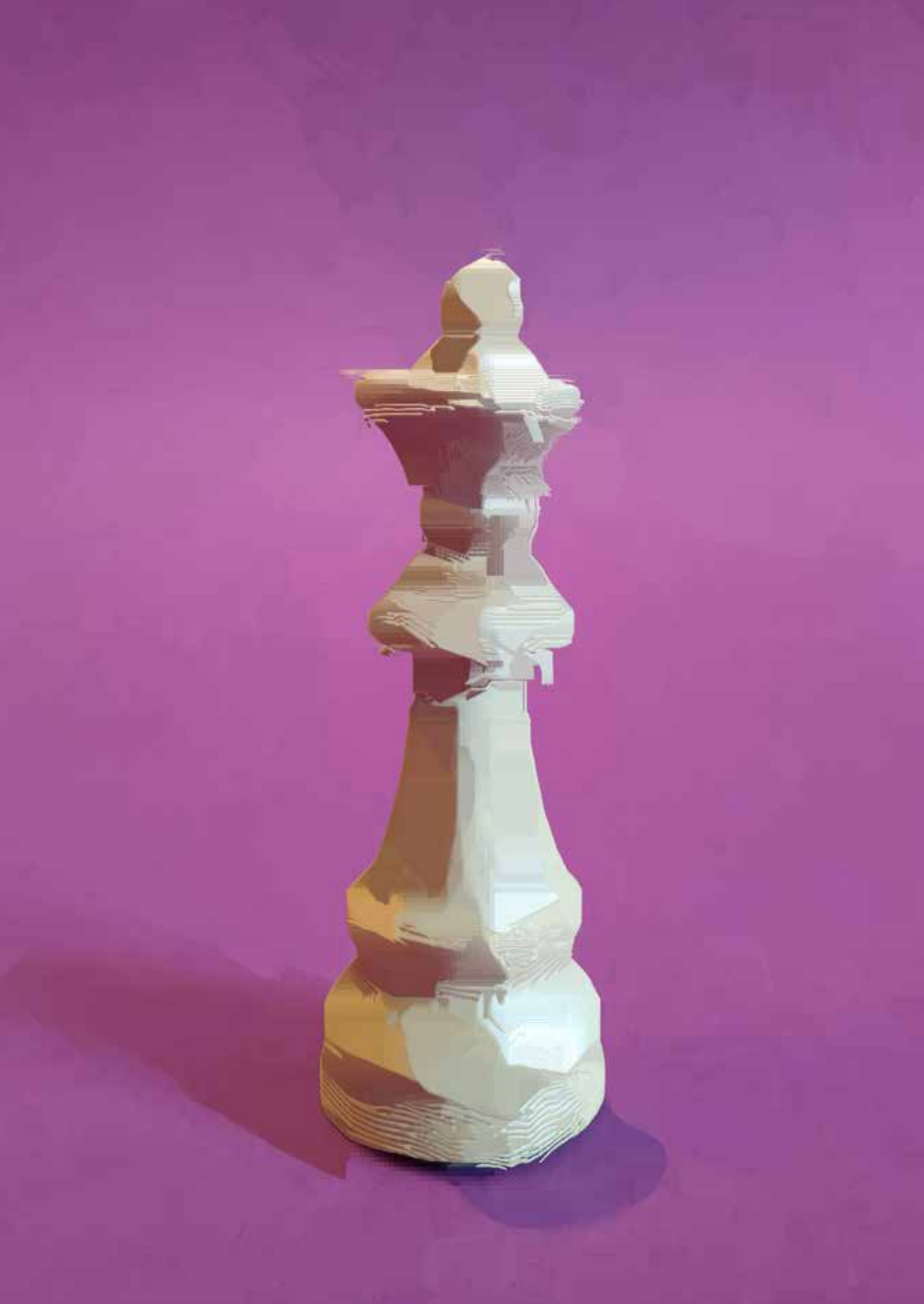
De eerste wet stelt dat robots mensen geen kwaad mogen doen. Het uitgangspunt van de tweede wet is dat mensen de baas zijn over robots, tenzij dat betekent dat een robot een mens schaadt. Dus wanneer een robot bijvoorbeeld de opdracht krijgt van een mens om iemand anders te vermoorden, dan zou hij volgens Asimov's wetten deze opdracht niet mogen uitvoeren. In deze wetten komt pas op de derde plaats dat de robot zichzelf beschermt. Dat betekent dat een robot zichzelf moet opofferen als hij daarmee een mens kan beschermen of als hij daartoe de opdracht krijgt. Deze wetten kunnen gezien worden als richtlijnen waar makers van robots zich aan moeten houden.

Asimov's wetten vormen een mooi begin voor het nadenken over hoe en waarvoor we kunstmatige intelligentie willen toepassen, maar ze zijn niet afdoende. Er zijn tal van voorbeelden van ontwerpkeuzes bij het ontwikkelen van AI-toepassingen waarbij Asimov's wetten geen leidraad bieden. Wat moet een AI-toepassing bijvoorbeeld doen als het onmogelijk is om geen menselijk letsel te veroorzaken? Je zou kunnen zeggen dat een AI-toepassing dan altijd moet kiezen voor de optie die de minst mogelijke schade veroorzaakt. Maar dat is niet altijd even gemakkelijk. Wat is bijvoorbeeld minder erg: dat een zelfrijdende auto een kind raakt of een burgemeester? Andere dilemma's ontstaan wanneer een AI-toepassing, behalve fysiek letsel, ook psychologische schade kan veroorzaken, zoals angst, inbreuk op privacy of beperking van vrijheid. Het is vaak lastig om deze verschillende vormen van schade met elkaar te vergelijken. Toch komen makers van AI-toepassingen regelmatig voor dergelijke vraagstukken te staan en is het onvermijdelijk dat ze hierin keuzes moeten maken.

De laatste jaren is er steeds meer aandacht voor de ethische gevolgen van kunstmatige intelligentie, zowel nationaal als internationaal (Cowls & Floridi, 2018; Kool et al., 2018; Taddeo & Floridi, 2018). In navolging van Asimov zijn er verschillende instanties die principes voor kunstmatige intelligentie in de samenleving hebben geïntroduceerd, bijvoorbeeld de *Asimolar AI Principles* van het *Future of Life Institute* (Asilomar AI Principles, 2017), de *General Principles* van de *Institute of Electrical and Electronic Engineers* (IEEE) (IEEE Global Initiative on Ethics of Autnomous and Intelligent Systems, 2017) en de ethische principes in de *Statement on AI, Robotics and 'Autonomous' Systems* van een werkgroep van de Europese Commissie (European Group on Ethics in Science and New Technologies, 2018). Ook zijn er verschillende onderzoeksprogramma's gericht op ethiek en kunstmatige intelligentie. De Europese Commissie stelt bijvoorbeeld geld beschikbaar in het *Responsible Research and Innovation* programma (European Commission, 2018a) en de Stichting Internet Domeinregistratie Nederland (SIDN) stelt onderzoeksgeld beschikbaar voor onderzoek binnen het thema *Responsible AI* (SIDN, 2018).

Hoewel ethiek en kunstmatige intelligentie meer aandacht krijgen zijn er nog veel open vragen op dit gebied. Er bestaan nog weinig concrete methodes, aanpakken en richtlijnen die makers van AI-toepassingen daadwerkelijk ondersteunen bij het rekening houden met de ethische gevolgen van hun werk. De methodes en technieken die beschikbaar zijn op dit gebied, zijn zelden toegespitst op kunstmatige intelligentie. Ook wordt er bij relevante opleidingen, zoals ontwerp- en informaticaopleidingen, relatief weinig aandacht besteed aan dit onderwerp. Zoals eerder al genoemd, is het vanwege de grote impact van kunstmatige intelligentie op de samenleving, juist bij het creëren van AI-toepassingen belangrijk om rekening te houden met ethische consequenties ervan.

Deze openbare les gaat over het verantwoord ontwerpen van toepassingen met kunstmatige intelligentie. Verantwoord ontwerpen houdt daarbij in dat ontwerpers zich bewust zijn van hun verantwoordelijkheid en rekening houden met de ethische implicaties van hun werk op mensen en de samenleving. In deze les zal ik dieper ingaan op waarom dat belangrijk is en wat ervoor nodig is. Om dat te doen, zal ik eerst een uitleggen wat kunstmatige intelligentie is en de belangrijkste ontwikkelingen in het vakgebied bespreken. Daarna zal ik ingaan op de ethische implicaties van toepassingen met kunstmatige intelligentie. Vervolgens zal ik de bespreken welke rol ontwerp hierin kan spelen. Tot slot zal ik de onderzoeksagenda van het lectoraat Artificial Intelligence & Society presenteren, dat zich de komende jaren zal bezighouden met verantwoord ontwerp van toepassingen met kunstmatige intelligentie.



HOOFDSTUK 2

Kunstmatige intelligentie

De term 'kunstmatige intelligentie' werd voor het eerst gebruikt in 1956, toen een paar Amerikaanse wetenschappers een voorstel indienden voor een onderzoeksproject over kunstmatige intelligentie. Deze wetenschappers veronderstelden dat alle aspecten van menselijk leren en menselijke intelligentie in principe door machines nagebootst zouden kunnen worden. Ze stelden voor om hier gedurende twee maanden met een groep van wiskundigen en informatici over te brainstormen en ideeën te ontwikkelen voor 'denkende machines'. De verwachting was dat er significante vooruitgang geboekt zou worden wanneer een groep van zo'n tien wetenschappers hier een hele zomer aan zou kunnen werken. De deelnemers van het AI-zomerproject waren inderdaad erg enthousiast, maar twee maanden bleken niet genoeg om het probleem van kunstmatige intelligentie op te lossen. In die zomer is het vakgebied van de kunstmatige intelligentie ontstaan.

Er wordt dus al vanaf de jaren 50 van de vorige eeuw onderzoek gedaan naar kunstmatige intelligentie, maar tot voor kort kwam kunstmatige intelligentie weinig 'in het wild' voor. Het beeld dat veel mensen hebben van kunstmatige intelligentie is daardoor vooral gevormd door films en boeken, zoals bijvoorbeeld *Frankenstein, 2001: A Space Odyssey*, *The Terminator*, *Wall-E*, *Her* en *Ex Machina*. In veel van deze science fiction verhalen ontwikkelt een robot of een intelligent systeem zich tot een wezen met mensachtige eigenschappen en gevoelens. Regelmatig zijn deze slimme wezens erop uit om te heersen over de mensheid of om de wereld te vernietigen. Deze voorstelling van zaken strookt gelukkig niet met de werkelijkheid, maar geeft veel mensen wel een vertekend beeld van kunstmatige intelligentie.

Kunstmatige intelligentie is dus niet een allesvernietigende entiteit, maar wat is het dan wél? In dit hoofdstuk geef ik een introductie in kunstmatige intelligentie en bespreek ik de belangrijkste ontwikkelingen in het vakgebied. Om uit te leggen wat kunstmatige intelligentie is, zal ik eerst ingaan op het begrip intelligentie zelf.

2.1 Wat is intelligentie?

Vele filosofen, psychologen, neurowetenschappers en informatici hebben zich reeds gebogen over de vraag "Wat is intelligentie?", maar een eenduidig antwoord op die vraag is er nog niet. Een van de redenen die het beantwoorden van de vraag

wat intelligentie is zo lastig maakt, is dat de betekenis van het begrip intelligentie relatief is. Of bepaald gedrag als intelligent wordt beschouwd, hangt namelijk af van wie of wat dat gedrag vertoont. Een kleuter die kan lezen beschouwen we bijvoorbeeld als intelligent, maar bij een volwassen persoon vinden we dat niet meer dan normaal. Dolfijnen noemen we intelligente dieren omdat ze gereedschap gebruiken en zichzelf kunnen herkennen in een spiegel. Mensen die datzelfde kunnen, bestempelen we niet per se als intelligent. Iemand die vlot uit het hoofd kan machtsverheffen en worteltrekken vinden we vaak wél intelligent, maar een computer die goed kan rekenen noemen we weer niet intelligent.

Menselijke intelligentie

Van alle vormen van intelligentie is die van mensen het meest bestudeerd (Anderson, 1996). Vrijwel alle psychologen zijn het erover eens dat intelligentie te maken heeft met abstract redeneren, logisch nadenken, problemen oplossen en kennis vergaren. De meeste psychologen beschouwen ook zaken als geheugen, taalvaardigheid, rekenvaardigheid, het kunnen aanpassen aan de omgeving en doelgericht kunnen handelen als onderdeel van intelligentie. Veel, maar niet alle, psychologen vatten onder intelligentie ook creativiteit en emotionele en sociale intelligentie, waarbij emotionele en sociale intelligentie verwijzen naar de capaciteit om eigen en andermans gedachten en gevoelens te kennen, begrijpen en beïnvloeden.

Bovenstaande opsomming laat zien dat intelligentie uit verschillende componenten bestaat. De mate waarin iemand intelligent is kan per onderdeel verschillen. Iemand kan bijvoorbeeld heel goed zijn in rekenen en wiskunde, maar sociaal minder handig. Een ander kan veel talent hebben om muziek te maken en te componeren, maar minder goed zijn in het onthouden van feiten. Bij mensen wordt intelligentie vaak uitgedrukt in een IQ-score. Deze reductie van intelligentie tot één enkele score doet tekort aan de rijkheid van intelligentie. Dat intelligentie uit verschillende componenten bestaat, maakt het lastiger om te definiëren wat intelligentie is. Niet iedereen is het erover eens welke onderdelen er wel en niet bij intelligentie horen.

Een van de discussiepunten over of iets wel of niet bij intelligentie hoort, gaat over bewustzijn. Onze hersenactiviteiten zijn op te delen in bewuste processen en onbewuste processen. Bedenken wat je vanavond wilt eten is bijvoorbeeld een bewust proces waarin je verschillende opties bedenkt (bijvoorbeeld lasagne, Thaise curry en couscous) en de voor- en nadelen tegen elkaar afweegt (wat is sneller klaar, lekkerder of gezonder). Maar wanneer je op de fiets stapt om boodschappen te gaan doen, gebeurt er veel onbewust. Je houdt evenwicht door voortdurend bij te sturen, je neemt je omgeving waar en bepaalt in minder dan een seconde of je nog net wel of niet meer voor een auto langs kunt. De processen voltrekken zich 'automatisch' zonder dat je er heel bewust over na hoeft denken. Toch zou je het resultaat van deze processen, naar de supermarkt fietsen om boodschappen te

doen, intelligent gedrag kunnen noemen. In de praktijk refereert het woord intelligentie echter toch vaak aan bewuste processen. De term cognitie – het vermogen om kennis te verwerven door waarnemen en nadenken – wordt vaak gebruikt voor het geheel aan bewuste en onbewuste processen.

Samenvattend kan gezegd worden dat intelligentie een complex begrip is dat zich lastig laat definiëren. Er is geen duidelijke grens tussen welke eigenschappen er nou wel en niet onder intelligentie vallen, intelligentie is lastig om te meten en intelligentie ziet er ook nog eens anders uit bij verschillende entiteiten met intelligentie. Zo onderscheiden volwassen mensen zich van baby's en dieren doordat ze taal, bewustzijn en een vrije wil hebben. Mensen en dieren onderscheiden zich van computers en machines doordat ze hersenen hebben. Hersenen bestaan uit biologische cellen en de intelligentie van mensen en dieren wordt daarom ook wel natuurlijke intelligentie genoemd. Dat staat tegenover de intelligentie van machines en computers die ontstaat in chips gemaakt van siliconen: artificiële of kunstmatige intelligentie.

De Turingtest

Zoals gezegd is menselijke intelligentie de meest uitgebreid bestudeerde vorm van intelligentie. Ook binnen het vakgebied van kunstmatige intelligentie vormt menselijke intelligentie de belangrijkste maatstaf voor wat intelligentie is. Alan Turing, een succesvolle codekraker tijdens de Tweede Wereldoorlog en pionier in de kunstmatige intelligentie, stelde als eerste de vraag: "Kunnen machines denken?". Om daar antwoord op te geven bedacht hij in de jaren 50 van de vorige eeuw de Turingtest (Turing, 2009). Deze test houdt in dat een vragensteller met iemand in gesprek gaat en moet achterhalen of die 'iemand' een mens of een machine is. Het gesprek vindt plaats door te chatten via een toetsenbord en scherm, zodat de vragensteller alleen maar tekst ziet en niet degene die de vragen beantwoordt. Wanneer een machine zodanig reageert dat de vragensteller niet weet of de ander een mens of een machine is, dan is die machine geslaagd voor de Turingtest en kan die machine volgens Turing denken.

Later zijn er vele varianten bedacht op de Turingtest, waarbij steeds mensen moeten raden of zij te maken hebben met een mens of een machine. Daarbij kan het bijvoorbeeld gaan om een tegenspeler in een computergame, een virtuele assistent of een adviessysteem, al dan niet door mensen bestuurd. Een Turingtest beproeft dus altijd of mensen vinden dat een machine menselijk gedrag laat zien.

Niet iedereen vindt de Turingtest een goede manier om te toetsen of machines kunnen denken of intelligent zijn. Dat een machine zich vóórdoet als mens betekent nog niet dat die machine ook echt denkt zoals mensen denken, of bewustzijn en gevoelens heeft zoals mensen dat hebben. Het is niet zo moeilijk om een robot te

maken die, wanneer de zon schijnt, een glimlach krijgt en zegt dat hij blij is, maar daarmee *voelt* de robot nog niet daadwerkelijk blijdschap. De meeste mensen zouden deze simpele robot dan ook niet direct emoties of intelligentie toeschrijven.

De Chinese kamer

De filosoof John Searle is een beroemde criticus van de Turingtest. Hij is van mening dat een machine die zich gedraagt als een mens niet per definitie intelligent is en om zijn standpunt te verduidelijken bedacht hij een beroemd gedachte-experiment genaamd de Chinese kamer (Searle, 1980).

De Chinese kamer is een afgesloten kamer met daarin een verzameling Chinese naslagwerken en man die geen Chinees begrijpt en spreekt. Via een gleuf krijgt de man briefjes met daarop vragen in het Chinees. In de naslagwerken staat (in het Chinees) een verzameling van alle mogelijke vragen met bij elke vraag een juist antwoord. De man heeft alle tijd van de wereld en vergelijkt de Chinese tekens op de briefjes met die in het boek. Zodra hij de vraag op een briefje in het boek vindt, tekent hij het bijbehorende antwoord uit het boek na op een nieuw briefje. Vervolgens duwt hij dit nieuwe briefje met Chinese tekens door een andere gleuf weer naar buiten. Op deze manier geeft hij steeds goede antwoorden op vragen in het Chinees. Voor de buitenstaander lijkt het alsof de man in de kamer Chinees begrijpt en spreekt, terwijl hij dat in werkelijkheid niet doet.

Het experiment van de Chinese kamer laat zien dat een systeem dat zich gedraagt *alsof* hij taal begrijpt, taal niet daadwerkelijk hoeft te begrijpen. Op dezelfde wijze hoeft een robot die zich gedraagt alsof hij bewustzijn en gevoelens heeft, nog niet de ervaring van bewustzijn en gevoelens te hebben. Maar volgens dezelfde redenering kan een mens eigenlijk van niemand, behalve van zichzelf, weten of hij *echte* gevoelens ervaart. Net als machines kan men bij een mens ook niet anders dan afgaan op zijn of haar gedrag om te bepalen of diegene intelligent is. Misschien komen we er dus wel nooit achter of machines *echt* kunnen denken en ervaren. Volgens de informaticus Edsger Dijkstra is dat ook niet zo belangrijk. Hij noemde de vraag of machines kunnen denken net zo zinnig als de vraag of onderzeeërs kunnen zwemmen. Ook heeft de vraag AI-onderzoekers niet tegengehouden in het ontwikkelen van kunstmatige intelligentie.

