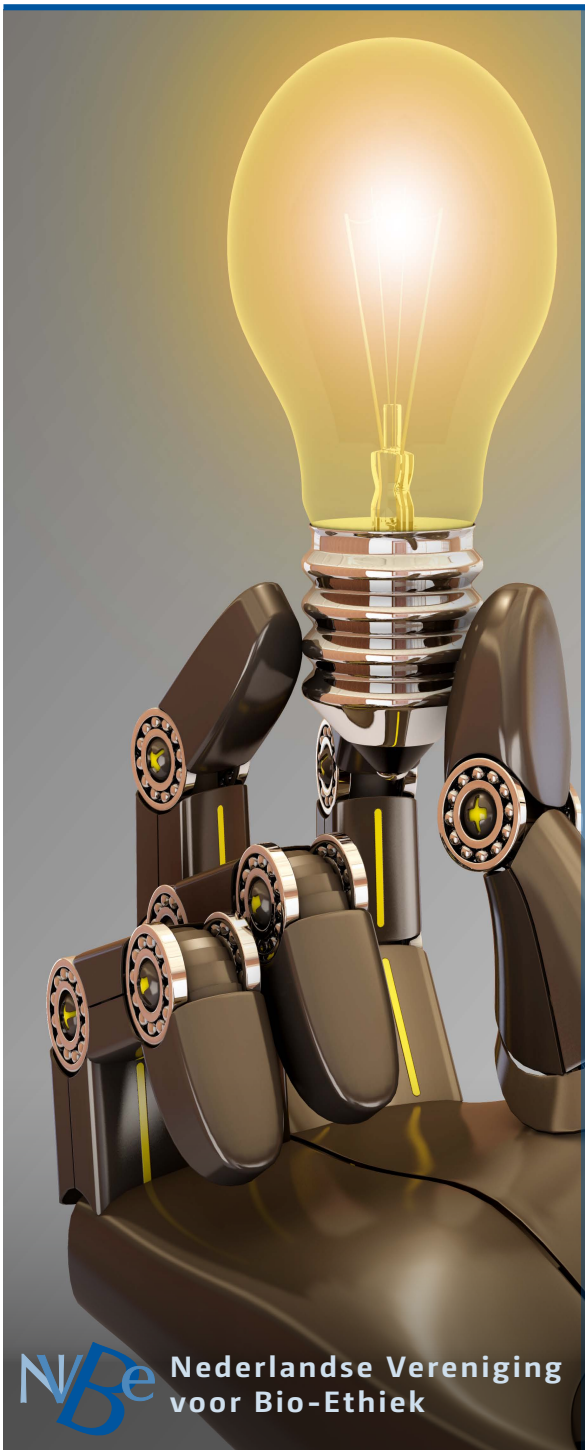
Preadvies

uitgebracht t.b.v. het jaarsymposium van de
Nederlandse Vereniging voor Bio-ethiek 11 juni 2021



Siri, wat adviseer jij?

Over het gebruik van kunstmatige intelligentie voor morele oordeelsvorming

Katleen Gabriels

Nederlandse Vereniging voor Bio-Ethiek

uitgebracht t.b.v. het jaarsymposium van de
Nederlandse Vereniging voor Bio-ethiek 11 juni 2021

Siri, wat adviseer jij?

Over het gebruik van kunstmatige intelligentie voor morele oordeelsvorming

Katleen Gabriels

Nederlandse Vereniging
voor Bio-Ethiek

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Katleen Gabriels

Siri, wat adviseer jij?
Over het gebruik van kunstmatige intelligentie voor morele oordeelsvorming

Preadvies uitgebracht t.b.v. het jaarsymposium van de Nederlandse Vereniging voor Bio-ethiek op 11 juni 2021 door Katleen Gabriels – Online: NVBe
ISBN: 9789080259409
Trefwoorden: kunstmatige intelligentie, morele oordeelsvorming, artificial moral agent, artificial moral advisor, casuïstiek
Omslagontwerp en lay out binnenwerk: Ger Palmboom
Omslagfoto: Shutterstock (afbeeldings-ID 567406477)

Copyright 2021, Katleen Gabriels – Valkenburg aan de Geul

Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand of openbaar gemaakt in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen of enig andere manier, zonder voorafgaande toestemming van de auteur.

Inhoudsopgave

3 Inleiding
9 Theoretisch kader
25 Vier casussen
39 Hoe kunnen we een kunstmatige morele actor programmeren?
49 Advies
61 Besluit
63 Referenties
67 Dankwoord
69 Over de auteur

Inleiding

Wat is de mens? Deze vraag was één van de leidvragen in het werk van Verlichtingsfilosoof Immanuel Kant. Filosofen Alberto Giubilini en Julian Savulescu (2018) omschrijven de mens als een *ondermaatse informatieverwerker*. Door stress, tijdgebrek en beperkte mentale mogelijkheden nemen we niet altijd alle relevante informatie op die nodig is om tot een onderbouwd moreel oordeel te komen. Om dit tekort op te vangen, opperen Giubilini en Savulescu dat we ons kunnen laten bijstaan door kunstmatige intelligentie (AI). In tegenstelling tot de mens kan AI enorm veel data verzamelen, analyseren en onthouden, wat volgens hen tot beter geïnformeerde en efficiëntere morele oordelen leidt. In Kant zijn tijd was er van AI natuurlijk nog geen sprake, maar zelf schreef hij in 1784: 'Uit het kromme hout waaruit de mens is opgebouwd is nog nooit iets rechts getimmerd'. Enige hulp om wat rechtlijniger te worden, kan dus welkom zijn.

De AI-toepassing waar Giubilini en Savulescu voor pleiten, namelijk, *artificial moral advisors* of kunstmatige morele adviseurs, is vandaag de dag nog niet beschikbaar. Spraakassistenten als Alexa en Siri kunnen wel helpen met het weerbericht, maar niet met moreel advies. Toen ik Siri onlangs voor de grap om morele bijstand vroeg, kwam Siri niet verder dan 'dat is een interessante vraag' en 'hmm... daar heb ik geen antwoord op. Kan ik je ergens anders mee helpen?'.

AI helpt ons al wel in andere domeinen. Chatbot Woebot is een Engelstalige AI-'therapeut' die, op basis van inzichten uit de cognitieve gedragstherapie, ontwikkeld werd door klinisch psychologen van de universiteit van Stanford (Fitzpatrick, Darcy, & Vierhile, 2017).[1] In tegenstelling tot een menselijke thera-

1 https://woebothealth.com/ Geraadpleegd op 18 april 2021.

peut is Woebot altijd beschikbaar om te chatten; de app kan gedownload worden via Google Play of de Apple App Store. Woebot 'luistert' en stelt vragen, maar kan een therapeut (nog?) niet vervangen. Gebruikers die met Woebot in gesprek gaan, kunnen uit standaard antwoordopties kiezen, en er is een mogelijkheid tot een open chatgesprek.

Er zijn ook beloftevolle doorbraken bij medische oordeelsvorming. Onderzoekers trainden algoritmen[2] bijvoorbeeld met 130.000 foto's van huidkanker (Esteva et al., 2017). Wat blijkt? De algoritmen zijn even goed als ervaren dermatologen in het voorspellen van huidkanker. Er zijn vergelijkbare doorbraken bij borst- en slokdarmkanker. Een aantal ziekenhuizen maakt al gebruik van AI-technieken bij het voorspellen van ziektes. AZ Klina in het Belgische Brasschaat gebruikt bijvoorbeeld AI om verdacht borstweefsel op te sporen. AI helpt ook mee in de strijd tegen corona. Algoritmen scannen longfoto's om te zien of patiënten besmet zijn met COVID-19. Het Nederlandse CAD4COVID (*Computer Aided Detection for COVID-19*) werd internationaal gratis ter beschikking gesteld, als hulpmiddel bij het detecteren van het virus op röntgenfoto's (Gabriels, 2021). Stel dat de inzet van AI steevast tot betere diagnoses leidt, kan het dan op termijn onethisch worden om er *geen* gebruik van te maken?