2.2 Artificial agents

Het vakgebied kunstmatige intelligentie houdt zich bezig met het creëren van artefacten die intelligentie vertonen, ook wel intelligente of *artificial agents* genoemd (Russell & Norvig, 2003). Deze artificial agents vertonen intelligent gedrag, maar ze 'denken' niet op dezelfde manier als mensen. Soms vinden AI-onderzoekers inspiratie in menselijk denken (zo zijn artificiële neurale netwerken bijvoorbeeld geïnspireerd op hoe hersenen werken), maar dat hoeft

niet. Toch is menselijk denken erg belangrijk voor het vakgebied kunstmatige intelligentie, omdat veel artificial agents interactie hebben met mensen. Om die interactie soepel te laten verlopen, moeten agents bijvoorbeeld emoties of intenties van mensen kunnen herkennen of met mensen kunnen communiceren in natuurlijke taal. Daarom zijn onder andere de vakgebieden psychologie en taalwetenschappen belangrijk voor kunstmatige intelligentie.

In de vorige sectie is beschreven dat menselijke intelligentie uit verschillende onderdelen bestaat. Datzelfde geldt voor kunstmatige intelligentie. Een klein aantal onderzoekers stelt zichzelf tot doel om agents te ontwikkelen die op alle denkbare taken en onderdelen van intelligentie minstens even goed presteren als de mens. Dit wordt *general AI* of *artificial general intelligence* genoemd. General AI bestaat op dit moment niet en het is onzeker of dat überhaupt ooit bereikt zal worden. De meeste AI-onderzoekers houden zich bezig met ontwikkelen van agents die goed zijn in specifieke taken, bijvoorbeeld taal begrijpen, gezichten herkennen of voetballen. Dit heet ook wel *narrow AI* en alle huidige AI-toepassingen vallen hieronder.

Doordat artificial agents in zulke verschillende taken gespecialiseerd zijn, hebben ze vele verschijningsvormen. Ze kunnen een fysiek, een virtueel of helemaal geen lichaam hebben en ze kunnen er meer of minder menselijk uitzien. Een agent met een fysiek lichaam wordt ook wel een robot genoemd. In films, speelgoed of op plaatjes worden vaak robots afgebeeld die er menselijk of in elk geval mensachtig uitzien. Maar er zijn ook heel veel voorbeelden van robots die er niet menselijk uitzien, bijvoorbeeld zelfrijdende auto's, autonome stofzuigers of smart homes. Voorbeelden van artificial agents met een virtueel lichaam zijn game karakters, virtuele coaches of persoonlijke assistenten. Ook deze kunnen er meer of minder menselijk uitzien. De verschijningsvorm van Siri (de spraakassistent van Apple) bestaat slechts uit wat bewegende gekleurde golfjes en een stem. Tot slot zijn er dus agents die helemaal geen lichaam hebben. Voorbeelden hiervan zijn spamfilters, gezichtsherkenningssoftware, text-to-speech engines en agents die aandelen op de beurs kopen en verkopen.

Hoe divers artificial agents er ook uitzien, er zijn twee kenmerken die voor alle agents gelden. Ten eerste bevinden agents zich altijd in een omgeving. Voor agents met een fysiek lichaam (robots) is dat een fysieke omgeving, ofwel, de echte wereld. Voor agents met een virtueel of geen lichaam is dat een virtuele omgeving, bijvoorbeeld een game of een emailprogramma. Ten tweede geldt voor alle agents dat ze hun omgeving kunnen waarnemen, erover kunnen redeneren en de omgeving kunnen beïnvloeden door te handelen.

Waarnemen doen agents met behulp van sensoren zoals camera's, microfoons, thermometers of bewegingssensoren. Zo hoort een virtuele assistent of je tegen

haar praat en merkt een slimme thermostaat hoe warm het is. Sommige virtuele agents kunnen alleen maar digitale informatie waarnemen, bijvoorbeeld doordat ze een tekstberichtje of digitale foto binnenkrijgen. Waarnemen is niet altijd eenvoudig voor artificial agents, omdat data die binnenkomen geïnterpreteerd moeten worden. Interpreteren van data is bijvoorbeeld het omzetten van geluidssignalen naar tekst of het herkennen van een auto in een verzameling pixels.

Redeneren is het 'denken' van artificial agents. Daarbij combineert een agent haar waarnemingen met de kennis die zij al heeft. Bijvoorbeeld, een virtuele assistent hoort iemand zeggen: "Kun je een alarm zetten voor over 8 minuten?" Een virtuele assistent begrijpt niet wat hier gezegd wordt op de manier waarop mensen deze vraag begrijpen. De eerste stap van de assistent is dan ook om de betekenis van deze zin uit de woorden af te leiden. Lastig daarbij is dat er veel verschillende manieren zijn om dezelfde vraag te stellen. Voor de assistent is het niet evident dat de vragen 'Mag er over 8 minuten een alarm afgaan?' en 'Een alarm over 8 minuten graag' dezelfde betekenis hebben. Wanneer de virtuele assistent dan toch heeft afgeleid dat het gaat om een alarm over 8 minuten, moet zij beredeneren hoe laat het alarm moet afgaan. De assistent weet dat het nu 18:45 uur is. Door beide twee gegevens met elkaar te combineren, kan de assistent beredeneren dat zij om 18:53 uur een alarm moet laten afgaan. Er zijn verschillende manieren om het redeneerproces van een agent te programmeren. In de volgende sectie wordt dieper ingegaan op hoe dat redeneren werkt.

Handelen kunnen agents op verschillende manieren doen. Een robot kan bijvoorbeeld voorruit rijden, bliepgeluiden laten horen, met zijn ogen knipperen of een voorwerp oppakken. Handelen hoeft zich niet altijd in de fysieke wereld af te spelen. Het kan bijvoorbeeld ook gaan om het versturen van een bericht, een online aankoop doen of het geven van een advies of aanbeveling.

2.3 Verschillende benaderingen van kunstmatige intelligentie

Het vakgebied kunstmatige intelligentie houdt zich bezig met het maken van artificial agents die kunnen waarnemen, redeneren en handelen. Sinds het ontstaan heeft het vakgebied vele ontwikkelingen doorgemaakt, waarin steeds verschillende technieken centraal stonden en waarin onderzoekers zich richtten op verschillende problemen. In deze sectie worden kort drie van de belangrijkste benaderingen in het vakgebied besproken: symbolische kunstmatige intelligentie, belichaamde cognitie en machine learning.

Symbolische kunstmatige intelligentie

Een manier om kennis te representeren in een computerprogramma is met behulp van regels en logica (Stefik, 2014). Kennis over 'wat een vogel is' kan bijvoorbeeld

worden gerepresenteerd met de regel: een vogel is een dier met twee vleugels, twee poten en een snavel en kan vliegen. Het computerprogramma kan vervolgens met logica afleiden dat een dier met vier poten en een slurf geen vogel is. Wanneer je heel veel regels aan het computerprogramma toevoegt, kan het programma steeds meer vragen goed beantwoorden en wordt het op die manier steeds intelligenter. Een belangrijke uitdaging bij deze aanpak is om, wanneer een computerprogramma uit heel erg veel regels bestaat, het programma op een efficiënte manier naar de juiste regel te laten zoeken. Deze benadering om intelligentie te creëren valt onder de symbolische kunstmatige intelligentie, omdat het gebruikmaakt van abstracte, 'symbolische' representaties van de werkelijkheid.

De meest succesvolle toepassing van symbolische kunstmatige intelligentie is die van expertsystemen. Daarbij wordt zoveel mogelijk kennis van menselijke experts, bijvoorbeeld dokters of hypotheekadviseurs, in 'als-dan' regels gevat. Bijvoorbeeld, *als* iemand last heeft van niezen, hoesten, een verstopte neus, hoofdpijn en lichte vermoeidheid, *dan* duidt dat mogelijk op verkoudheid. Door een verzameling van zulke algemene kennisregels toe te passen op een specifiek geval, is een expertsysteem in staat om een diagnose te stellen, advies te geven of een voorspelling te doen. Expertsystemen zijn regelmatig commercieel toegepast. Een belangrijke bevinding daarbij was dat mensen adviezen en diagnoses van expertsystemen alleen overnemen als ze gepaard gaan met een uitleg over hoe het systeem tot een conclusie is gekomen (Clancey, 1983).

Er is veel vooruitgang geboekt binnen de symbolische kunstmatige intelligentie, maar deze stroming kent ook een aantal belangrijke beperkingen. Het representeren van kennis in symbolen en regels is moeilijk en kost erg veel tijd, mede doordat experts vaak moeite hebben om hun kennis expliciet te maken. Ook hebben symbolische, logische benaderingen moeite met het omgaan met uitzonderingen en onzekerheid. Zo bestaan er vogels die niet kunnen vliegen, zoals pinguïns en struisvogels, en leggen sommige vogels geen eieren omdat ze toevallig onvruchtbaar zijn. Het is ondoenlijk om voor alle mogelijke situaties nieuwe kennisregels op te stellen, terwijl de wereld vol zit met uitzonderingen en onzekerheden. In het midden van de jaren 80 van de vorige eeuw begonnen onderzoekers daarom naar andere benaderingen te kijken. Inmiddels wordt symbolische kunstmatige intelligentie ook wel GOFAI genoemd, *Good Old-Fashioned Artificial Intelligence* (Haugeland, 1989). Symbolische kunstmatige intelligentie wordt nog steeds toegepast, maar meestal in combinatie met andere benaderingen.

Belichaamde cognitie

Aan het eind van de jaren 80 ontstond een tegenreactie op symbolische kunstmatige intelligentie, met Rodney Brooks als belangrijkste vertegenwoordiger.

Brooks is van mening dat het niet nodig is om kennis te vatten in expliciete regels en symbolen en dat de meeste intelligentie juist ontstaat door de interactie van een entiteit met haar omgeving (Brooks, 2002). In plaats van logica en redeneren zou onderzoek naar kunstmatige intelligentie zich volgens hem moeten richten op robots in de echte wereld. Schaken werd lange tijd als een van de belangrijkste uitdagingen van kunstmatige intelligentie gezien, maar Brooks vond voetbal een veel interessantere uitdaging voor kunstmatige intelligentie. Schaken vereist alleen maar abstract nadenken, voetbal vereist daarnaast ook nog eens snelheid, coördinatie en timing.

Voor Brooks is het dus belangrijk dat een AI-systeem een 'lichaam' heeft en ontstaat intelligentie door de interactie van een systeem met de omgeving. Een voorbeeld hiervan is een mier die een geurenspoor achterlaat, waardoor hij anderen de weg wijst naar voedsel en hij zelf de weg kan terugvinden naar de mierenhoop. De mier 'weet' niet waar hij is en 'vertelt' andere mieren niet waar ze heen moeten op de manier waarop mensen dat weten en doen, maar toch is de mier in staat om intelligent gedrag te vertonen. Dit wordt ook wel belichaamde cognitie genoemd. Brooks heeft met zijn collega's verschillende robots ontwikkeld volgens dit principe. Eerst als onderzoeker, later als oprichter van het bedrijf iRobot dat de bekende stofzuigerrobot Roomba op de markt heeft gebracht. Zijn ideeën bleken later vooral geschikt voor relatief simpel robotgedrag, maar zijn wel van grote invloed geweest op de robotica en kunstmatige intelligentie.

Een belangrijk deelgebied binnen robotica is dat van de sociale robotica (Breazeal, 2004). Sociale robots zijn robots die communiceren en interactie hebben met mensen door sociaal gedrag te vertonen. Dit kan zijn doordat deze robots natuurlijke taal spreken, maar dat hoeft niet: sociale interactie kan ook plaatsvinden door het vertonen van gezichtsuitdrukkingen, gebaren en lichaamshoudingen die een bepaalde emoties, gemoedstoestand of intentie uitdrukken. Het uiterlijk van sociale robots heeft vaak menselijke kenmerken, zoals een hoofd, lichaam, ogen of oren. Dit maakt het gemakkelijker om deze robots op sociale wijze met mensen te laten communiceren.

Onderzoek naar robotica blijft een van de meest lastige disciplines binnen de kunstmatige intelligentie. In tegenstelling tot in de virtuele wereld is het in de echte wereld lastig om de omgeving te controleren, terwijl de omgeving juist in de echte wereld complex en onvoorspelbaar is. Robots kunnen indrukwekkende taken verrichten, maar evenaren de mens op veel vlakken bij lange na nog niet.

Machine learning

Machine learning is een deelgebied van kunstmatige intelligentie dat zich bezighoudt met het ontwikkelen van software die kan leren van ervaringen (Domingos, 2015;

Witten, Frank, Hall & Pal, 2016). De term 'algoritme' of 'zelflerend algoritme' wordt vaak gebruikt als het over machine learning gaat. Een algoritme is een reeks van instructies die van een beginpunt naar een doel leiden. Het algoritme om een temperatuur uitgedrukt in graden Fahrenheit om te rekenen naar graden Celsius is bijvoorbeeld: trek 32 van het begingetal af en deel het resultaat vervolgens door 1,8. Dit omrekenalgoritme van graden Fahrenheit naar Celsius is echter te simpel om van kunstmatige intelligentie te kunnen spreken. Maar naarmate algoritmes ingewikkelder worden en bijvoorbeeld ook kunnen leren, kunnen ze steeds moeilijkere taken volbrengen en daarmee steeds intelligenter gedrag vertonen.

Met behulp van machine learning kan een algoritme bijvoorbeeld handgeschreven tekst leren herkennen. Bij aanvang ziet handgeschreven tekst er voor het algoritme uit als wat nietszeggende lijnen en figuren. Het algoritme moet nu leren om te herkennen welke figuren bij welke letter horen. Dat kan door het algoritme te 'trainen'. Trainen bestaat eruit het algoritme heel veel voorbeelden te laten zien van handgeschreven tekst met het 'juiste antwoord' (digitale tekst) erbij. Na elk voorbeeld past het algoritme zichzelf een beetje aan en gaat het langzaam herkennen welke combinatie van lijnen een 'a', 'b' of een 'c' voorstellen. Er hoeft dus niet expliciet aan het algoritme verteld te worden dat een 'i' eruitziet als een verticale streep met een punt erboven, dat leert het zelf. Deze manier van leren lijkt op hoe kleine kinderen leren. Door veel te oefenen en bij anderen af te kijken leren kinderen kruipen, lopen en praten zonder dat ze expliciet verteld is hoe dat moet. Hoe meer voorbeelden (data) het algoritme te zien krijgt, des te beter het een taak leert uit te voeren. Voor machine learning zijn dus veel data nodig.

Er bestaan verschillende vormen van machine learning. Bij *supervised learning* zijn de trainingsdata van het algoritme 'gelabeld', dat wil zeggen dat de voorbeelden vergezeld zijn van het 'juiste antwoord'. Bij *unsupervised learning* zijn de trainingsdata 'ongelabeld'. Het algoritme gaat dan zelf op zoek naar structuur in de data door te kijken welke voorbeelden meer en minder op elkaar lijken. Een laatste vaak voorkomende vorm van machine learning is *reinforcement learning*. Daarbij heeft een computerprogramma interactie met een omgeving en moet het een bepaald doel in die omgeving bereiken (bijvoorbeeld een computerspelletje winnen). Het programma weet of het zijn doel wel of niet bereikt, maar moet zelf ontdekken wat de beste strategie is om dat doel te bereiken. Het kan een uitdaging zijn om voldoende (gelabelde) data te vinden om algoritmes te trainen.

Lerende algoritmes hebben verschillende vormen. Een populaire benadering binnen machine learning zijn artificiële neurale netwerken. Deze netwerken zijn geïnspireerd op hoe onze hersenen werken en bestaan uit een netwerk van 'knopen' (neuronen) en verbindingen tussen die knopen. Bij elke verbinding wordt met een getal aangegeven hoe sterk de verbinding tussen de betreffende twee

knopen is. Tijdens het trainen worden deze verbindingen sterker of zwakker. Kennis wordt in een neuraal netwerk gerepresenteerd door de verzameling van verbindingen en getallen. Neurale netwerken kunnen complexere problemen oplossen wanneer ze meer 'lagen' hebben (groter zijn). Het leren van neurale netwerken met veel lagen heet *deep learning*.

Er wordt al decennialang onderzoek gedaan naar machine learning, maar de afgelopen jaren is indrukwekkende vooruitgang geboekt. Bekende voorbeelden zijn IBM's Watson die in 2011 de beste menselijke spelers van de televisiequiz Jeopardy overwon (Best, 2013) en DeepMind's Alpha Go die in 2016 de wereldkampioen van het bordspel Go versloeg (DeepMind, 2018). Go is vele malen complexer dan schaken en bijna niemand had deze overwinning zo snel verwacht. Andere voorbeelden van toepassingen met machine learning zijn het herkennen van afwijkingen op röntgenfoto's en MRI-scans, het herkennen van objecten en gezichten in plaatjes of videobeelden en het leren kennen van iemands voorkeuren en interesses en op basis daarvan aanbevelingen doen voor producten, films of berichten.

De successen van machine learning zijn te danken aan een aantal ontwikkelingen. Ten eerste zorgden nieuwe inzichten in deep learning ervoor dat algoritmes effectiever kunnen leren. Ten tweede zijn er door goedkopere sensoren en dataopslag de laatste jaren volop data beschikbaar, een belangrijke voorwaarde voor machine learning. Ten derde kost het trainen van algoritmes veel *processing power* en ook die is veel goedkoper geworden. Machine learning wordt tegenwoordig volop toegepast: van spamfilters tot gezichtsherkenning en van gepersonaliseerde aanbevelingen tot automatisch vertalen.

Een nadeel van machine learning is dat systemen met machine learning erg ondoorzichtig zijn. Doordat kennis niet op één plek gerepresenteerd wordt (maar bijvoorbeeld als een neuraal netwerk), is het lastig te traceren hoe een algoritme tot een bepaalde uitkomst komt. Dit maakt het lastiger om de betrouwbaarheid van een resultaat goed in te schatten.

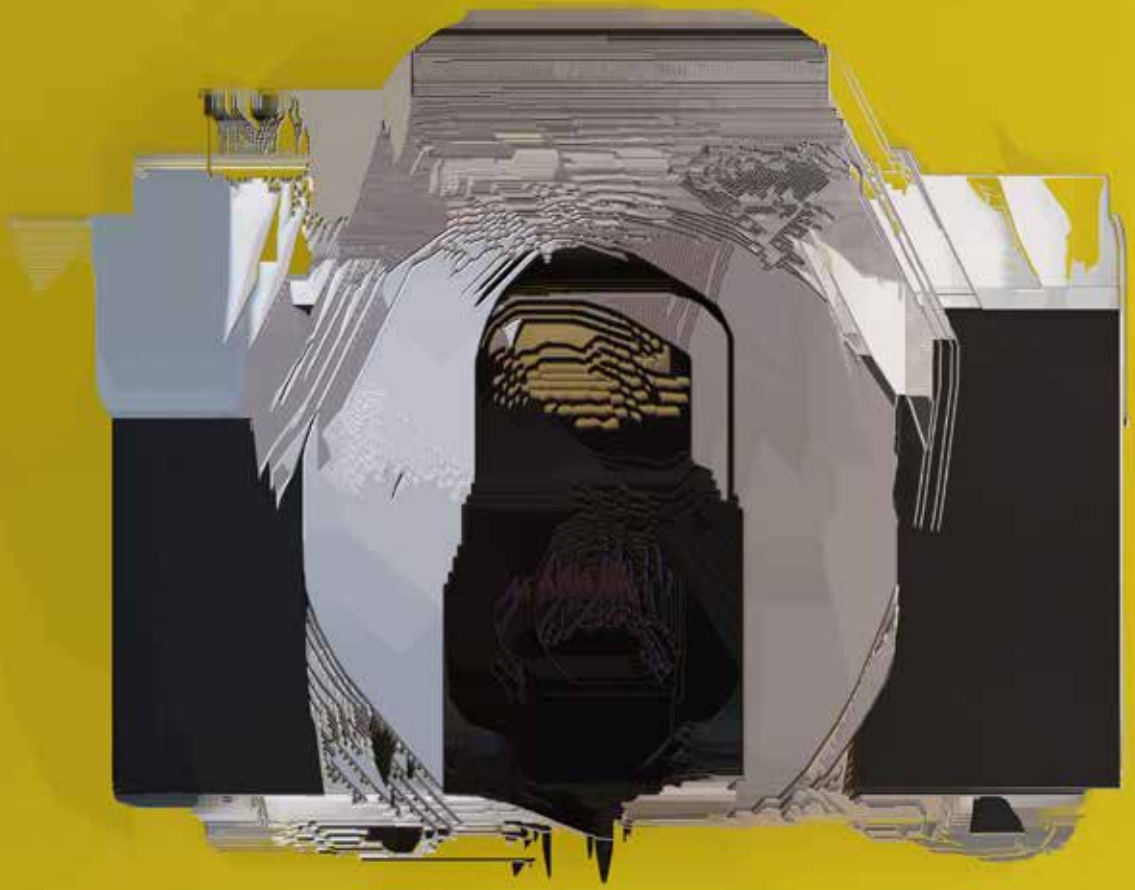
2.3 Wat er al kan (en wat nog niet)

We leven nu al in een wereld vol met kunstmatige intelligentie: Siri, Alexa, Google search, gezichtsherkenning, gepersonaliseerde advertenties, navigatiesoftware, zelfstandig inparkerende auto's en ga zo maar door. AI-toepassingen kunnen veel taken betrouwbaar uitvoeren, net zo goed of vaak zelfs beter dan mensen. Deze taken zijn specialistisch en afgebakend en daarmee vallen alle huidige AI-toepassingen onder het eerdergenoemde narrow AI. De tegenhanger general AI, waarbij kunstmatige intelligentie alles kan wat mensen kunnen of meer, bestaat (nog) niet.

Er wordt veel gespeculeerd over de vraag of general AI er ooit gaat komen. Er zijn onderzoekers die waarschuwen voor de *singularity*, het punt waarop AI-systemen slimmer worden dan mensen (Bostrom, 2014; Kurzweil, 2010). AI-systemen zouden dan hun intelligentie gebruiken om zichzelf te verbeteren, waardoor ze slimmer worden en zichzelf nog verder zouden verbeteren, enzovoort. Hiermee zouden AI-systemen op een gegeven moment *superintelligence* bereiken, intelligentie die ver boven die van mensen uitstijgt. Er zijn ook veel gerenommeerde AI-onderzoekers die verwachten dat de singulariteit nooit gaat gebeuren (Broussard, 2018). Zij geloven dat dergelijke scenario's onmogelijk of hoogst onwaarschijnlijk zijn, al was het maar omdat mensen bij elektrische apparaten de stekker uit het stopcontact kunnen trekken. Vrijwel alle experts zijn het er in elk geval over eens dat er in de komende decennia in elk geval geen sprake zal zijn van singulariteit (Tegmark, 2017).

Een grote uitdaging die de ontwikkeling van general AI in de weg staat, is *common sense*. Een kind van tien jaar oud snapt dat het een glazen vaas niet van de tafel moet duwen, dat het ongepast is om te hard te lachen als iemand huilt en dat het een baby niet te hard moet knijpen. Een AI-toepassing begrijpt dat niet zomaar en dat maakt het lastig voor AI-toepassingen om om te gaan met nieuwe en onvoorspelbare omgevingen. Wanneer je een zorgrobot uit een ziekenhuis weghaalt en meeneemt naar een restaurant, heeft hij geen idee meer wat hij moet doen en welk gedrag daar passend is. Siri verstaat misschien wel welke woorden je uitspreekt, maar begrijpt niet wat je zegt. Een goed gesprek voeren met Siri lukt niet. Zalando toont met behulp van kunstmatige intelligentie advertenties op basis van eerdere aankopen, maar goed kledingadvies geven met kunstmatige intelligentie gaat niet. Common sense zal ook in de toekomst een grote uitdaging blijven in het ontwikkelen van AI-toepassingen (Broussard, 2018).

Ook al blijft kunstmatige intelligentie beperkt tot narrow AI, dat neemt niet weg dat het de komende jaren een steeds grotere stempel gaat drukken op de wereld waarin we leven. Een groeiend aantal taken die nu nog door mensen worden uitgevoerd, zal overgenomen worden door AI-toepassingen. Hóe dat onze maatschappij verder gaat beïnvloeden staat nog niet vast. Zoals geschetst in de inleiding zou dat kunnen leiden tot een veiligere, aangenamere wereld, maar ook tot een ondoorzichtige samenleving waarin veel wordt bepaald door AI-toepassingen, maar zonder dat we precies weten hoe en waarom. Kunstmatige intelligentie heeft nu al veel impact op de maatschappij en van beide mogelijke toekomstrichtingen zijn tekenen zichtbaar. In het volgende hoofdstuk ga ik daarom dieper in op de maatschappelijke implicaties van kunstmatige intelligentie die nu al zichtbaar zijn.



HOOFDSTUK 3

Maatschappelijke implicaties

Er gaat geen dag meer voorbij waarin we niet een zoekopdracht uitvoeren met Google Search, onze tijdslijn checken op Instagram, Facebook of LinkedIn, of een aanbeveling krijgen van Spotify, YouTube of Netflix. Ook groeit het aantal auto's dat zelf kan inparkeren en het gebruik van software die beelden van surveillancecamera's automatisch scant op veiligheidsdreigingen. We komen vrijwel dagelijks in aanraking met kunstmatige intelligentie. Met de groeiende verspreiding van kunstmatige intelligentie wordt de impact ervan op onze samenleving steeds groter. Deze impact kan zowel positief als negatief zijn. Aan de ene kant wordt kunstmatige intelligentie bijvoorbeeld ingezet om gezondheidszorg te verbeteren of energie te besparen, maar aan de andere kant zijn er toepassingen die de autonomie van mensen beperken of groepen mensen discrimineren.

Een manier om de maatschappelijke implicaties van kunstmatige intelligentie voor mensen en de samenleving in kaart te brengen is aan de hand van waarden. Een waarde is datgene wat een persoon of groep personen belangrijk vindt, bijvoorbeeld veiligheid, vriendschap, nieuwsgierigheid en creativiteit (Friedman, Kahn, Borning & Huldtgren, 2013). Waarden vormen daarmee een goede kapstok om de positieve en negatieve gevolgen van kunstmatige intelligentie in kaart te brengen: AI-toepassingen hebben positieve gevolgen voor mens en samenleving wanneer ze waarden bevorderen en een negatieve impact wanneer ze waarden aantasten.

Niet iedereen vindt alle waarden even belangrijk of geeft op dezelfde manier invulling aan een waarde. Daarmee zijn soms verschillende standpunten mogelijk over of het effect van een AI-toepassing vooral positief of negatief is. Ethiek kan helpen om na te denken over of de maatschappelijke implicaties van AI-toepassingen wel of niet gewenst zijn. Ethiek is een tak van de filosofie die gaat over 'juist' handelen (Deigh, 2010). Daarbij horen vragen als: Wat is goed en fout? Wat is het juiste om te doen? Welk gedrag is lovenswaardig en welk is verachtelijk?

Binnen de ethiek bestaan verschillende theorieën om na te denken over wat goed en fout handelen is. De drie belangrijkste ethische theorieën zijn: utilitarisme, deontologie en deugdenethiek. Deze theorieën leiden echter niet altijd tot dezelfde conclusie over wat in een bepaalde situatie het juiste is om te doen.

Het utilitarisme streeft er bijvoorbeeld naar om zoveel mogelijk geluk voor zoveel mogelijk mensen te creëren. Dit lijkt logisch, maar zou kunnen betekenen dat een gezond persoon opgeofferd moet worden als er met haar hart, lever, longen en nieren vier mensenlevens gered kunnen worden. Vier levens kan namelijk meer geluk opleveren dan één leven. Deontologie zegt dat dit niet zomaar mag. Het leidende principe van deze theorie is dat je anderen zou moeten behandelen zoals je iedereen zou willen behandelen, inclusief jezelf. Dus als je vindt dat er mensen zijn, bijvoorbeeld jijzelf of je familie, die hun organen niet hoeven te doneren tijdens hun leven, dan mag je dat van niemand vragen. Deugdenethiek kijkt vooral naar de deugden, goede karaktereigenschappen, die ten grondslag liggen aan een handeling. Een persoon die z'n organen doneert is bijvoorbeeld moedig, hulpvaardig of altruïstisch, en daarmee is het een juiste handeling. De uitkomst van de handeling (zijn er mensenlevens gered?) en het achterliggende principe (zou iedereen organen moeten donderen?) zijn daarbij minder belangrijk.

Bovenstaand voorbeeld laat zien dat ethiek niet altijd tot eenduidige antwoorden leidt op de vraag wat goed en fout is. Dit geldt ook voor de maatschappelijke implicaties van AI-toepassingen; soms is het niet duidelijk of de implicaties van een toepassing vooral positief of negatief zijn voor de samenleving en soms zijn er verschillende standpunten mogelijk. Toch krijgen makers van AI-toepassingen met lastige kwesties te maken en moeten ze er keuzes over maken. Ook al biedt ethiek niet altijd eenduidige antwoorden, het kan in zulke gevallen kaders schaffen om na te denken over wat goed handelen is en geeft inzicht in welke overwegingen daarbij belangrijk zijn. Het doel van dit hoofdstuk is dan ook niet om antwoord te geven op alle ethische vragen rondom AI-toepassingen, maar wel om lastige kwesties te beschrijven en daar vragen over te stellen.

In dit hoofdstuk schets ik enkele van de meest belangrijke maatschappelijke gevolgen van kunstmatige intelligentie aan de hand van waarden. Eerst bespreek ik de gevolgen van fysieke AI-toepassingen, vervolgens die van niet-fysieke AI-toepassingen en ten slotte de gevolgen van met elkaar verbonden AI-toepassingen. Dit is nadrukkelijk niet een uitputtend overzicht van alle mogelijke maatschappelijke implicaties van kunstmatige intelligentie.

3.1 Robots: veiligheid en verantwoordelijkheid

Fysieke AI-toepassingen, ofwel robots, vervullen vele verschillende functies (Gates, 2007; Vergunst & Mols, 2017). In de zorg, bijvoorbeeld, houden sociale robots ouderen gezelschap en ondersteunen ze verplegend personeel, in het onderwijs fungeren ze als tutor en in winkels en hotels verwelkomen ze nieuwe klanten. Maar ook apparaten als zelfrijdende auto's, slimme koelkasten en intelligente lantaarnpalen vallen onder robots.

Dull, dirty en dangerous

Robots worden nooit moe, raken niet verveeld of afgeleid, ervaren geen frustratie of pijn en kunnen overleven op plekken waar wij dat niet kunnen. Dat maakt robots erg geschikt om saaie, zware en gevaarlijke klussen op te knappen. Industriële robots fabriceren en assembleren bijvoorbeeld al een hele tijd producten in fabrieken. In oorlogsgebieden worden militaire robots ingezet om verkenningen uit te voeren, spullen te dragen of te vechten. In de ruimtevaart worden robots gebruikt om het heelal te verkennen, zoals bijvoorbeeld *Mars rovers*.

Robots kunnen mensen dus pijn en moeite besparen en veiligheidsrisico's bij hen wegnemen. Dit kan een groot verschil maken in het verkeer. In Nederland komen jaarlijks honderden mensen om in het verkeer. Het grootste deel van die verkeersongelukken wordt veroorzaakt door menselijke fouten, bijvoorbeeld doordat ze niet goed opletten en vervolgens te weinig tijd hebben om adequaat op een gevaarlijke situatie te reageren. In de toekomst zouden veel van dit soort ongelukken voorkomen kunnen worden door kunstmatige intelligentie in te zetten, bijvoorbeeld door auto's automatisch te laten remmen als ze een bijna-botsing detecteren. Kunstmatige intelligentie zou zo bijdragen aan veiligheid. Wanneer auto's volledig zelfstandig konden rijden, zou dat nog een ander voordeel met zich meebrengen ten opzichte van gewone auto's. Ook mensen die geen auto kunnen rijden, zouden zich zelfstandig met de auto kunnen verplaatsen, zoals minderjarigen, ouderen of mensen met een beperking. Deze mensen zouden hiermee meer vrijheid en autonomie krijgen, doordat ze niet afhankelijk zijn van andere mensen die hen moeten halen of brengen.

Ook in de operatiekamer van het ziekenhuis komen de eigenschappen van robots soms goed van pas. Robots kunnen preciezer en gedetailleerder werken dan menselijke chirurgen. Voor patiënten kan dit bijvoorbeeld leiden tot minder bloedverlies of een kortere hersteltijd. Ook kunnen robots hun concentratie continue volledig vasthouden tijdens een urenlange ingreep, wat de kans op fouten verkleint. De afgelopen jaren hebben robots in ziekenhuizen verschillende operaties succesvol uitgevoerd.

Ongelukken door robots

Robots kunnen veel, maar helaas zijn ze niet feilloos. Net als mensen maken ook robots soms fouten. AI in 1979 is het eerste dodelijke slachtoffer gevallen door een ongeluk met een industriële robot in een Fordfabriek (Kravets, 2010). Het eerste ongeluk door een zelfrijdende auto met dodelijke afloop vond plaats in 2016, toen een Tesla op een vrachtauto inreed omdat de auto ten onrechte had aangenomen dat het witte oppervlak aan de zijkant van de vrachtauto onderdeel was van de lucht (Yadron & Tynan, 2016). In de gezondheidszorg kunnen robots een foute diagnose stellen of de verkeerde medicijnen toedienen (Sharkey & Sharkey, 2012; Van Wynsberghe, 2013).

De toenemende inzet van robots zorgt voor steeds meer veiligheidsrisico's. Waar robots eerst vooral werden ingezet in de beschermde omgeving van fabrieken, treden ze nu steeds meer onze leefwereld binnen. Ook worden robots in toenemende mate voor belangrijke en risicovolle taken ingezet, waardoor de consequenties steeds groter worden als er iets fout gaat. Dit stelt ons voor een aantal lastige ethische dilemma's.

AI-toepassingen kunnen nog zo goed in elkaar zitten en nog zo uitgebreid getest zijn, het is nooit mogelijk om voor de volledige 100% te garanderen dat er geen ongelukken zullen gebeuren. Betekent dat dat we ze nooit in mogen zetten als er kans is op (dodelijke) slachtoffers? Waarschijnlijk niet. Er bestaat bijvoorbeeld ook een kans, hoe klein ook, dat de dijken in Nederland doorbreken, maar ondanks die wetenschap blijven veel mensen toch graag onder zeeniveau wonen. Dat heeft ermee te maken dat de kans op een dijkdoorbraak zo ontzettend klein is. Het is dus zinvoller om ons af te vragen welke risico's van AI-toepassingen we aanvaardbaar vinden dan of een AI-toepassing volledig risicovrij is.

Als het bijvoorbeeld gaat om zelfrijdende auto's zijn er verschillende standpunten mogelijk. Zouden we zelfrijdende auto's moeten toestaan op het moment dat ze *minder* ongelukken maken dan mensen? Sommige voorstanders van zelfrijdende auto's betogen dat we op dat moment zelfs moreel verplicht zijn om over te stappen op zelfrijdende auto's. Er zijn echter ook onderzoekers die beargumenteren dat zelfrijdende auto's onze menswaardigheid kunnen aantasten, omdat veel mensen het erger vinden om iemand te verliezen door een dodelijk ongeluk veroorzaakt door een machine dan door een mens.

Verantwoordelijkheid voor AI-toepassingen

Het inzetten van kunstmatige intelligentie voor taken met risico's op ongelukken roept nog een belangrijke vraag op: wie is er verantwoordelijk voor de gevolgen als er iets mis gaat? Net als bij andere producten is de ontwerper van een AI-toepassing verantwoordelijk voor ontwerpfouten en de fabrikant voor productiefouten. Toch is

het niet altijd gemakkelijk om een verantwoordelijke aan te wijzen, doordat er zoveel verschillende partijen betrokken zijn bij het ontwerpen, ontwikkelen, fabriceren en gebruiken van een product: het *many hands problem* (Thompson, 1980). Bij AI-toepassingen, die zelfstandig beslissingen nemen en taken uitvoeren, kan het nog lastiger zijn om een verantwoordelijke aan te wijzen in het geval er iets fout gaat (Murphy & Shields, 2012; Noorman & Johnson, 2014). Dit specifieke probleem voor AI-toepassingen wordt de *responsiblity gap* genoemd (Matthias, 2004).

Een mogelijk uitgangspunt is dat de gebruiker van een AI-toepassing altijd verantwoordelijk moet zijn voor de acties van die toepassing. Dit zou betekenen dat de gebruiker dan ook in de gelegenheid moet worden gesteld om verantwoordelijkheid te nemen, bijvoorbeeld door te zorgen dat bij belangrijke beslissingen de mens altijd kan blijven ingrijpen en het AI-toepassing kan overstemmen als dat nodig is. Maar wat als dat juist leidt tot meer ongelukken? Of wanneer het gaat om situaties waarin een bovenmenselijk snelle reactie vereist is? Soms kunnen mensen een AI-toepassing niet goed bijsturen. Bovendien is het de vraag of mensen gedurende lange tijd alert kunnen blijven, bijvoorbeeld tijdens een lange autorit waarin ze niet zelf rijden.

Een andere optie is om robots zelf verantwoordelijk te maken voor hun gedrag. Sommige onderzoekers pleiten ervoor om AI-systemen een juridische status te geven, net zoals bedrijven en organisaties, die als rechtspersoon fungeren en als zodanig aansprakelijk worden gesteld voor fouten. Zelfrijdende auto's zouden dan bijvoorbeeld verplicht een verzekering moeten hebben om de weg op te gaan. Hoe beter een model is getest en hoe langer het goede prestaties heeft geleverd, des te lager de premie wordt.

Het debat over kunstmatige intelligentie en verantwoordelijkheid speelt zich voor een belangrijk deel af in het militaire domein. Voorstanders van militaire robots beargumenteren dat robots betrouwbaarder en nauwkeuriger kunnen vechten dan mensen en zorgen voor minder burgerslachtoffers en andere nevenschade (Arkin, 2009). Met de huidige technologie zijn militaire robots, ook wel autonome wapensystemen genoemd, in staat om zelfstandig een gebied te monitoren, vijanden erin te detecteren, op die vijanden te richten en vervolgens te vuren. Er is echter ontzettend veel weerstand tegen het automatiseren van deze *kill chain*, waarvan de campagne *Stop Killer Robots* de meeste media-aandacht heeft gegenereerd. Volgens veel mensen overschrijdt het vermoorden van mensen door robots een grens die de menselijke waardigheid aantast en zouden mensen beslissingen over levens van andere mensen nooit uit handen mogen geven (Docherty, 2012).

3.2 Algoritmes: rechtvaardigheid en inzichtelijkheid

Niet-fysieke AI-toepassingen worden vaak virtuele agents of algoritmes genoemd. De toepassing van virtuele agents en algoritmes heeft verschillende ethische gevolgen.