Algoritmen worden ook ingezet voor de preventie van zelfdoding. Een Belgische studie naar patronen van suïcidale gedachten en gedrag van eerstejaarsstudenten aan de KU Leuven, maakte gebruik van algoritmen om risico's op zelfdoding in kaart te brengen (Mortier et al., 2017). Eerstejaarsstudenten vulden vragenlijsten in en een aantal van hen bleef betrokken bij vervolgstudies. Volgens de onderzoekers is de technologie een beloftevolle manier om grote risico's tijdig op te sporen tijdens de studie aan de universiteit. Algoritmen scannen ook elektronische patiëntendossiers om patiënten met een groot risico op zelfdoding te voorspellen (Van Mens et al., 2020). In tegenstelling tot artsen kunnen algoritmen heel snel elektronische patiëntendossiers doorzoeken, wat nuttig kan zijn in situaties waar sprake is van grote tijdsdruk (Biller-Andorno & Biller, 2019).

Centrale vraagstelling

AI presteert in steeds meer domeinen beter dan de mens. Waarom zou ethiek, en in het bijzonder morele oordeelsvorming, een uitzondering zijn (zie o.a. Biller-

2 Centrale begrippen zoals 'algoritmen' worden toegelicht in hoofdstuk 1.

Andorno & Biller, 2019)? Mensen zijn feilbaar en beïnvloedbaar: als we ons gevleid voelen of gestrest zijn, dan kan dit een nadelig effect hebben op de kwaliteit van onze oordelen en reflecties. Bovendien zeggen we vaak dat we iets belangrijk vinden, zoals duurzaamheid, maar passen we ons gedrag er vervolgens niet op aan. Volgens Giubilini & Savulescu (2018) slagen we er vaak niet in om alle informatie in ogenschouw te nemen die relevant is voor een rationele of morele beslissing, of om vast te houden aan belangrijke principes die we onderschrijven. En zelfs als de eerste twee problemen niet spelen, kan bijvoorbeeld wilszwakte ons ervan weerhouden om datgene te doen waarvan we overtuigd zijn dat het moreel gezien het juiste is om te doen. We zijn volgens de auteurs daarom naast slechte verwerkers van informatie, ook ondermaatse morele beoordelaars (*moral judges*) en ondermaatse morele actoren. AI zou deze problemen volgens de auteurs kunnen verhelpen.

Deze gedachtenlijn vormt de aanleiding voor dit preadvies. De centrale vraag is: **in hoeverre kan AI ingezet worden bij morele oordeelsvorming?** Het vormen en onderbouwen van morele oordelen spelen een belangrijke rol in het dagelijks leven – we vellen ze aan de lopende band – en in het werk van professionals, waaronder bio-ethici. In hoeverre is het realistisch dat AI hen in de toekomst bijstaat bij morele oordeelsvorming? Als AI sneller en consistenter tot een moreel oordeel kan komen, wordt het dan op termijn onethisch om géén AI in bijvoorbeeld een medisch-ethische toetsingscommissie plaats te laten nemen?

In dit preadvies worden met behulp van casuïstiek vier typen zogeheten *Artificial Moral Agents* (AMA's) onderzocht. De typen AMA worden op een continuüm geplaatst, met aan het ene uiterste AI die data aanlevert, zodat mensen tot een beter moreel oordeel kunnen komen, en aan het andere uiterste AI die zélf een moreel oordeel vormt en moreel kan handelen. De nadruk ligt op het gebruik van AI voor morele oordeelsvorming. Bij de bespreking van de verschillende typen AMA zal ik steeds de brug slaan naar hun (mogelijke) bijdragen op het vlak van moreel advies. Daarbij ga ik ook in op een aantal sub-vragen:

- Welke 'onderdelen' van moraliteit moet AI leren, om te kunnen spreken van een kunstmatige morele actor en adviseur?
- Kunnen we moraliteit aan AI aanleren en zo ja, hoe gaat dat in zijn werk?
- Welke technische, conceptuele en ethische uitdagingen brengt het gebruik van AI voor morele oordeelsvorming met zich mee?