Data, data en nog meer data

Zoals besproken in hoofdstuk 2 hebben lerende algoritmes data, liefst veel data, nodig om goed te worden in een taak. Tegelijkertijd hebben wij algoritmes nodig om een weg te vinden in de gigantische hoeveelheden data die beschikbaar zijn. De afgelopen twee jaar zijn er meer data bijgekomen dan in de hele geschiedenis daarvoor. Elke minuut komt er 400 uur aan videomateriaal bij op YouTube (DigiVid360, 2018). Op Facebook worden elke dag meer dan 300 miljoen foto's geüpload (Zephoria, 2018). De omvang van het internet is immens. Als we handmatig websites zouden moeten doorzoeken als we informatie nodig hebben, dan zou één zoekopdracht zo een paar dagen kunnen kosten. Gelukkig kunnen computers heel snel heel veel data verwerken en zorgen slimme algoritmes dat die data op een efficiënte manier worden doorzocht. Google Search, de meest gebruikte zoekmachine ter wereld, voert per seconde gemiddeld zo'n 40.000 zoekopdrachten uit (Internet live stats, 2018).

Hoe meer data er beschikbaar zijn, des te beter kunnen online algoritmes hun resultaten op ons afstemmen. In de eerste plaats doordat algoritmes ons steeds beter leren kennen, maar ook doordat algoritmes weten wat personen met vergelijkbare voorkeuren leuk of interessant vinden. Deze personalisatie van zoekresultaten kan ervoor zorgen dat we nieuwe series, films, boeken of muziek ontdekken waar we anders nooit tegenaan zouden lopen. Veel aanbevelingsalgoritmes zijn zo ontworpen dat ze serendipiteit bevorderen; je wordt af en toe blootgesteld aan iets compleet nieuws waardoor de kans ontstaat dat je onverwachts iets moois, leuks of interessants ontdekt.

Naast het doorzoeken van informatie zijn algoritmes goed in het ordenen en classificeren van informatie. Dat kan ons een hoop tijd en frustratie schelen. Zo hoeven we spammail niet allemaal zelf te lezen en worden onze foto's automatisch netjes geordend in logische categorieën. Voor ons is dit vooral handig, maar in ziekenhuizen kunnen algoritmes die röntgenfoto's en MRI-scans analyseren en betere diagnoses stellen dan artsen van levensbelang zijn.

Algoritmische bias

Algoritmes kunnen ons een hoop werk besparen en vaak kunnen ze zelfs betere beslissingen nemen dan mensen. Algoritmes presteren echter niet feilloos. Veel algoritmes hebben een *bias*, dat wil zeggen dat de uitkomsten van het algoritme mede worden bepaald door vooroordelen (Bozdag, 2013). Dat is niet zo erg

wanneer Netflix een serie aanraadt die toch niet zo leuk blijkt, je hebt in dat geval hoogstens een uurtje tijd verloren. Bias in algoritmes vormt echter wel een serieus probleem als algoritmes worden gebruikt om beslissingen te nemen met een grote impact, zoals of iemand in aanmerking komt voor een baan, lening, vervroegde vrijlating of een hypotheek, en dat gebeurt nu al op grote schaal (O'Neil, 2016).

Een zaak die in de VS veel aandacht heeft gekregen gaat over software die voorspelt hoe groot de kans is dat een verdachte of gevangene opnieuw een misdaad begaat. Rechters gebruiken deze voorspellingen bij het bepalen van de zwaarte van een straf en of iemand in aanmerking komt voor vervroegde vrijlating. Uit onderzoek door ProPublica (Angwin & Larson, 2016) bleek dat de misdaadrisico's systematisch hoger werden ingeschat voor zwarte dan voor witte mensen. De reden hiervoor is dat de algoritmes getraind zijn met data uit het verleden, waarin zwarte mensen vaker opnieuw een misdaad pleegden dan witte mensen. Door data van andere mensen uit het verleden te gebruiken, kregen zwarte gevangen nu op voorhand minder kans om in aanmerking te komen voor vervroegde vrijlating. Hier is sprake van etnisch profileren en deze software toont een racistische bias.

Er zijn vele voorbeelden van bias in algoritmes. Zo bleek een beeldherkenningsapplicatie van Google zwarte mensen op foto's te identificeren als gorilla's (Liedtke, 2015). Een algoritme op een vacaturewebsite toonde vaker vacatures voor goedbetaalde banen aan mannen dan aan vrouwen (Gibbs, 2015). De politie gebruikt algoritmes om te voorspellen waar misdaden zullen plaatsvinden en waar ze dus het beste kunnen patrouilleren (Perry, 2013). Het probleem met dat laatste is dat wanneer de politie eenmaal meer aanwezig is in een bepaalde buurt, er ook meer misdaden gevonden zullen worden in die buurt. De algoritmes zullen vervolgens een nog hogere waarschijnlijkheid op misdaden in die buurt voorspellen. Sommige mensen, vaak in arme buurten, zullen daardoor onevenredig vaak met politiecontroles te maken krijgen. Algoritmes met bias ondermijnen daarmee een gelijkwaardige behandeling en rechtvaardigheid.

Ook op sociale media, waar algoritmes bepalen welke content gebruikers te zien krijgen, zijn algoritmes soms bevooroordeeld. Onderzoek wijst bijvoorbeeld uit dat haatzaaiende berichten en 'fake news' vaak hoog in tijdslijnen komen omdat die meer aandacht genereren (Shao et al., 2017). Verder zorgen algoritmes op sociale media ervoor dat verschillende gebruikers verschillende informatie te zien krijgen en vaak krijgen gebruikers alleen maar informatie te zien die aansluit bij hun interesses en wereldbeeld. Mensen hebben hierdoor een gebrek aan andere perspectieven en komen in een zogenaamde *filterbubbel* (Pariser, 2011). Sociale media vormen voor veel mensen hun belangrijkste bron van nieuws. Algoritmes op sociale media kunnen de publieke opinie dus sterk beïnvloeden en zelfs effect

hebben op de uitslag van politieke verkiezingen. Hiermee vormen algoritmes op sociale media een bedreiging voor de democratie. Democratie vereist immers dat burgers goed geïnformeerd zijn.

De reden dat algoritmes fouten maken en bevoordeeld zijn, heeft vaak te maken met onvolkomenheden in de data waarmee een algoritme wordt getraind. Stel dat een bedrijf een algoritme wil maken dat voorspelt hoe succesvol een toekomstige werknemer zal zijn. Dit algoritme zou getraind kunnen worden met data van eerdere werknemers. Maar als in het verleden alleen mannen topposities bereikten, bijvoorbeeld doordat er nauwelijks vrouwen in het bedrijf werkten, dan zal het algoritme leren dat mannen een grotere kans op succes hebben dan vrouwen. De data waarmee het algoritme leert, zijn in dit geval bevoordeeld. Het komt ook voor dat algoritmes getraind worden op datasets met incomplete of onjuiste data. Een algoritme is geen formule die data kan omtoveren tot nuttige kennis en informatie: *garbage in* betekent *garbage out*. In de praktijk wordt echter vaak – ten onrechte – gedacht dat algoritmes en data neutraal en objectief zijn en nemen veel mensen de uitkomsten van algoritmes klakkeloos over (Broussard, 2018; O'Neil, 2016).

Inzichtelijkheid

Vaak is niet duidelijk hoe een algoritme werkt en waarom de toepassing ervan tot een bepaalde uitkomst leidt (Pasquale, 2015). Algoritmes worden daarom ook wel *black boxes* genoemd. Zoals hierboven beschreven komt het regelmatig voor dat algoritmes bevooroordeeld zijn en bepaalde groepen mensen uitsluiten. Hierbij gaat het soms om beslissingen die grote impact hebben, zoals of iemand in aanmerking komt voor een hypotheek, lening of een baan. Door het gebrek aan inzichtelijkheid en transparantie van algoritmes is het echter lastig om erachter te komen hoe beslissingen worden genomen. Daarmee is het moeilijk om aan te tonen dat een beslissing niet is genomen op de juiste gronden en wordt het ook lastiger om onjuiste beslissingen aan te vechten.

Het gebrek aan inzichtelijkheid van algoritmes kent verschillende oorzaken. Veel bedrijven willen hun algoritmes niet vrijgeven omdat ze bang zijn dat ze daarmee hun concurrentiepositie aantasten. Het algoritme van Facebook is bijvoorbeeld topgeheim. Een andere belangrijke reden voor het gebrek aan inzichtelijkheid van algoritmes is dat veel algoritmes erg complex zijn. Soms is er niet sprake van één algoritme, maar van een netwerk van meerdere algoritmes die elkaar beïnvloeden. Dus ook al zou je toegang hebben tot de code van deze algoritmes, dan geeft dat nog steeds geen inzicht in hoe het algoritme tot een bepaalde uitkomst leidt. Dit maakt het lastig om discriminatie door de inzet van algoritmes te bestrijden.

3.3 Veelomvattendheid: privacy, beveiliging en banen

In de bovenstaande secties zijn ethische implicaties besproken van fysieke en virtuele AI-toepassingen. Hierbij ging het steeds om afzonderlijke producten of diensten. AI-toepassingen worden op steeds grotere schaal ingezet en het komt regelmatig voor dat verschillende AI-toepassingen, al dan niet met een fysiek lichaam, met elkaar in verbinding staan. Dit levert nieuwe kansen en bedreigingen op. Deze sectie gaat in op de maatschappelijke implicaties die ontstaan door het verbinden en op grote schaal gebruiken van AI-toepassingen.

Verbondenheid van AI-toepassingen

AI-toepassingen die met elkaar verbonden zijn, kunnen informatie met elkaar uitwisselen. Zowel fysieke als virtuele AI-toepassingen kunnen met elkaar worden verbonden. Een voorbeeld van verbonden AI-toepassingen zijn de verschillende diensten met kunstmatige intelligentie van grote technologiebedrijven als Facebook en Google. Deze diensten wisselen informatie met elkaar uit om betere, meer gepersonaliseerde service te bieden. Bijvoorbeeld, een emailapplicatie die weet dat een gebruiker op reis gaat naar Berlijn kan dit doorgeven aan een kaartapplicatie, die vervolgens aan de gebruiker voorstelt om de kaart van Berlijn te downloaden. Een ander voorbeeld is een zoekmachine die, op basis van een gebruikersprofiel met iemands voorkeuren en interesses, resultaten van een zoekopdracht specifiek afstemt op de persoon die de zoekopdracht uitvoert. Om zo'n gebruikersprofiel op te bouwen wordt vaak informatie gecombineerd die is verzameld binnen verschillende AI-toepassingen.

Een ander bekend voorbeeld van verbonden AI-toepassingen is het *Internet of Things (IoT)* (Van Berkel et al., 2018). In dit netwerk zijn verschillende 'dingen', zoals slimme koelkasten en lantaarnpalen, met elkaar en vaak ook nog met virtuele agents verbonden. Deze toepassingen kunnen berichten naar elkaar kunnen versturen en zo hun acties beter op elkaar afstemmen. Als een slimme fruitschaal bijvoorbeeld merkt dat de appels bijna op zijn, dan kan hij een virtuele inkoopagent alvast een seintje sturen dat hij nieuwe appel moet bestellen. Het IoT wordt op diverse plekken toegepast. Huizen met veel IoT -toepassingen heten vaak *smart homes* en steden met veel IoT-toepassingen worden vaak *smart cities* genoemd.

In smart homes kunnen verbonden AI-toepassingen onze levens op allerlei manieren gemakkelijker maken. 's Ochtends kunnen ze bijvoorbeeld de verwarming op tijd aanzetten, de wekker af laten gaan op het moment dat je net een slaapcyclus doorlopen hebt, het licht aandoen, een muziekje uitkiezen dat past bij je stemming van die dag, koffiezetten, de broodrooster aandoen en die berichten uit de krant selecteren die jij interessant vindt. Naast het verhogen van gemak, kan een smart home ook een hoop energie besparen. Google Nest claimt op basis van klantgegevens in Nederland dat een besparing van zo'n 16% mogelijk is door

gebruik van de slimme Nest-thermostaat (Nest, 2018). Tot slot kunnen mensen met een visuele, auditieve of fysieke beperking met intelligente technologie worden geholpen bij hun dagelijkse activiteiten (Domingo, 2012).

Op een grotere schaal, in een smart city, kan verbondenheid van AI-toepassingen ook grote duurzaamheidsvoordelen opleveren. Meer inzicht in energieverbruik van publieke voorzieningen als straatverlichting, transportdiensten en camera's kan helpen om deze efficiënter in te richten en zo om energie te besparen (Zanella, Bui, Castellani, Vangelista & Zorzi, 2014). Sensoren in de stad kunnen luchtkwaliteit monitoren en passende maatregelen nemen als dat nodig is, bijvoorbeeld door het verkeer om te leiden als bepaalde grenswaarden overschreden worden (Farooq, Waseem, Mazgar, Khairi & Kamal, 2015). Intelligente technologie in de stad kan ook zorgen voor meer veiligheid. Politie kan bewerkte data gebruiken voor effectievere handhaving, opsporing en hulpverlening.

Privacy

Onderlinge verbondenheid van AI-toepassingen kan ervoor zorgen dat verschillende toepassingen beter met elkaar samenwerken om individuen en de samenleving als geheel beter van dienst te zijn. Maar het koppelen van verschillende AI-toepassingen verhoogt ook de kans dat onze privacy wordt aangetast.

AI-toepassingen verzamelen heel erg veel data over ons. Online AI-toepassingen nemen waar naar welke muziek we vaak luisteren, welke websites we bezoeken en wie onze vrienden en kennissen zijn; smartphones weten hoe vaak we op onze telefoon kijken, met wie we appen en ze registreren continu waar we zijn; smart watches houden bij hoeveel we bewegen en hoe hoog onze hartslag is; banken hebben een overzicht van al onze digitale geldtransacties en winkels weten aan welke spullen we ons geld uitgeven; door persoonsgebonden ov-kaarten en automatische nummerbordherkenning valt te traceren waar we reizen; openbare ruimtes hangen vol met surveillancecamera's en gezichtsherkenningssoftware kan ons daarop automatisch herkennen. Om maar eens wat te noemen. Data uit al die afzonderlijke bronnen op zich kunnen onze privacy al aantasten, maar het combineren van al die data door verschillende AI-toepassingen met elkaar te verbinden, levert een nog veel grotere bedreiging voor onze privacy op.

Het credo 'ik heb niets te verbergen' wordt regelmatig genoemd door mensen die weinig problemen zien in het opgeven van privacy (Martijn & Tokmetzis, 2016). Maar vrijwel niemand wil intieme gesprekken met een geliefde met de hele wereld delen. Of informatie privacygevoelig is, hangt sterk af van de context waarin het getoond wordt (Nissenbaum, 2009). Informatie die in een bepaalde situatie niet belastend lijkt, kan ergens anders of op een ander tijdstip wel belastend zijn. In Nederland is het misschien geen probleem om homoseksualiteit openbaar te

maken, maar in sommige landen staat daar de doodstraf op. Video's van een uitbundig feestje kunnen grappig zijn om te delen met vrienden, maar toekomstige werknemers knappen er misschien op af. Data die eenmaal opgeslagen zijn, worden niet zomaar gewist. Dit besef kan leiden tot een *chilling effect* waarin mensen banger zijn om fouten te maken en bepaald gedrag zullen vermijden (Marder, Joinson, Shankar & Houghton, 2016). Dit gaat ten koste van vrijheid. Angst voor fouten kan ook creativiteit en innovatie belemmeren. Ook kan continue monitoring als betuttelend ervaren worden.

Een belangrijke vraag in de discussie omtrent privacy is wíe er toegang heeft tot persoonsgegevens. Dat bepaalt voor een belangrijk deel hoe er met die data wordt omgegaan. Eerder is al genoemd dat grote technologiebedrijven veel verschillende diensten aanbieden waarbij gebruikersdata worden verzameld. Met name in de Westerse wereld loopt de '*Big 5*' van Amerikaanse technologiebedrijven daarin voorop: Facebook, Google, Apple, Amazon en Microsoft. Deze bedrijven beheren veel van de infrastructuur die wij dagelijks massaal gebruiken voor ons werk en in onze vrije tijd, en beschikken daardoor over een schat aan data van heel veel mensen. Helaas blijken bedrijven niet altijd even zorgvuldig met persoonsgegevens om te gaan. De laatste tijd zijn er verschillende schandalen geweest rondom het oneigenlijk gebruiken van gebruikersdata, bijvoorbeeld het schandaal waarbij Facebook data aan Cambridge Analytica heeft verkocht (Granville, 2018).

Andere spelers met toegang tot veel persoonsgegevens zijn overheden. Overheden hebben geen winstoogmerk, maar ook overheden kunnen data op een manier gebruiken die niet wenselijk is. De Nederlandse politie gebruikt bijvoorbeeld data om criminaliteit te voorkomen (Andringa, 2018). Een risico hierbij is dat er sprake is van etnisch profileren, wat ten koste gaat van een gelijke behandeling (zie vorige sectie). Met name de Chinese overheid gaat ver in het verzamelen en gebruiken van data over haar inwoners. Zij heeft een sociaal kredietsysteem geïntroduceerd dat voor elke bewoner op basis van over hem of haar verzamelde data een score (sociaal krediet) berekent (Rollet, 2018). Deze score is onder andere al gebruikt om te bepalen of bewoners met de high-speedtrein mogen reizen en of ze worden toegelaten tot privéschool.

In de VS hebben grote technologiebedrijven veel macht doordat ze beschikken over veel persoonsgegevens en in China is dat de overheid. De Europese Unie probeert daar een derde model tegenover te zetten door burgers inzicht en invloed te geven op wat er met hun data gebeurt. In mei 2018 is de Algemene Verordening Gegevensbescherming (AVG) van kracht gegaan, dit betreft nieuwe Europese wetgeving rondom privacy (European Commission, 2018b). Deze wet stelt dat gebruikers eigenaar blijven van hun eigen persoonsgegevens en dat ze het recht hebben op inzicht in welke data een partij of bedrijf over hen heeft verzameld.

Beveiliging

Eerder in dit hoofdstuk is genoemd dat robots onze veiligheid kunnen bedreigen doordat ze onbedoeld ongelukken kunnen veroorzaken. Naast deze niet-intentioneel veroorzaakte ongelukken zijn er ook veiligheidsrisico's die ontstaan door kwaadwillend gedrag van bijvoorbeeld hackers en cybercriminelen (NCSC, 2018). Zo kan de motor van een gehackte zelfrijdende auto op afstand worden uitgezet, een gehackte pacemaker worden ontregeld en de elektriciteitsvoorziening vanuit een gehackte energiecentrale worden stopgezet. Het gebeurt regelmatig dat computers van privépersonen gehackt worden en er losgeld betaald moet worden om weer toegang te krijgen tot bestanden. Ook wordt speelgoed met camera's gehackt zodat kwaadwillenden kinderen kunnen bespieden.

Door het verbinden van verschillende AI-toepassingen ontstaan nieuwe veiligheidsrisico's. Zo worden bij een *distributed denial-of-service*-aanval heel veel apparaten gehackt en vanuit dit 'botnet' wordt een grote hoeveelheid netwerkwerkverkeer naar bijvoorbeeld een webserver verstuurd waardoor de dienst niet meer beschikbaar is (Silva, Silva, Pinto & Salles, 2013).

Dé remedie tegen ongewenste toegang tot AI-toepassingen is goede cybersecurity. In de praktijk blijkt de beveiliging van veel IT-systemen echter zwak. Zo zijn er veel gebruikers die standaard ingestelde gebruikersnamen en wachtwoorden niet wijzigen, wat hacken aanzienlijk makkelijker maakt. Een van de problemen is dat er weinig prikkels zijn om goed te beveiligen (Harbers et al., 2018). Zo hebben eigenaren van gehackte apparaten er meestal geen last van dat een apparaat onderdeel is van een botnet, vaak merken ze het niet eens. De Cyber Security Raad in Nederland maakt zich hard voor betere beveiliging (www.cybersecurityraad.nl).

Banen

Wanneer kunstmatige intelligentie op steeds grotere schaal wordt toegepast, zullen er banen van mensen verloren gaan (Makridakis, 2017). Experts zijn het niet eens over de mate waarin dat zal gebeuren (Freese, Dekker, Kool & Van Est, 2018). Volgens sommigen gaat kunstmatige intelligentie drastische gevolgen hebben op de banenmarkt, terwijl anderen verwachten dat het zo'n vaart nog niet direct zal lopen. Ze wijzen erop dat al vanaf de industriële revolutie werk van mensen wordt overgenomen door machines en dat dat er tot nu toe nog niet toe heeft geleid dat we massaal minder zijn gaan werken. Hoewel er in het verleden banen verloren zijn gegaan, zijn er ook nieuwe banen bijgekomen en is de invulling van sommige beroepen veranderd.

Sommige taken zijn gemakkelijker over te nemen door AI-toepassingen dan andere. Zo zijn beroepen met veel sociale interactie, creativiteit en onvoorspelbaarheid lastiger te automatiseren dan banen met repetitief, voorspelbaar werk. Het lijkt er

dus op dat beroepen als leraar, verpleger, dokter, onderzoeker, ondernemer en ontwerper langer zullen meegaan dan banen als chauffeur (taxi, bus, vrachtwagen), schoonmaker, magazijnmedewerker en accountant. Een risico van het verdwijnen van veel banen is dat het de ongelijkheid in de samenleving verhoogt. Aan de andere kant kan het juist positief zijn dat de samenleving van zwaar, saai en vervelend werk wordt ontlast. Het automatiseren van werk is dus niet per se een slechte zaak, maar als er op grote schaal banen gaan verdwijnen is het wel noodzakelijk om serieus na te gaan denken over andere manieren om geld te verdelen, tijd in te vullen en betekenis te geven aan ons leven.

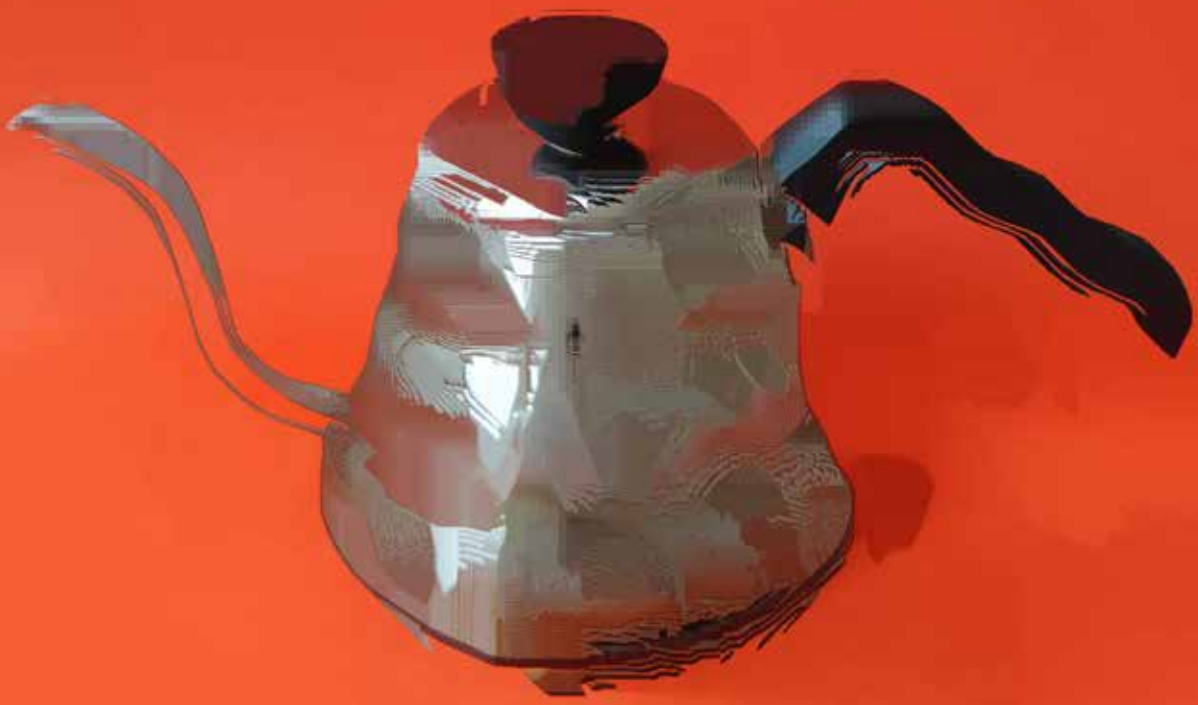
3.4 Deelconclusie

In dit hoofdstuk zijn zowel positieve als negatieve maatschappelijke gevolgen van AI-toepassingen besproken. Aan de ene kant bevorderen AI-toepassingen waarden als gemak, gezondheid, veiligheid, zelfstandigheid en duurzaamheid, maar aan de andere kant gaat dat soms ten koste van veiligheid, privacy, gelijkwaardigheid, vrijheid en creativiteit. Een groot deel van de negatieve impact van AI-toepassingen valt te verklaren door een van de twee volgende verschijnselen.

Ten eerste hebben AI-toepassingen met een lovenswaardig doel soms ongewenste bijeffecten. Bijvoorbeeld, een AI-toepassing die ouderen monitort is ervoor bedoeld om ouderen langer thuis te laten wonen en zo hun zelfstandigheid te vergroten. Grotere zelfstandigheid voor ouderen is een positieve implicatie van deze AI-toepassing. Echter, deze specifieke oplossing kan (onbedoeld) een inbreuk vormen op privacy van ouderen, doordat de ouderen de hele dag in de gaten worden gehouden. Het komt regelmatig voor dat er (AI-)toepassingen worden gecreëerd waarbij dergelijke onbedoelde bijeffecten van tevoren niet zijn voorzien.

Ten tweede kunnen veel negatieve gevolgen van AI-toepassingen worden verklaard doordat gebruikers vaak weinig mogelijkheden hebben tot interactie met de toepassing. AI-toepassingen die informatie filteren, zoals nieuws en zoekresultaten, zijn erg handig, maar deze toepassingen kunnen ook onjuiste informatie verspreiden of een vertekend wereldbeeld scheppen, wat vervolgens kan leiden tot polarisatie. Verbeterde mogelijkheid tot interactie met dergelijke toepassingen zou deze nadelige uitkomsten kunnen verminderen, bijvoorbeeld doordat mensen inzicht hebben in hoe een algoritme bepaalt welk nieuws zij te zien krijgen of doordat mensen in het systeem kunnen aangeven dat ze gevarieerde nieuwsberichten willen ontvangen.

Negatieve implicaties van AI-toepassingen door beide hierboven genoemde oorzaken kunnen (deels) worden voorkomen, door het ontwerp- en ontwikkelproces van AI-toepassingen anders in te richten. In het volgende hoofdstuk ga ik in op de rol die ontwerpers daarin kunnen spelen.



HOOFDSTUK 4

Verantwoord ontwerp

Kunstmatige intelligentie is een krachtig middel, met veel impact op mensen en de samenleving. Zoals beschreven in het vorige hoofdstuk kunnen AI-toepassingen de samenleving veel goeds, maar ook minder goede dingen kan brengen. In het eerste hoofdstuk is al genoemd dat dit afhangt van welke AI-toepassingen er ontworpen en toegepast worden. Hieruit volgt dat diegenen die AI-toepassingen creëren, grote invloed hebben op de ethische implicaties ervan. Ook is al genoemd dat deze invloed een verantwoordelijkheid met zich meebrengt voor de makers van AI-toepassingen (Friedman et al., 2013; Nissenbaum, 2001; Van den Hoven, 2007) en dat deze makers baat hebben bij concrete handvatten om invulling te geven aan deze verantwoordelijkheid.

In dit hoofdstuk ga ik dieper in op de manier waarop makers van AI-toepassingen, met name ontwerpers, invulling kunnen geven aan hun verantwoordelijkheid. Na het bespreken van de rol van ontwerpers in het creëren van AI-toepassingen, ga ik in dit hoofdstuk in op ontwerpen met kunstmatige intelligentie. Uit literatuur blijkt dat er in het ontwerpvakgebied nog relatief weinig bekend is over dit onderwerp, laat staan dat er veel kennis is opgebouwd over hoe dat op verantwoorde wijze zou moeten gebeuren. In de rest van dit hoofdstuk bespreek ik op welke wijze ontwerpers zouden kunnen bijdragen aan het verantwoord ontwerpen van AI-toepassingen en wat daar voor nodig is.

4.1 De rol van ontwerp

Voor het creëren van AI-toepassingen zijn verschillende expertises nodig. AI-toepassingen bestaan altijd uit software en soms ook uit hardware (bijvoorbeeld bij robots). Er zijn programmeurs nodig om de software te ontwikkelen en ingenieurs om de hardware te bouwen. Er is ook kennis over kunstmatige intelligentie nodig om de toepassing intelligent te maken, bijvoorbeeld van AI-wetenschappers, informatici of datawetenschappers. Daarnaast is er ontwerpkennis nodig om de gebruikers-ervaring van het product of de dienst vorm te geven. Dit is de expertise van *User eXperience (UX)* ontwerpers en interactieontwerpers. Vaak is er ook kennis nodig over het domein waarin de toepassing wordt gebruikt, bijvoorbeeld gezondheidszorg of logistiek. Om AI-toepassingen op verantwoorde wijze te ontwerpen, is ook kennis nodig over ethiek en ontwerp.

De kennis uit de verschillende hierboven genoemde gebieden moet samenkomen om tot een AI-toepassing te komen. In de praktijk gebeurt dat op verschillende manieren. In grote projecten komt het voor dat alle expertises door verschillende mensen vertegenwoordigd zijn, maar dat is lang niet altijd het geval. Zo zijn er UX-ontwerpers met programmeerkennis en robotingenieurs die ook wel iets weten van interactieontwerp. Ook komt het vaak voor dat er bij het creëren van een AI-toepassing gebruik wordt gemaakt van reeds bestaande producten of diensten, zoals een template voor een website, software die gezichten kan herkennen of een reeds bestaande robotarm. Het gegeven dat vaak meerdere mensen op verschillende tijdstippen en plekken aan *één* AI-toepassing bijdragen, maakt het er niet gemakkelijker op om op verantwoorde wijze toepassingen te creëren, maar betekent niet dat er geen aandacht aan moet worden besteed.

In het vakgebied van kunstmatige intelligentie wordt nagedacht over hoe kunstmatige intelligentie op een ethisch verantwoorde manier kan worden toegepast. De termen *responsible AI* en *AI for good* worden vaak in deze context gebruikt. Binnen het AI-vakgebied wordt relatief vaak gewerkt aan oplossingen voor directe gevolgen van geïsoleerde problemen, bijvoorbeeld naar wat een zelfrijdende auto moet doen wanneer een ongeluk dreigt (Conitzer, Sinnott-Armstrong, Borg, Deng & Kramer, 2017). Er wordt minder vaak gekeken naar de impact van die oplossingen op gebruikers, andere stakeholders en de samenleving. Ontwerpers hebben hier meer ervaring mee. Het AI-vakgebied zou kunnen profiteren van deze ervaring van ontwerpers.

De verantwoordelijkheid voor de ethische implicaties van AI-toepassingen geldt voor iedereen die bijdraagt aan het creëren van dergelijke toepassingen. Deze openbare les richt zich met name op ontwerpers. Het vorige hoofdstuk eindigt met de conclusie dat veel van de negatieve implicaties van AI-toepassingen ontstaan ofwel als ongewenst bijeffect ofwel door een gebrek aan interactie. Juist het ontwerpdomein beschikt over kennis en vaardigheden die kunnen helpen bij het voorkomen van negatieve impact van AI-toepassingen door (een van) deze twee oorzaken. Zo worden in de ontwerpliteratuur verschillende aanpakken beschreven voor het meenemen van ethische implicaties tijdens het ontwerpproces. Daarnaast is het gebied van interactieontwerp gericht op het ontwerpen van goede gebruikersinteractie.

Er worden vele soorten producten en diensten ontworpen. Het deelgebied van ontwerp dat met name van belang is voor het ontwerpen van AI-toepassingen is UX-ontwerp en gaat over het ontwerpen van digitale, interactieve producten en diensten (Garrett, 2010; Hartson & Pyla, 2012). UX-ontwerpers richten zich nadrukkelijk op de ervaring die een product of dienst voor de gebruiker creëert. Technologie wordt daarbij gezien als middel om een betekenisvolle ervaring te

scheppen. Een essentieel onderdeel van UX-ontwerp is het ontwerpen van de interactie tussen een gebruiker en het product of de dienst (Preece, Rogers & Sharp, 2015). Deze interactie kan verschillende vormen hebben, bijvoorbeeld via tekst, afbeeldingen (GUI: graphical user interface), spraak (CUI: conversational user interface) of gebaren.

Er is geen simpel stappenplan voor verantwoord ontwerpen. Dat komt onder andere doordat het ontwerpproces er verschillend uitziet bij verschillende projecten. Ontwerpers hebben vaak te maken met ontwerpopgaven waarbij niet alle aspecten van het probleem duidelijk zijn aan het begin van het ontwerpproces, de gevolgen van verschillende oplossingen moeilijk te overzien zijn en de interpretatie van het probleem kan verschuiven tijdens het ontwerpproces (Dorst, 2017). Een ontwerpproces is vaak rommelig, vooral aan het begin van het proces. Door deze complexiteit en onvoorspelbaarheid van ontwerpopdrachten bestaat er geen stappenplan of recept om gegarandeerd tot een goed resultaat te komen.

Ook al verschillen ontwerpprocessen onderling, er zijn verschillende activiteiten te onderscheiden die regelmatig voorkomen in een ontwerpproces:

1. **empathize**: inleven in het probleem, de belanghebbenden en de toepassingscontext;
2. **define**: definiëren van het probleem en de randvoorwaarden van de toepassing;
3. **ideate**: ideeën verzinnen voor de oplossing;
4. **prototype**: prototypes maken van de oplossing, uiteenlopend van een simpele schets tot een hoogwaardig prototype;
5. **test**: testen van het prototype met de doelgroep.

Het ontwerpproces is vaak iteratief, wat wil zeggen dat er meerdere cycli worden doorlopen waarin een prototype wordt bedacht, gemaakt en getest. Hierbij wordt het prototype steeds gedetailleerder uitgewerkt en dienen de uitkomsten van de voorgaande cyclus als startpunt voor het verbeteren van het prototype.

4.2 Ontwerpen met kunstmatige intelligentie

De term ‘ontwerpen met kunstmatige intelligentie’ is op twee manieren op te vatten. In de eerste betekenis, ontwerpen *met behulp van* kunstmatige intelligentie, wordt kunstmatige intelligentie ingezet tijdens het ontwerpproces om tot betere concepten te komen. Dat kan bijvoorbeeld door data over hoe gebruikers een bestaand product gebruiken, te analyseren met kunstmatige intelligentie. Het gebruik van data tijdens het ontwerpproces heet ook wel *data-driven design*. In deze openbare les gaat het echter om een andere betekenis van ‘ontwerpen met kunstmatige intelligentie’, namelijk het ontwerpen, al dan niet met behulp van kunstmatige intelligentie, van AI-toepassingen. AI-toepassingen kunnen een of

meerdere complexe taken zelfstandig uitvoeren, zoals een robot die kan navigeren of een webapplicatie die gepersonaliseerde aanbevelingen kan doen. Ontwerpers van dergelijke toepassingen verwerken dus kunstmatige intelligentie in de producten en diensten die ze ontwerpen. Kunstmatige intelligentie vormt daarmee als het ware een nieuw soort 'ontwerpmateriaal' waar ontwerpers mee kunnen werken (Holmquist, 2017).

Hoewel ontwerpers steeds vaker te maken krijgen met kunstmatige intelligentie als ontwerpmateriaal, zijn nog maar weinig ontwerpers daar goed op voorbereid. Uit een onderzoek in de VS, Engeland en Denemarken kwam naar voren dat 63% van de UX-designers in hun werk te maken kreeg met het ontwerpen van producten met machine learning (Dove, Halskov, Forlizzi & Zimmerman, 2017). Uit hetzelfde onderzoek bleek dat UX-designers slechts beperkt zijn voorbereid op het ontwerpen met machine learning. Slechts 8% van de ondervraagden is tijdens zijn opleiding voorbereid op het werken met machine learning. Ook gaven ondervraagden aan dat er een gebrek is aan 'design practices', zoals methodes, technieken en gereedschappen voor het werken met machine learning.

Er is onderzoek is nodig om methoden en technieken te ontwikkelen voor het omgaan met kunstmatige intelligentie als ontwerpmateriaal (Dove et al., 2017; Yang, Zimmerman, Steinfeld & Tomasic, 2016). Hoe brainstorm je bijvoorbeeld over een product dat zichzelf kan aanpassen? Hoe kun je snel een prototype maken van een intelligent product? Waar moet je op letten bij het testen?

Ook is meer kennis nodig over wat ontwerpers moeten weten en kunnen om met kunstmatige intelligentie te werken, en hoe ontwerp-opleidingen toekomstige ontwerpers daar het best op kunnen voorbereiden. Uit interviews met UX-ontwerpers die ervaren zijn in het werken met machine learning bleek dat ze er slechts abstracte kennis van hebben en veel samenwerken met datawetenschappers (Yang, Scuito, Zimmerman, Forlizzi & Steinfeld, 2018). Tekstboeken die gebruikt worden op universiteiten gaan vaak in op technische details van machine learning. Dit lijkt erop te wijzen dat ontwerp-opleidingen die aandacht besteden aan machine learning dat niet op de juiste manier doen. Ook wordt in veel ontwerp-opleidingen nog slechts beperkt aandacht besteed aan de samenwerking van ontwerpers met datawetenschappers, informatici of AI-onderzoekers.

Er is dus nog geen breed gedragen aanpak voor het ontwerpen van toepassingen met kunstmatige intelligentie. Laat staan dat er een geïntegreerde benadering bestaat voor hoe ontwerpers bij het ontwerpen van toepassingen met kunstmatige intelligentie rekening kunnen houden met de ethische gevolgen van die toepassingen. Dus naast het überhaupt werken met kunstmatige intelligentie,

hebben ontwerpers ondersteuning nodig bij het *verantwoord* ontwerpen met kunstmatige intelligentie. Het ontwerpvakgebied beschikt over een grote kennisbasis van aanpakken, methodes en inzichten waaruit geput kan worden om zo'n geïntegreerde benadering te ontwikkelen.