Opbouw van het preadvies

Hoofdstuk 1 start met een korte uitleg van belangrijke begrippen, waaronder de vier typen Artificial Moral Agents (AMA's). Ook komt een aantal valkuilen aan bod. Zo leidt het begrip 'agent' bijvoorbeeld vaak tot verwarring, omdat er bij een AMA van vrij handelingsvermogen (nog) geen sprake is. Bovendien, voordat we kunnen bepalen welke rol AI kan spelen bij morele oordeelsvorming, hebben we eerst meer inzicht nodig in hoe *de mens* tot morele oordelen komt, en wat daarvoor precies van belang is. Daarover bestaat echter geen eensgezindheid.

In hoofdstuk 2 bespreek ik vier uiteenlopende cases. De cases zijn primair bedoeld als voorbeelden om het onderscheid tussen de vier typen AMA's te verhelderen, en om te illustreren welke vragen en uitdagingen toenemende complexiteit en morele 'autonomie' van een AMA meebrengen. In elke volgende casus heeft de AMA een grotere rol bij morele oordeelsvorming. De eerste casus gaat over euthanasie. Ik bespreek of een AMA op basis van data-analyses de wil van een patiënt zou kunnen voorspellen als een wilsverklaring ontbreekt of deze voor interpretatie vatbaar is. En zou AI een rol kunnen spelen bij het bepalen of er sprake is van uitzichtloos en ondraaglijk lijden? De tweede casus gaat over de inzet van AMA's bij de coördinatie en distributie van goederen en hulp bij een ramp. De AI houdt daarbij rekening met een aantal verdelingscriteria, zoals dat van 'hoogste nood', maar heeft verder geen eigen inbreng. Casus drie betreft een AI-gesprekspartner die ons kan bijstaan bij het beslechten van morele dilemma's en die bovendien zwakke argumenten kan ontdekken en bijstellen. Tot slot (casus vier) schets ik een gedachte-experiment over een robot met (zelf)bewustzijn en bespreek ik of deze geschikt kan zijn als opvoeder van kinderen.

In het derde hoofdstuk bespreek ik, aan de hand van academische literatuur, hoe moraliteit in een machine geprogrammeerd kan worden. Drie manieren komen aan bod: top-down, bottom-up en hybride. Het aanleren van moraliteit met behulp van een bestaande ethische theorie is een top-down manier. Een AMA kan ook bottom-up leren, bijvoorbeeld uit interacties met de omgeving. Een hybridebenadering combineert de eerste twee. In dit hoofdstuk wordt duidelijk dat de uitdagingen bij het programmeren van moraliteit niet alleen technisch, maar ook conceptueel zijn, want hoe definieer je bijvoorbeeld 'goed' of 'kwaad'?

Op basis van inzichten uit hoofdstuk 1 tot en met 3 staat in het voorlaatste hoofdstuk mijn advies centraal. Eerst komt de vraag aan bod met welke data en ethische theorieën we de verschillende types AMA moeten trainen. In het

tweede deel bespreek ik of meer snelheid, consistentie en efficiëntie bij morele oordeelsvorming in elke context wenselijk is. Vervolgens ga ik in op de vraag wat het betekent als een AMA op het gebied van morele oordeelsvorming 'beter' presteert dan de mens. In het vierde deel staat verantwoordelijkheid centraal. Tot slot, in het vijfde deel, bespreek ik het belang van duidelijk omschreven begrippen, om verwarring te vermijden, zoals bij de term 'agent'.