4.3 Interactie met AI-toepassingen

In hoofdstuk 3 is geconcludeerd dat veel negatieve gevolgen van AI-toepassingen, naast dat het onbedoelde bijeffecten zijn van goedbedoelde producten of diensten, ontstaan doordat gebruikers weinig interactiemogelijkheden hebben met AI-toepassingen. Bij verantwoord ontwerp van AI-toepassingen is het dus belangrijk om te zorgen dat gebruikers wel over voldoende interactiemogelijkheden beschikken. Dat vereist dat gebruikers inzicht in en controle hebben over AI-toepassingen (Johnson, Bradshaw & Feltovich, 2018; Johnson et al., 2014)

Inzichtelijkheid

In hoofdstuk 2 is beschreven dat alle AI-toepassingen waarnemen, redeneren en handelen. Inzicht geven in AI-toepassingen kan gaan over alle drie van deze onderdelen. Ten eerste kan een AI-toepassing duidelijkheid geven over wat en wanneer hij waarneemt. Het rode lampje op een videorecorder geeft bijvoorbeeld aan dat de camera op dat moment aan het opnemen is. Ten tweede kan een AI-toepassing inzicht geven in haar redeneerprocessen, dit wordt ook wel *explainable AI* genoemd (Harbers, 2011; Van Lent, Fisher, & Mancuso, 2004). Een voorbeeld hiervan is een expertsysteem dat uitlegt hoe het tot een bepaald advies is gekomen of een aanbevelingsalgoritme dat inzicht geeft in welk profiel of binnen welke categorie hij je heeft geplaatst. Tot slot kan een AI-toepassing inzicht geven in zijn handelen door van tevoren of tijdens het uitvoeren van een actie duidelijk te maken wat hij gaat doen of doet. Een huishoudrobot kan bijvoorbeeld aankondigen dat hij naar de keuken gaat om een drankje op te halen. Zo wordt een gebruiker niet onnodig verrast door de robot en hoeft zij niet te gissen naar wat de robot van plan is.

Het inzichtelijk maken van kunstmatige intelligentie betekent niet dat we de onderliggende code hoeven te zien. Sterker nog, voor de meeste mensen wordt kunstmatige intelligentie daar helemaal niet inzichtelijk van. Inzichtelijk maken hoeft zelfs niet te betekenen dat we de hele werking van een AI-toepassing op abstract niveau snappen (Lipton, 2018). Ook dit kan ingewikkeld zijn en het is lang niet altijd de kennis waarnaar we op zoek zijn. Vaak is het voldoende om te begrijpen wat de doorslaggevende factor is geweest om tot een bepaalde uitkomst te komen. Of juist waarom het systeem niet voor een voor de hand liggend alternatief heeft gekozen. Om het gedrag van AI-toepassingen goed uit te leggen, is het met name belangrijk om te kijken naar hoe mensen intelligent gedrag begrijpen en verklaren (Malle, 2006).

Technisch gezien is het inzichtelijk maken van 'de doorslaggevende factor' niet altijd even eenvoudig. Sommige AI-benaderingen lenen zich beter voor explainable AI dan andere. Symbolische kunstmatige intelligentie (zie hoofdstuk 2) maakt gebruik van symbolische representaties die mensen iets zeggen. Een doel als 'huis schoonmaken' zou dan letterlijk in de code kunnen staan. Dergelijke representaties lenen zich goed voor het automatisch verklaren van gedrag (Harbers, 2011). Veel AI-toepassingen maken echter gebruik van machine learning, waarbij 'redeneren' vaak op veel impliciete wijze gerepresenteerd wordt. In een neuraal netwerk kan het bijvoorbeeld zo zijn dat één stukje kennis over het hele netwerk verspreid is. Er wordt onderzoek gedaan naar hoe uitkomsten van machine learning algoritmes verklaard kunnen worden (Burrell, 2016; Samek, Wiegand, & Müller, 2017)
Inzicht geven hoeft niet altijd de vorm van tekstuele uitleg te hebben, het kan ook gaan om een lampje, een piepje of het uiterlijk van een AI-toepassing. Een virtuele agent of robot communiceert met haar uiterlijk over waartoe zij in staat is.
Uit onderzoek blijkt dat mensen hogere verwachtingen hebben van robots die er geavanceerder en menselijker uitzien dan van robots met een eenvoudig, cartoon-achtig uiterlijk (Goetz, Kiesler & Powers, 2003; Harbers, Peeters & Neerincx, 2017). Onderzoek wijst zelfs uit dat het uiterlijk van een robot invloed heeft op hoe mensen moreel gedrag van die robot beoordelen: als een robot er geavanceerder uitziet, vinden mensen sneller dat de robot er zelf schuld aan heeft als hij een fout maakt (Malle, Scheutz, Arnold, Voiklis & Cusimano, 2015).

Betekenisvolle menselijke controle

Voor alle huidige AI-toepassingen geldt dat er altijd een mens is die controle heeft, er is altijd een mens die een robot of algoritme 'aan' zet. De mate waarin we vervolgens opdrachten moeten geven of bij kunnen sturen, verschilt sterk. Soms is een AI-toepassing zo ontworpen dat er eerst goedkeuring van een mens nodig is voordat de toepassing een actie uitvoert, bijvoorbeeld bij een militaire robot die op een vijand wil schieten. Dit wordt ook wel *human-in-the-loop* genoemd. Een ontwerp waarbij een AI-toepassing zelfstandig handelt, maar een mens nog wel in kan grijpen, heet *human-on-the-loop*. Een voorbeeld hiervan is een slimme thermostaat. De thermostaat bepaalt zelf wanneer de verwarming aangaat, maar het is altijd mogelijk om de temperatuur handmatig aan te passen. Een AI-toepassing die zelfstandig handelt en mensen geen mogelijkheid biedt om invloed uit te oefenen, vormt een *human-out-of-the-loop* systeem. Een zelfrijdende auto zonder stuur is hier een voorbeeld van.

Te weinig mogelijkheid tot controle op AI-toepassingen is ongewenst omdat mensen soms willen bijsturen, bijvoorbeeld als er een veiligheidsrisico ontstaat, of omdat sommige beslissingen zo belangrijk zijn dat we deze niet aan een machine over willen laten. Te veel controle is echter ook niet gewenst. AI-toepassingen kunnen ons allerlei saaie en zware taken uit handen nemen en we zitten er niet op

te wachten om bij elk klein stapje dat een systeem maakt in te moeten grijpen. Bij het creëren van mogelijkheden van gebruikers om controle uit te oefenen op het systeem moet dus gezocht worden naar een middenweg. Dit wordt vaak *meaningful human control* genoemd (Santoni de Sio & Van den Hoven, 2018).

Ook het uitoefenen van controle kan slaan op het waarnemen, redeneren en handelen van een AI-toepassing. Bij het waarnemen zou een gebruiker bijvoorbeeld de mogelijkheid kunnen krijgen om aan te geven welke informatie een algoritme wel en niet mee moet nemen om van te leren. Het kan bijvoorbeeld zijn dat iemand het leuk vindt om kattenfilmpjes of 'slechte' Netflixseries te bekijken, maar daar geen aanbevelingen over wil krijgen. Het zou handig zijn om in zo'n geval het leermechanisme even te pauzeren. Een ander voorbeeld is dat iemand die niet wil dat zijn of haar afkomst meegenomen wordt door een algoritme dat vacatures aanbeveelt, de mogelijkheid heeft om die factor uit te sluiten. Dit vereist uiteraard inzicht bij de gebruiker in welke factoren het algoritme allemaal meeneemt.

Gebruikers zouden meer controle op het redeneren van AI-toepassingen kunnen krijgen als ze de mogelijkheid hebben om bij een algoritme aan te geven wat ze belangrijk vinden. Sommige gebruikers van sociale media willen bijvoorbeeld vooral nieuws te zien krijgen terwijl andere gebruikers vooral op de hoogte willen blijven van wat hun vrienden doen. Het zou prettig zijn als het algoritme daar rekening mee zou houden. De mogelijkheid om voorkeuren aan een algoritme door te geven zou tegelijk voor meer inzicht en bewustzijn zorgen, doordat zo'n optie gebruikers eraan herinnert dat er op de achtergrond een algoritme aan het werk is. Een ander voorbeeld van controle op redeneren is dat een algoritme of robot de afleidingen die hij heeft gemaakt, bijvoorbeeld "je houdt van tennis" of "je hebt honger", ter verificatie voorlegt aan de gebruiker. De gebruiker kan dan aangeven of de aanname klopt en op die manier meer controle over de agent krijgen.

Controle over het handelen van een AI-toepassing gaat vooral over de mogelijkheid om in te grijpen. In een zelfrijdende auto zou dat kunnen zijn doordat de bestuurder te allen tijde op de rem kan trappen of het stuur over kan nemen. Een ander voorbeeld is dat een virtuele agent die spullen koopt en verkoopt, om goedkeuring vraagt van de gebruiker voordat hij tot een transactie overgaat.

Zowel voor het vergroten van inzichtelijkheid als het verschaffen van mogelijkheden om controle uit te oefenen, is een goed ontwerp van de interactie vereist. Ontwerpers kunnen hierin een grote bijdrage leveren en daarmee een deel van de negatieve gevolgen van AI-toepassingen wegnemen. Op dit moment bestaan er voor veel AI-toepassingen nog geen standaardoplossingen voor het ontwerpen van inzichtelijkheid en controle. Onderzoek kan zich richten op het in kaart brengen van goede oplossingen en het daaruit extraheren van *design practices,* zodat ontwerpers niet bij elke AI-toepassing vanaf nul hoeven te beginnen.

4.4 Ontwerpaanpak

Om verantwoord te ontwerpen is het belangrijk om al tijdens het ontwerpproces rekening te houden met de (onbedoelde) ethische implicaties van een oplossing. Door hier tijdens het ontwerpproces al aandacht aan te besteden, kunnen er bewust keuzes worden gemaakt en is er nog voldoende ruimte om aanpassingen te doen. Achteraf, als een toepassing al in werking is getreden, is het een stuk lastiger om ongewenste implicaties weg te nemen.

Er zijn verschillende ontwerpmethodologieën ontwikkeld die op systematische wijze rekening houden met ethische implicaties van ontwerpen, zoals *critical design* (Bardzell, Bardzell, Forlizzi, Zimmerman & Antanitis, 2012), *reflective design* (Sengers, Boehner, David & Kaye, 2005) en *value sensitive design* (Friedman et al., 2013). Value sensitive design (VSD) is daarvan de meest uitgewerkte aanpak en biedt verschillende methodes en technieken om invulling te geven aan verantwoord ontwerpen (Friedman, Hendry & Borning, 2017).

VSD houdt op systematische wijze rekening met menselijke waarden bij het ontwerpen van technologie. Een waarde wordt daarbij gedefinieerd als datgene wat een persoon of groep personen belangrijk vindt in het leven (Friedman et al., 2017, 2013). Voorbeelden van waarden zijn privacy, vriendschap, gezondheid, traditie, veiligheid, vrijheid, gelijkwaardigheid, openheid en creativiteit.

VSD maakt onderscheid tussen waarden die expliciet ondersteund worden door technologie, waarden van stakeholders en waarden van ontwerpers. Een expliciet ondersteunde waarde is bijvoorbeeld veiligheid als het gaat om een surveillancesysteem of vriendschap als het gaat om een socialmediaplatform. Stakeholders kunnen indirecte of directe stakeholders zijn, waarbij directe stakeholders zelf interactie hebben met de technologie en indirecte stakeholders zelf geen interactie hebben met de technologie maar er wel door worden beïnvloed. Patiënten zijn bijvoorbeeld indirecte stakeholders van het elektronisch patiëntendossier. Waarden van directe en indirecte stakeholders kunnen ondersteund of ondermijnd worden door technologie. Een belangrijk begrip in VSD is een *value tension*. Hiervan is sprake als twee of meer waarden met elkaar in conflict zijn. Dat kan gaan om een waardenconflict binnen een stakeholdergroep (gezondheid versus gemak), tussen stakeholdergroepen (privacy versus openheid) of tussen de ontwerper en een stakeholdergroep (esthetiek versus gebruiksgemak).

VSD beschikt over een verzameling van concrete methoden en technieken om stakeholders, waarden en effecten van technologie daarop te identificeren, analyseren en voorspellen, zoals de waarde- en stakeholderanalyse (Friedman et al., 2017), value scenarios (Nathan, Klasnja & Friedman, 2007), envisioning cards (Friedman & Hendry, 2012) en value stories (Harbers, Detweiler & Neerincx, 2015).

Daarnaast is er inzicht en kennis opgedaan door VSD in verschillende contexten toe te passen, bijvoorbeeld voor blinde en dove ov-reizigers (Azenkot et al., 2011), daklozen (Woelfer, Iverson, Hendry, Friedman & Gill, 2011) en treinverkeersleiders (Harbers & Neerincx, 2017). VSD is niet 'af', maar vormt een rijke bron van kennis en ervaring waaruit ontwerpers kunnen putten om verantwoord te ontwerpen.

Naast algemene methodologieën voor verantwoord ontwerpen zoals VSD zijn er aanpakken die zich richten op een specifiek ethisch aspect. Een bekend voorbeeld van zo'n aanpak is *privacy by design* (Cavoukian, 2011; Schep, 2016), waarbij een vertaalslag is gemaakt van abstractere ethische principes over privacy naar concrete richtlijnen voor ontwerpers. Voorbeelden van deze richtlijnen zijn dat een ontwerp erop gericht moet zijn om privacy-inbreuk te voorkomen in plaats van het achteraf op te lossen, dat default-instellingen altijd zo zijn ingesteld dat ze privacy beschermen en dat er securitymaatregelen worden genomen voor elke stap waarin data worden verwerkt. Het toepassen van privacy by design resulteert in privacy-vriendelijke producten en diensten.

Andere voorbeelden van aanpakken voor verantwoord ontwerpen die zich richten op een specifiek ethisch aspect zijn *sustainable design* (Walker, 2012), dat zich richt op het ontwerpen van producten en diensten die sociale, economische en ecologische duurzaamheid bevorderen, en *inclusive design* (Clarkson, Coleman, Keates & Lebbon, 2013), dat focust op het ontwerpen van producten en diensten die toegankelijk en bruikbaar zijn voor alle mensen, zoals ouderen én jongeren, personen met én zonder een handicap, mannen én vrouwen, mensen van verschillende afkomst en mensen die verschillende talen spreken.

Bovenstaande aanpakken helpen met het voorspellen en in kaart brengen van mogelijke ethische gevolgen van technologie tijdens het ontwerpproces, en met het voorkomen van (onbedoelde) negatieve impact van technologie op de samenleving. Hiermee zijn ze van grote waarde voor verantwoord ontwerpen van AI-toepassingen. Geen van de benaderingen richt zich echter specifiek op het ontwerpen van technologie met kunstmatige intelligentie. Toekomstig onderzoek is nodig om te bepalen in hoeverre bovenstaande aanpakken geschikt zijn voor het ontwerpen van AI-toepassingen en op welke manier deze bestaande aanpakken uitgebreid of aangepast zouden kunnen worden om ontwerpers te ondersteunen bij het verantwoord ontwerpen van AI-toepassingen.

4.5 Ontwerpcontext

Naast de ontwerpaanpak die wordt gevolgd tijdens het ontwerpen, heeft de context waarin ontworpen wordt, ook invloed op ethische implicaties van een AI-toepassing. Veel AI-toepassingen worden ontworpen in bedrijven of door zelfstandigen. In die context wordt bijvoorbeeld bepaald welke opdrachten wel en

niet worden aangenomen, op welke manier ontwerpprocessen worden ingericht en welke reeds bestaande AI-componenten er wel of niet in een AI-toepassing worden verwerkt. Bedrijven kunnen ervoor kiezen om zich te richten op het oplossen van maatschappelijke opgaven, zogenoemde *grand societal challenges*. Ontwerp leent zich uitstekend voor het onderzoeken en oplossen van dergelijke complexe uitdagingen (Rutten & Schijvens, 2015). Bedrijven kunnen ook de expliciete keuze maken om naast economische belangen ook andere belangen mee te nemen, zoals duurzaamheid en menselijk welzijn (Raworth, 2017). Maatschappelijk Verantwoord Ondernemen Nederland is bijvoorbeeld een netwerk van bedrijven die dit nastreven (mvonederland.nl).

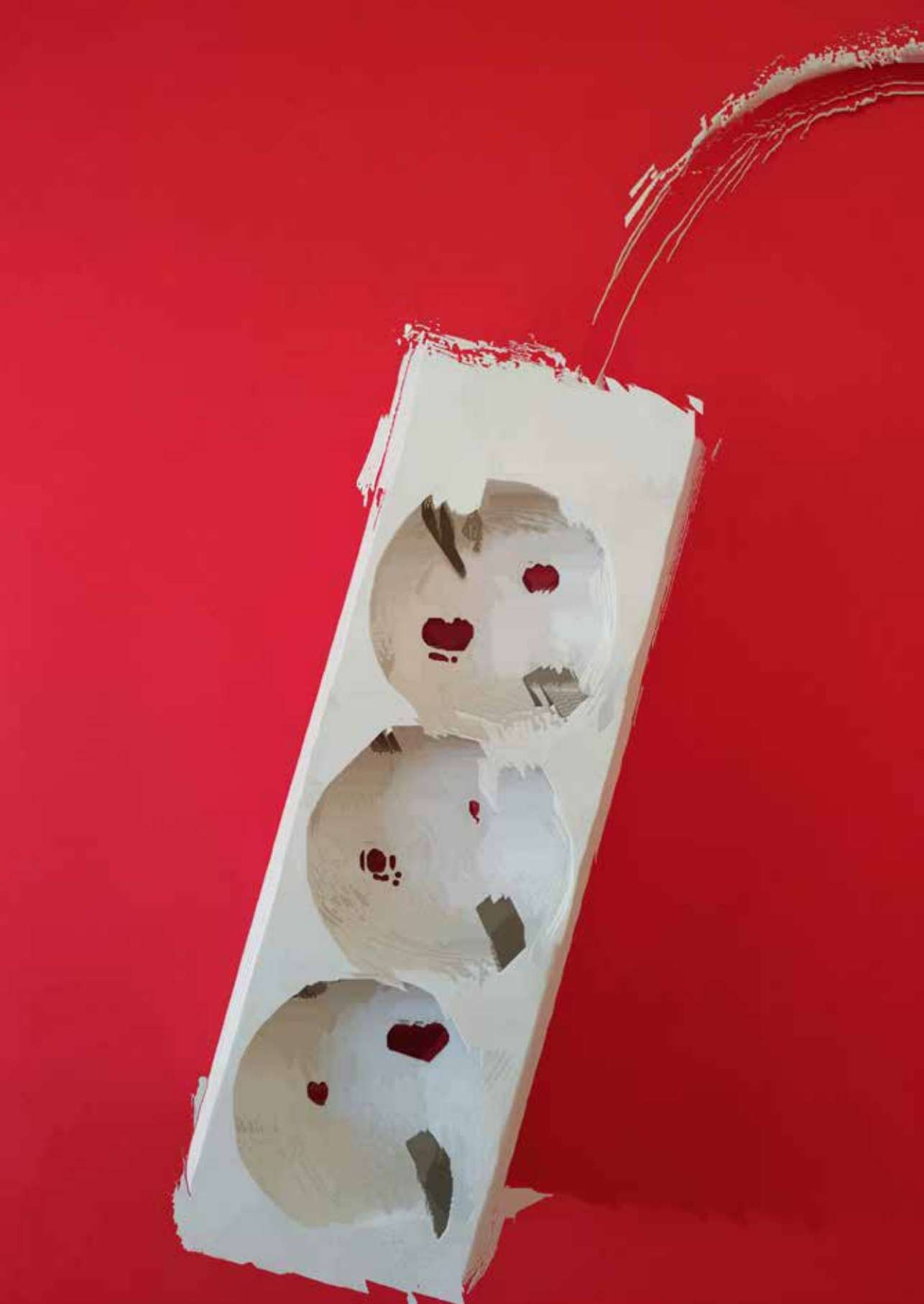
Bedrijven die AI-toepassingen ontwerpen en ontwikkelen, kunnen op verschillende manieren verantwoord ontwerp bevorderen. Zo kan een bedrijf zijn werknemers stimuleren om aandacht te besteden aan verantwoord ontwerp, bijvoorbeeld door te werken met een ontwerpaanpak die daar aandacht aan besteedt of door gebruik te maken van open source software. Een zelfstandige of een bedrijf kan er ook voor kiezen om sommige opdrachten wel en andere niet aan te nemen, of sommige producten wel en andere niet te gaan ontwikkelen. De samenleving heeft bijvoorbeeld meer baat bij een AI-toepassing die energie bespaart dan een AI-toepassing die de aandacht van mensen langer vast weet te houden al waren ze dat eigenlijk niet van plan. Rekening houden met maatschappelijke implicaties kan bedrijven zelfs een competitief voordeel opleveren (Gurzawska et al., 2017).

Bedrijven en ontwerpers kunnen ook invloed uitoefenen op de samenstelling van ontwerp- en ontwikkelteams (Shilton & Anderson, 2017). Ontwerpers hebben eigen meningen en voorkeuren die van invloed zijn op hun werk. Dit is onvermijdelijk, maar het helpt om deze eigen voorkeuren van ontwerpers expliciet te maken en er rekening mee te houden. Om ervoor te zorgen dat de belangen van stakeholders voldoende behartigd worden, kunnen er bijvoorbeeld stakeholders bij het ontwerpproces worden betrokken of zelfs aan het ontwerpteam worden toegevoegd (Schuler & Namioka, 1993). Ook kan het waardevol zijn om te zorgen voor diversiteit in een ontwerpteam (Fletcher-Watson et al., 2018). Zo werd een algoritme voor het herkennen van gezichten vooral getraind met foto's van witte mensen, waardoor gezichten van zwarte mensen minder goed herkend werden (Garcia, 2016). Dit zou waarschijnlijk niet gebeurd zijn wanneer er sprake was geweest van een gemengd team van witte en zwarte ontwerpers.

Naast bedrijven hebben overheden invloed op ontwerpers die AI-toepassingen ontwerpen. Wet- en regelgeving kan de prikkels geven die verantwoord ontwerpen of maatschappelijk verantwoord ondernemen aantrekkelijker maken voor bedrijven (Gurzawska et al., 2017; Harbers et al., 2018). Een voorbeeld van zo'n maatregel is het introduceren van standaarden of keurmerken, bijvoorbeeld voor producten die

aan bepaalde beveiligingseisen of inzichtelijkheidseisen voldoen. In hoofdstuk 3 is al genoemd dat in mei 2018 de nieuwe Europese privacywetgeving van kracht is gegaan – de Algemene Verordening Gegevensbescherming (AVG). Deze wetgeving geeft burgers bijvoorbeeld het recht om bij bedrijven op te vragen welke informatie deze bedrijven over hen hebben opgeslagen en om te verzoeken die data te verwijderen. Dit is direct van invloed op het ontwerp van AI-toepassingen die persoonsgegevens verwerken.

In de softwareontwikkeling is *usability* of gebruiksvriendelijkheid van software gemeengoed geworden. Het vakgebied is ervan doordrongen dat gebruiksgemak belangrijk is, er zijn standaarden voor ontwikkeld en het wordt breed toegepast. Op vergelijkbare wijze zou de verantwoordelijkheid van ontwerpers en het belang van ethiek in ontwerpen door kunnen dringen tot de ontwerp- en ontwikkelwereld. Dit vereist samenwerking tussen bedrijven, overheden en onderzoeks- en onderwijsinstellingen. In de medische wetenschappen is al veel langer aandacht voor ethiek en zijn processen rondom ethische beslissingen goed geregeld. Het vakgebied van ontwerp zou hier een voorbeeld aan kunnen nemen. Wellicht sluiten toekomstige ontwerpers, net als geneeskundestudenten, hun studie af met een eed, in hun geval voor verantwoord ontwerp.



HOOFDSTUK 5

Onderzoeksagenda

In de voorgaande hoofdstukken heb ik een overzicht gegeven van de huidige ontwikkelingen op het raakvlak van kunstmatige intelligentie, ethiek en ontwerp. Daaruit blijkt dat er verschillende obstakels zijn bij het ontwerpen met en toepassen van kunstmatige intelligentie zodanig dat het de samenleving ten goede komt. Het lectoraat *Artificial Intelligence & Society* wil zich de komende jaren daarom richten op het wegnemen van deze obstakels door onderzoek te doen naar verantwoord ontwerpen van toepassingen met kunstmatige intelligentie. Dit laatste hoofdstuk geeft een overzicht van hoe en met wie het lectoraat dat wil aanpakken en wat het daarmee hoopt te bereiken.

5.1 Context van het onderzoek

Het onderzoek van het lectoraat Artificial Intelligence & Society vindt plaats in Kenniscentrum Creating 010 van Hogeschool Rotterdam. Kenniscentrum Creating 010 doet praktijkgericht onderzoek naar maatschappelijke transformaties die samenhangen met digitalisering en ontwikkelingen in informatie- en communicatietechnologie. Daarbij stelt Kenniscentrum Creating 010 mensen in hun sociale context centraal en wordt aandacht besteed aan de rol van ontwerpers en ontwikkelaars van technologie en van degenen die deze technologie toepassen. Het Privacy Lab van Kenniscentrum Creating 010 doet onderzoek naar het verder ontwikkelen en toepassen van privacy by design. Het kenniscentrum vormt daarmee een zeer geschikte omgeving voor het onderzoek van het lectoraat.

Relatie met onderwijs

Kenniscentrum Creating 010 werkt nauw samen met het Instituut voor Communicatie, Media en Informatietechnologie (CMI) van Hogeschool Rotterdam. Het CMI-instituut verzorgt de volgende vijf opleidingen: Communicatie, Informatica, Technische Informatica, Communication and Multimedia Design (CMD) en Creative Media and Game Technologies (CMGT). Het onderzoek van het lectoraat Artificial Intelligence & Society is van belang voor al deze opleidingen. Communicatieprofessionals hebben namelijk vaak te maken met algoritmes die een rol spelen bij het verspreiden van content, bijvoorbeeld in sociale media. Daarnaast kunnen algoritmes een rol spelen in het vormgeven van communicatie tussen mensen en machines. Informatici en technische informatici ontwikkelen

software; daarbij maken ze steeds vaker gebruik van kunstmatige intelligentie, ofwel door dit zelf te ontwikkelen of door bestaande AI-oplossingen in hun software te verwerken.

Het lectoraat is in het bijzonder van belang voor de opleidingen CMD en CMGT. Beide opleidingen vallen onder het hbo-domein Creative Technologies en leiden studenten op tot ontwerpers van interactieve technologie (creativetechnologies.nl). Studenten Creative Technologies ontwerpen en maken interactieve, digitale producten en diensten, zoals websites, apps, games, IoT-producten of een combinatie daarvan. De mens staat daarbij centraal. Het hbo-domein Creative Technologies heeft vastgelegd over welke competenties studenten moeten beschikken om hun diploma te halen. De competenties vallen in de volgende vier categorieën:

1. **technologisch**: goede technische kennis en analysevaardigheden, talent voor ontwerpen en prototypen, testen en implementeren;
2. **ontwerpend**: goed onderzoeks- en analysevermogen, in staat om te conceptualiseren en vorm te geven;
3. **organiserend**: ondernemende houding en vaardigheden, projectmatige werkinzet en goede communicatie;
4. **professioneel:** goed lerend en reflecterend vermogen, goed verantwoordelijkheidsgevoel.

Het onderzoek van het lectoraat draagt bij aan het invulling geven aan deze competenties. Studenten Creative Technologies dienen op de hoogte te zijn van actuele technologische ontwikkelingen. Hieronder vallen technische ontwikkelingen binnen het AI-vakgebied. Daarnaast dienen ze op professionele wijze, vanuit een goed verantwoordelijkheidsgevoel te werken. Het onderzoek naar verantwoord ontwerpen zoals beschreven in deze openbare les kan helpen om daar een invulling aan te geven.

Relatie met de beroepspraktijk

Zowel Kenniscentrum Creating 010 als het CMI-instituut werken veel samen met bedrijven uit de regio Rotterdam, met name in de sectoren ICT en creatieve industrie. Voor veel van deze bedrijven zijn de ontwikkelingen op het gebied van kunstmatige intelligentie erg relevant. Dat is bijvoorbeeld te merken aan de afstudeeronderwerpen van studenten Creative Technologies. Steeds meer studenten Creative Technologies studeren af op een onderwerp gerelateerd aan kunstmatige intelligentie. In het nu lopende onderzoeksproject ‘Filterbubbles, nou en?’ van het lectoraat Artificial Intelligence & Society blijkt ook dat verschillende bedrijven te maken krijgen met ethische dilemma’s en interesse tonen voor verantwoord omgaan en inzetten van (intelligente) technologie (Harbers, Mazerant & Voskuyl, 2017). Er liggen dus duidelijke vragen vanuit de beroepspraktijk ten grondslag aan het voorgenomen onderzoek van het lectoraat.

Het is belangrijk om bedrijven te betrekken bij het onderzoek naar verantwoord ontwerpen van AI-toepassingen. Bedrijven zijn de plek waar het uiteindelijk gebeurt. Op dit moment zijn lang niet alle bedrijven goed op de hoogte van de ethische implicaties van AI-toepassingen en de rol die ontwerp daarin speelt. Daarnaast verschillen bedrijven in de mate waarin ze openstaan voor verantwoord ontwerp. Door de belangen van deze bedrijven te verkennen en samen met hen op te trekken, kan het lectoraat ervoor zorgen dat het onderzoek zo relevant mogelijk is voor deze bedrijven. Daarnaast kan samenwerking zorgen voor meer kennis binnen bedrijven over verantwoord ontwerpen en het draagvlak voor verantwoord ontwerp vergroten.

Relatie met overheden

Zoals reeds genoemd in het vorige hoofdstuk zijn ook overheden van invloed op ontwerpers die in de praktijk met kunstmatige intelligentie werken. Kunstmatige intelligentie, machine learning, data, cybersecurity en privacy zijn onderwerpen die hoog op de politieke agenda staan. Overheden kunnen maatregelen betreffende deze onderwerpen doorvoeren die van invloed zijn op het werk van ontwerpers van intelligente technologie. In de toekomst kan de overheid mogelijk besluiten, eventueel in Europees verband, om nieuwe wet- en regelgeving in te voeren voor producten en diensten die kunstmatige intelligentie bevatten. Voor creatief technologen is het in de eerste plaats van belang om goed op de hoogte te zijn van dergelijke plannen. Daarnaast is het belangrijk dat overheden, bedrijven en kennisinstellingen die zich bezighouden met dit onderwerp, met elkaar in gesprek gaan.

5.2 Inhoud en aanpak

In het onderzoek naar verantwoord ontwerpen met kunstmatige intelligentie zal het lectoraat zich op de volgende drie onderwerpen richten: het ontwerpproces, het ontwerpresultaat en de ontwerpcontext.

Om het *ontwerpproces* te onderzoeken, zal ten eerste worden bestudeerd hoe ervaren ontwerpers te werk gaan wanneer ze ontwerpen met kunstmatige intelligentie als ontwerpmateriaal. Hoe pakken ontwerpers dat aan? Waar lopen ze tegenaan? In hoeverre denken ze na over de ethische gevolgen van hun ontwerpkeuzes? Vervolgens zullen er op basis van de bevindingen, aangevuld met literatuuronderzoek, methodes en tools ontwikkeld worden die ontwerpers ondersteunen om op verantwoorde wijze AI-toepassingen te ontwerpen. Een methode of tool is bijvoorbeeld een stappenplan, inspiratiekaarten, workshopformat of een website. Deze methodes en tools zullen geëvalueerd en verbeterd worden door ontwerpers ermee te laten werken. Het evalueren van dergelijke tools en methodes kan uitstekend plaatsvinden met studenten, want zij zijn degenen die er in de toekomst mee moeten werken. Daarnaast kan het voor studenten een leerzame ervaring zijn om te reflecteren op de methodes die ze gebruiken in een ontwerpproces.

Het onderzoek naar het *ontwerpresultaat* van verantwoord ontwerpen met kunstmatige intelligentie zal zich richten op het verzamelen en analyseren van toepassingen met kunstmatige intelligentie. Dit kan gaan om producten of diensten in de markt, maar ook om concepten die door studenten of docenten zijn ontworpen. Door meerdere toepassingen empirisch te evalueren en kritisch te analyseren, kunnen patronen ontdekt worden in welke oplossingen meer en minder goed werken. Hieruit kunnen richtlijnen en zogenaamde *design patterns* geabstraheerd worden die ontwerpers kunnen hergebruiken bij nieuwe ontwerpen. Deze richtlijnen en *design patterns* kunnen bijvoorbeeld gaan over hoe de uitkomst van een intelligent algoritme het best aan een gebruiker uitgelegd kan worden of over hoe een robot eruit moet zien om de juiste verwachtingen te scheppen bij een gebruiker.

Het onderzoek naar het ontwerpproces en ontwerpresultaat kan nog zulke mooie methodes, tools, richtlijnen en design patterns opleveren die verantwoord ontwerpen ondersteunen, ze zullen alleen impact hebben als ze daadwerkelijk worden toegepast in de praktijk. Het laatste onderwerp waar het lectoraat onderzoek naar doet is dan ook de *ontwerpcontext* van ontwerpers die werken met kunstmatige intelligentie. De meeste afgestudeerde studenten Creative Technologies van Hogeschool Rotterdam werken in dienst bij een bedrijf in de creatieve industrie in de regio Rotterdam (Voskuyl & Rutten, 2018). Door in gesprek te gaan met studenten en bedrijven wil het lectoraat onderzoeken wat er nu al gedaan wordt aan verantwoord ontwerpen en waar behoefte aan is.

De resultaten van de drie verschillende onderwerpen beïnvloeden elkaar en zijn niet van elkaar te isoleren. Door het ontwerpproces anders in te richten, is de kans groot dat er andere ontwerpresultaten uitkomen. In verschillende ontwerpcontexten zal dezelfde ontwerpmethode tot verschillende resultaten leiden. Daarnaast is elke ontwerpvraag uniek, wat het lastig maakt om resultaten van verschillende projecten goed met elkaar te vergelijken. Er zal daarom gewerkt worden volgens een *Research through Design*-aanpak, waarbij kennis wordt verworven door het uitvoeren van ontwerpactiviteiten en door deze activiteiten met de daaruit voortvloeiende artefacten te analyseren en erop te reflecteren (Stappers & Giaccardi, 2017). Bij Research through Design wordt dus op een exploratieve manier onderzoek gedaan door aan concrete casussen te werken. Het onderzoek van het lectoraat zal ontwerpactiviteiten bevatten die gericht zijn op het ontwerpen van toepassingen met kunstmatige intelligentie, maar ook op het ontwerpen van methoden en tools om ontwerpers te ondersteunen in het verantwoord ontwerpen met kunstmatige intelligentie.

Vanwege de breedheid en complexiteit van het onderwerp is het van belang om verschillende partijen bij het onderzoek te betrekken. Daarbij wordt gestreefd naar een combinatie van zowel verschillende rollen (student, docent, onderzoeker, ondernemer en ambtenaar) als expertisegebieden (informatica, ontwerp, communicatie en bedrijfskunde).

Het lectoraat beoogt met dit onderzoek diverse effecten te bereiken. In de eerste plaats is het de bedoeling dat het onderzoek kennis oplevert over verantwoord ontwerpen met kunstmatige intelligentie die van nut is voor ontwerpers en toegepast kan worden in de praktijk. De opgedane kennis wordt ook gepubliceerd in wetenschappelijke tijdschriften, zodat het met de academische gemeenschap gedeeld wordt. Verder kunnen kennis en ontwikkelde ontwerpmethodes gebruikt worden in het onderwijs om de curricula van de opleidingen aan het CMI-instituut up-to-date te houden, met name de Creative Technologies-opleidingen CMD en CMGT. De kennis kan ook van belang zijn voor overheden die nieuwe wet- en regelgeving omtrent kunstmatige intelligentie ontwikkelen.

Naast dat het onderzoek kennis oplevert, is het doen van onderzoek zelf op zich ook waardevol. Het meedoen aan onderzoeksprojecten is een leerzame ervaring voor studenten. Het uitvoeren van onderzoek traint de onderzoeksvaardigheden van docent-onderzoekers en helpt hun vakkennis bij te houden. Daarnaast helpt het onderzoek om meer bewustzijn te genereren over (het belang van) dit onderwerp en mensen aan het denken te zetten over hun omgang met kunstmatige intelligentie, zowel in het bedrijfsleven als in het onderwijs, bij de overheid en in de samenleving. Tot slot kan het onderzoek nieuwe (ideeën voor) AI-toepassingen opleveren die waardevol zijn voor de samenleving.

5.3 Conclusie

In het eerste hoofdstuk zijn twee mogelijke scenario's geschetst van een toekomst met kunstmatige intelligentie, waar bij de ene de positieve en de ander de negatieve gevolgen van kunstmatige intelligentie overheersen. Kunstmatige intelligentie is sterk in opkomst en de komende jaren zullen bepalend zijn voor de manier waarop zij wordt toegepast en tot wat voor samenleving dat zal leiden. Zoals betoogd in deze openbare les, spelen ontwerpers daarin een cruciale rol. Het uiteindelijke doel van het lectoraat is om handvatten te ontwikkelen voor ontwerpers, waarmee ze kunnen bijdragen aan een toekomst waarin kunstmatige intelligentie de samenleving ten goede komt.

Het is moeilijk en ambitieus om technologische ontwikkelingen een richting op te sturen en soms lijkt de ontwikkeling van 'de technologie' een onvermijdelijk proces te zijn. Maar ieder ontwerp is een nieuwe kans om een product of dienst van betekenis te creëren en 'technologie' is de collectieve uitkomst van alle ontwerpen. Het is niet mogelijk om de toepassing van kunstmatige intelligentie met een paar simpele ingrepen de goede kant op te sturen, maar iedere stap in de goede richting is er één en op de lange termijn kan een kleine stap dingen in beweging zetten en uiteindelijk een groot verschil maken. Dit zal niet vanzelf gaan. Naast kunstmatige intelligentie moet er menselijk verstand bij.

Literatuur

Anderson, J. R. (1996). *The Architecture of Cognition*. Psychology Press.

Andringa, R. (2018, 7 september). Politie wil zakkenrollers en plofkrakers vangen met data. *NOS*. Geraadpleegd op 24 oktober 2018: https://nos.nl/artikel/2250767-politie-wil-zakkenrollers-en-plofkrakers-vangen-met-data.html.

Angwin, J., & Larson, J. (2016, 30 december). Bias in Criminal Risk Scores Is Mathematically Inevitable Researchers Say. *ProPublica*. Geraadpleegd op 24 oktober 2018: https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.

Arkin, R. (2009). *Governing Lethal Behavior in Autonomous Robots*. Chapman and Hall/CRC.

Asilomar AI Principles. (2017). Principles developed in conjunction with the 2017 Asilomar conference (Benevolent AI 2017). Future of Life Institute. Geraadpleegd op 24 oktober 2018: https://futureoflife.org/ai-principles/.

Asimov, I. (1950). *I, Robot, Robot series*. Bantam Books.

Azenkot, S., Prasain, S., Borning, A., Fortuna, E., Ladner, R. E., & Wobbrock, J. O. (2011). Enhancing Independence and Safety for Blind and Deaf-blind Public Transit Riders. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3247–3256). New York, NY, USA: ACM.

Bardzell, S., Bardzell, J., Forlizzi, J., Zimmerman, J., & Antanitis, J. (2012). Critical Design and Critical Theory: The Challenge of Designing for Provocation. In *Proceedings of the Designing Interactive Systems Conference* (pp. 288–297). New York, NY, USA: ACM.

Best, J. (2013). IBM Watson: The inside story of how the Jeopardy-winning supercomputer was born, and what it wants to do next. *TechRepublic*. Geraadpleegd op 24 oktober 2018: https://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, *15*(3), 209–227.

Breazeal, C. L. (2004). *Designing Sociable Robots*. MIT Press.

Brooks, R. (2002). *Flesh and Machines: How Robots Will Change Us*. Knopf Doubleday Publishing Group.

Broussard, M. (2018). *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512.

Cavoukian, A. (2011). *Privacy by design in law, policy and practice* (A white paper for regulators, desicion-makers and policy-makers).

Clancey, W. J. (1983). The epistemology of a rule-based expert system —a framework for explanation. *Artificial Intelligence*, *20*(3), 215–251.

Clarkson, P. J., Coleman, R., Keates, S., & Lebbon, C. (2013). *Inclusive Design: Design for the Whole Population*. Springer Science & Business Media.

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral Decision Making Frameworks for Artificial Intelligence. *AAAI*, 4831–4835.

Cowls, J., & Floridi, L. (2018). *Prolegomena to a White Paper on an Ethical Framework for a Good AI Society* (SSRN Scholarly Paper No. ID 3198732). Rochester, NY: Social Science Research Network.

DeepMind. (2018). The story of AlphaGo so far. Geraadpleegd op 24 oktober 2018: https://deepmind.com/research/alphago/.

Deigh, J. (2010). *An Introduction to Ethics*. Cambridge University Press.

DigiVid360. (2018, 2 mei). 2018 YouTube Statistics. Geraadpleegd op 24 oktober 2018: http://www.digivid360.com/blog/video-strategy/2018-youtube-statistics/.

Docherty, B. (2012). Losing Humanity | The Case against Killer Robots. *Human Rights Watch*.

Domingo, M. C. (2012). An overview of the Internet of Things for people with disabilities. *Journal of Network and Computer Applications*, *35*(2), 584–596.

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Penguin Books Ltd.

Dorst, K. (2017). *Notes on Design: How Creative Practice Works*. BIS Publishers.

Dove, G., Halskov, K., Forlizzi, J., & Zimmerman, J. (2017). UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 chi conference on human factors in computing systems* (pp. 278–288). ACM.

European Commission. (2018a). Responsible Research and Innovation. Geraadpleegd op 24 oktober 2018: https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation.

European Commission. (2018b). 2018 reform of EU data protection rules. Geraadpleegd op 24 oktober 2018: https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en.

European Group on Ethics in Science and New Technologies. (2018). *Statement on Artificial Intelligence, Robotics and Autonomous Systems*. European Commission.

Farooq, M. U., Waseem, M., Mazgar, S., Khairi, A., & Kamal, T. (2015). A Review on Internet of Things (IoT). *International Journal of Computer Applications*, *113*(1), 1–7.

Fletcher-Watson, S., De Jaegher, H., van Dijk, J., Frauenberger, C., Magnee, M., & Ye, J. (2018). Diversity computing. *Interactions*, *25*(5), 28–33.

Freese, C., Dekker, R., Kool, L., & Van Est, R. (2018). *Robotisering en automatisering op de werkvloer - Bedrijfskeuzes bij technologische innovaties*. Den Haag: Rathenau Instituut.

Friedman, B. (1996). Value-sensitive Design. *Interactions*, *3*(6), 16–23.

Friedman, B., & Hendry, D. G. (2012). The Envisioning Cards: A Toolkit for Catalyzing Humanistic and Technical Imaginations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1145–1148). New York, NY, USA: ACM.

Friedman, B., Hendry, D. G., & Borning, A. (2017). A Survey of Value Sensitive Design Methods. *Foundations and Trends® in Human–Computer Interaction*, *11*(2), 63–125.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value Sensitive Design and Information Systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. E. Gorman (Red.), *Early engagement and new technologies: Opening up the laboratory* (pp. 55–95). Dordrecht: Springer Netherlands.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347.

Garcia, M. (2016). Racist in the Machine - The Disturbing Implications of Algorithmic Bias. *World Policy Journal*, *33*(4), 111–117.

Garrett, J. J. (2010). *The Elements of User Experience: User-Centered Design for the Web and Beyond*. Pearson Education.

Gates, B. (2007). A Robot in Every Home. *Scientific American*, *296*(1), 58–65.

Gibbs, S. (2015, 8 juli). Women less likely to be shown ads for high-paid jobs on Google, study shows. *The Guardian*. Geraadpleegd op 24 oktober 2018: https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study.

Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication* (pp. 55–60).

Granville, K. (2018, 19 maart). Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens. *The New York Times*. Geraadpleegd op 24 oktober 2018: https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.

Gurzawska, A., Mäkinen, M., Brey, P., Gurzawska, A., Mäkinen, M., & Brey, P. (2017). Implementation of Responsible Research and Innovation (RRI) Practices in Industry: Providing the Right Incentives. *Sustainability*, *9*(10), 1759.

Harbers, M. (2011). *Explaining agent behavior in virtual training*. Utrecht University.

Harbers, M., Bargh, M., Pool, R., Van Berkel, J., Van den Braak, S., & Choenni, S. (2018). A Conceptual Framework for Addressing IoT Threats: Challenges in Meeting Challenges. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Harbers, M., Detweiler, C., & Neerincx, M. A. (2015). Embedding Stakeholder Values in the Requirements Engineering Process. In S. A. Fricker & K. Schneider (Red.), *Requirements Engineering: Foundation for Software Quality* (pp. 318–332). Springer International Publishing.

Harbers, M., Mazerant, K., & Voskuyl, I. (2017). *Filterbubbels, nou en?* Rotterdam University of Applied Sciences.

Harbers, M., & Neerincx, M. A. (2017). Value sensitive design of a virtual assistant for workload harmonization in teams. *Cognition, Technology & Work*, 19(2-3), 329-343.

Harbers, M., Peeters, M. M., & Neerincx, M. A. (2017). Perceived autonomy of robots: Effects of appearance and context. In *A World with Robots* (pp. 19–33). Springer.

Hartson, R., & Pyla, P. S. (2012). *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. Elsevier.

Haugeland, J. (1989). *Artificial Intelligence: The Very Idea*. MIT Press.

Holmquist, L. E. (2017). Intelligence on Tap: Artificial Intelligence As a New Design Material. *Interactions*, *24*(4), 28–33.

IEEE Global Initiative on Ethics of Autnomous and Intelligent Systems. (2017). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (EADv2)*. IEEE.

Internet live stats. (2018). Google Search Statistics. Geraadpleegd op 24 oktober 2018: http://www.internetlivestats.com/google-search-statistics/.

Johnson, M., Bradshaw, J. M., & Feltovich, P. J. (2018). Tomorrow's Human–Machine Design Tools: From Levels of Automation to Interdependencies. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 77–82.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive Design: Designing Support for Interdependence in Joint Activity. *J. Hum.-Robot Interact.*, *3*(1), 43–69.

Kool, L., Dujso, E., & Van Est, R. (2018). *Doelgericht digitaliseren – Hoe Nederland werkt aan een digitale transitie waarin mensen en waarden centraal staan*. Den Haag: Rathenau Instituut.

Kravets, D. (2010, 25 januari). Jan. 25, 1979: Robot Kills Human. *Wired*. Geraadpleegd op 24 oktober 2018: https://www.wired.com/2010/01/0125robot-kills-worker/.

Kurzweil, R. (2010). *The Singularity is Near*. Gerald Duckworth & Co.

Liedtke, M. (2015, 2 juli). Google's new app blunders by calling black people 'gorillas.' *The Seattle Times.* Geraadpleegd op 24 oktober 2018: https://www.seattletimes.com/nation-world/google-apologizes-after-app-tagged-black-people-gorillas/.

Lin, P., Bekey, G., & Abney, K. (2009). *Robots In War: Issues Of Risk And Ethics*. NAVAL ACADEMY ANNAPOLIS MD, NAVAL ACADEMY ANNAPOLIS MD.

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Queue*, 16(3), 30.

Lovelock, J. D., Tan, S., Hare, J., Woodward, A., & Priestley, A. (2018). *Forecast: The Business Value of Artificial Intelligence, Worldwide, 2017-2025*. Gartner.

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, *90*, 46–60.

Malle, B. F. (2006). *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. MIT Press.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117–124). New York, NY, USA: ACM.

Marder, B., Joinson, A., Shankar, A., & Houghton, D. (2016). The extended 'chilling' effect of Facebook: The cold reality of ubiquitous social networking. *Computers in Human Behavior*, *60*, 582–592.

Martijn, M., & Tokmetzis, D. (2016). *Je hebt wél iets te verbergen*. de Correnspondent.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183.

Murphy, R., & Shields, J. (2012). *The role of autonomy in DoD systems*. Defense Science Board Task Force Report.

Nathan, L. P., Klasnja, P. V., & Friedman, B. (2007). Value Scenarios: A Technique for Envisioning Systemic Effects of New Technologies. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems* (pp. 2585–2590). New York, NY, USA: ACM.

NCSC. (2018). *Cyber Security Assessment Netherlands | CSAN 2018*. Den Haag: Ministerie van Justitie en Veiligheid.

Nest. (2018). Hoe bespaart de 3e generatie Nest Learning Thermostat energie? Geraadpleegd op 24 oktober 2018: https://nest.com/nl/support/article/eu-savings.

Nissenbaum, H. (2001). How computer systems embody values. *Computer*, *34*(3), 120–119.

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Noorman, M., & Johnson, D. G. (2014). Negotiating autonomy and responsibility in military robots. *Ethics and Information Technology*, *16*(1), 51–62.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown/Archetype.

Pariser, E. (2011). *The Filter Bubble: What The Internet Is Hiding From You*. Penguin UK.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

Perry, W. L. (2013). *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Rand Corporation.

Preece, J., Rogers, Y., & Sharp, H. (2015). *Interaction Design: Beyond Human-Computer Interaction*. John Wiley & Sons.

Raworth, K. (2017). *Doughnut Economics: Seven Ways to Think Like a 21st-Century Economist*. Chelsea Green Publishing.

Rollet, C. (2018, 5 juni). The odd reality of life under China's all-seeing credit score system. *Wired UK*. Geraadpleegd op 24 oktober 2018: https://www.wired.co.uk/article/china-social-credit.

Russell, S. J., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Rutten, P. W. M., & Schijvens, L. J. M. A. (2015). Ontwerpers onderzoeken de toekomst. *Stimuleringsfonds Creatieve Industrie. Jaarverslag 2014*, 144-149.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ArXiv:1708.08296*.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, *5*.

Schep, T. (2016). *Design my privacy*. BIS Publishers.

Schuler, D., & Namioka, A. (1993). *Participatory Design: Principles and Practices*. CRC Press.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424.

Sengers, P., Boehner, K., David, S., & Kaye, J. "Jofish." (2005). Reflective Design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility* (pp. 49–58). New York, NY, USA: ACM.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K., Flammini, A., & Menczer, F. (2017). The spread of low-credibility content by social bots. *ArXiv:1707.07592*. Geraadpleegd op 24 oktober 2018: http://arxiv.org/abs/1707.07592.

Sharkey, A., & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*, *14*(1), 27–40.

Shilton, K., & Anderson, S. (2017). Blended, Not Bossy: Ethics Roles, Responsibilities and Expertise in Design. *Interacting with Computers*, *29*(1), 71–79.

SIDN. (2018). Responsible AI. Geraadpleegd op 24 oktober 2018: https://www.sidnfonds.nl/responsible-ai/.

Silva, S. S. C., Silva, R. M. P., Pinto, R. C. G., & Salles, R. M. (2013). Botnets: A survey. *Computer Networks*, *57*(2), 378–403.

Singer, P. W. (2009). *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. Penguin.

Stappers, P., & Giaccardi, E. (2017). Research through Design. In *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* Interaction Design Foundation.

Stefik, M. (2014). *Introduction to Knowledge Systems*. Elsevier.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751–752.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. Random House.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Thompson, D. F. (1980). Moral Responsibility of Public Officials: The Problem of Many Hands. *American Political Science Review*, *74*(4), 905–916.

Turing, A. M. (2009). Computing Machinery and Intelligence. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer* (pp. 23–65). Dordrecht: Springer Netherlands.

Van Berkel, J., Pool, R., Harbers, M., Oerlemans, J. J., Bargh, M., & Van den Braak, S. (2018). *(Verkeerd) verbonden in een slimme samenleving: Het Internet of Things: kansen, bedreigingen en maatregelen* (No. 2734). Den Haag: Ministerie van Justitie en Veiligheid.

Van de Weijer, B. (2018, 14 maart). Deze auto rijdt zelfstandig. Wat moeten we ons daar nou precies bij voorstellen? Volkskrant. Geraadpleegd op 24 oktober 2018: https://www.volkskrant.nl/wetenschap/deze-auto-rijdt-zelfstandig-wat-moeten-we-ons-daar-nou-precies-bij-voorstellen-~b9650b48/.

Van den Hoven, J. (2007). ICT and Value Sensitive Design. In P. Goujon, S. Lavelle, P. Duquenoy, K. Kimppa, & V. Laurent (Red.), *The Information Society: Innovation, Legitimacy, Ethics and Democracy In honor of Professor Jacques Berleur SJ* (pp. 67–72). Boston, MA: Springer.

Van Lent, M., Fisher, W., & Mancuso, M. (2004). An Explainable Artificial Intelligence System for Small-unit Tactical Behavior. In *Proceedings of the 16th Conference on Innovative Applications of Artifical Intelligence* (pp. 900–907). San Jose, California: AAAI Press.

Van Noort, W. (2018, 14 augustus). Googles AI verslaat oogartsen in het stellen van diagnoses. *NRC*. Geraadpleegd op 24 oktober 2018: https://www.nrc.nl/nieuws/2018/08/14/googles-ai-verslaat-oogartsen-in-stellen-diagnose-a1613048.

Van Wynsberghe, A. (2013). Designing Robots for Care: Care Centered Value-Sensitive Design. *Science and Engineering Ethics*, *19*(2), 407–433.

Verbeek, P.-P. (2006). Materializing Morality: Design Ethics and Technological Mediation. *Science, Technology, & Human Values*, *31*(3), 361–380.

Vergunst, N., & Mols, B. (2017). *Hallo robot: De machine als medemens*. Nieuw Amsterdam.

Voskuyl, I., & Rutten, P. (2018). *Arbeidspraktijk Creatief Technologen alumnimonitor 2017*. Rotterdam University of Applied Sciences.

Walker, S. (2012). *Sustainable by Design : Explorations in Theory and Practice*. Routledge.

Wallach, W. (2015). *A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control*. New York: Basic Books.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Woelfer, J. P., Iverson, A., Hendry, D. G., Friedman, B., & Gill, B. T. (2011). Improving the Safety of Homeless Young People with Mobile Phones: Values, Form and Function. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1707–1716). New York, NY, USA: ACM.

Yadron, D., & Tynan, D. (2016, June 30). Tesla driver dies in first fatal crash while using autopilot mode. *The Guardian*. Geraadpleegd op 24 oktober 2018: https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk.

Yang, Q., Scuito, A., Zimmerman, J., Forlizzi, J., & Steinfeld, A. (2018). Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference* (pp. 585–596). New York, NY, USA: ACM.

Yang, Q., Zimmerman, J., Steinfeld, A., & Tomasic, A. (2016). Planning Adaptive Mobile Experiences When Wireframing. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (pp. 565–576). New York, NY, USA: ACM.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., & Zorzi, M. (2014). Internet of Things for Smart Cities. *IEEE Internet of Things Journal*, *1*(1), 22–32.

Zephoria. (2018, 19 september). Top 20 Facebook Statistics - Updated September 2018. Geraadplaagd op 24 oktober 2018: https://zephoria.com/top-15-valuable-facebook-statistics/.

Eerdere uitgaven

van Hogeschool Rotterdam Uitgeverij

Inaugural lecture

Auteur	Prof Dr Ben van Lier CMC
ISBN	9789493012028
Verschijningsdatum	oktober 2018
Aantal pagina's	78
Prijs	€ 14,95

Measurement Numeracy Education for Prospective Elementary School Teachers

Auteur	Mark van Houwelingen
ISBN	9789493012004
Verschijningsdatum	oktober 2018
Aantal pagina's	144
Prijs	€ 19,95

Schillen van het verschil

Auteur	Tina Rahimi
ISBN	9789493012011
Verschijningsdatum	juni 2018
Aantal pagina's	86
Prijs	€ 14,95

Zorg voor Communicatie

Auteur	Karin Neijenhuis
ISBN	9789051799859
Verschijningsdatum	maart 2018
Aantal pagina's	116
Prijs	€ 14,95

Next Strategy

Auteur	Arjen van Klink
ISBN	9789051799712
Verschijningsdatum	november 2017
Aantal pagina's	102
Prijs	€ 14,95

Visie op de toekomst van de Nederlandse procesindustrie

Auteur	Marit van Lieshout
ISBN	9789051799682
Verschijningsdatum	oktober 2017
Aantal pagina's	68
Prijs	€ 14,95

Slim bewegen tussen haven en stad

Auteur	Ron van Duin
ISBN	9789051799675
Verschijningsdatum	oktober 2017
Aantal pagina's	84
Prijs	€ 14,95

Techniek is belangrijk, maar het zijn mensen die het verschil maken

Auteur	Hans van den Broek
ISBN	9789051799644
Verschijningsdatum	oktober 2017
Aantal pagina's	84
Prijs	€ 14,95

#DuurzaamRenoveren

Auteur	Haico van Nunen
ISBN	9789051799651
Verschijningsdatum	oktober 2017
Aantal pagina's	100
Prijs	€ 14,95

Bewegen naar gezondheid

Auteur	Maarten Schmitt
ISBN	9789051799632
Verschijningsdatum	september 2017
Aantal pagina's	86
Prijs	€ 14,95

Studiesucces

Auteur	Ellen Klatter
ISBN	9789051799583
Verschijningsdatum	juni 2017
Aantal pagina's	96
Prijs	€ 14,95

Exemplaren zijn te downloaden via www.hr.nl/onderzoek/publicaties. Hier zijn ook eerder verschenen uitgaven van Hogeschool Rotterdam Uitgeverij beschikbaar.

Maaike Harbers

Verstand erbij

Verantwoord ontwerp van toepassingen met kunstmatige intelligentie

9 789493 012035

Zoekmachines helpen ons dagelijks met zoeken naar relevante informatie, spamfilters houden ongewenste email buiten zicht, we krijgen gepersonaliseerde aanbevelingen voor nieuwsartikelen, filmpjes, muziek en series, de thermostaat leert wanneer de verwarming aan moet en stofzuigerrobots houden onze huizen schoon.

De toepassing van kunstmatige intelligentie is sterk gegroeid in de afgelopen jaren. Dit heeft zowel wenselijke als minder wenselijke maatschappelijke gevolgen. Robots en algoritmes kunnen bijvoorbeeld gezondheidszorg verbeteren en steden verduurzamen, maar ze kunnen ook ongelukken veroorzaken, etnisch profileren en discrimineren. In deze openbare les zal Maaike Harbers ingaan op de rol die ontwerpers van toepassingen met kunstmatige intelligentie hierin spelen. Ontwerpers beïnvloeden, met hun ontwerpkeuzes, wat de gevolgen zijn van die toepassingen. Door verantwoorde keuzes te maken tijdens het ontwerpproces, kunnen ontwerpers bijdragen aan een inzet van kunstmatige intelligentie die de samenleving ten goede komt.

Maaike Harbers is lector Artificial Intelligence & Society bij Kenniscentrum Creating 010 van Hogeschool Rotterdam en hoofddocent bij de opleiding Creative Media & Game Technologies van Hogeschool Rotterdam.

OPENBARE LES

praktijkgericht **onderzoek**

Hogeschool Rotterdam Uitgeverij

HOGESCHOOL ROTTERDAM