In het vijfde en laatste hoofdstuk volgt het besluit. Om alvast een tipje van de sluier op te lichten: gezien de beperkte mogelijkheden van AI om tot morele oordeelsvorming te komen, moet de eindverantwoordelijkheid naar mijn idee bij de mens blijven liggen.

Dit preadvies heeft niet de ambitie om uitputtend te zijn. Over veel onderwerpen en begrippen die aan bod komen, bestaat geen eensgezindheid. Bovendien zijn de vragen groots en complex en leiden pogingen om ze te beantwoorden direct weer tot nieuwe vragen. Volledigheid is onmogelijk. Ook zijn er geen 'pasklare' antwoorden. Het doel van dit preadvies is vooral om discussie te stimuleren over het (eventuele) gebruik van AI bij morele oordeelsvorming, door in kaart te brengen wat er is, wat er kan zijn en wat *sciencefiction* zal (moeten) blijven.

1 Theoretisch kader

1.1 Kunstmatige intelligentie: de belangrijkste begrippen op een rij

Kunstmatige intelligentie (AI) begon in de jaren 1950 als onderzoeksveld om menselijke intelligentie in een machine na te bootsen. De opzet was om machines op een vergelijkbaar niveau als de mens taken te laten uitvoeren, zoals redeneren. Dat kan door er top-down regels in te programmeren (regelgestuurde AI) of door de machine bottom-up te laten leren van data (datagestuurde AI).[3] Vandaag de dag is AI vooral een verzamelnaam voor datagestuurde AI, namelijk "technieken zoals machine learning, deep learning en neurale netwerken die tot doel hebben om computers te laten leren uit hun ervaringen" (Jacobs et al., 2019, p. 14). Hieronder licht ik de belangrijkste begrippen kort toe.

Machine learning is een techniek om AI te ontwikkelen. In je mailbox 'leren' algoritmen bijvoorbeeld gaandeweg wat spamberichten zijn en vervolgens plaatsen ze die zelf in de spamfolder. Een *algoritme* bestaat uit een aantal opeenvolgende beslissingsmomenten, waarin informatie (input) wordt verwerkt tot een resultaat (output). Wanneer ik 's morgens voor de kledingkast sta, voer ik ook een soort van algoritme uit: als input verwerk ik een heleboel informatie, zoals de weersvoorspelling voor die dag, mijn planning, eerdere ervaringen (zijn deze schoenen comfortabel?), enzovoort. Op basis daarvan neem ik een aantal beslissingen die leiden tot output van het algoritme: de keuze voor de specifieke outfit die ik vandaag zal dragen (Gabriels, 2019). Machine learning-algoritmen worden getraind met grote hoeveelheden data en leren daaruit (*self-learning*), zonder dat ze expliciete regels hoeven te volgen. Algoritmen kunnen veel databronnen tegelijk verwerken en er

3 In de literatuur zijn ook 'datagebaseerde' en 'regelgebaseerde' AI veelgebruikte termen. Hoe moraliteit in AI kan worden geprogrammeerd, komt uitgebreider aan bod in hoofdstuk 3.

patronen in zoeken. Je kan een AI-systeem bijvoorbeeld leren hoe een kat eruitziet, door het te trainen met een grote hoeveelheid kattenfoto's. Op den duur kan het systeem zelf een kat *herkennen* en *labelen*.

Deep learning is een onderdeel van machine learning en verwijst naar kunstmatige neurale netwerken (ANN, met de A van *artificial*), die vergelijkbaar zijn met de neurale structuren in onze hersenen. ANN zoeken patronen en complexe verbanden op diepliggende niveaus in datasets, om zo voorspellingen te kunnen doen. De algoritmen in het eerdergenoemde onderzoek over het diagnosticeren van huidkanker leerden aan de hand van deep learning-technieken. Deep learning bestaat al sinds de jaren 1980, maar het domein is recent in een stroomversnelling geraakt, door de enorme hoeveelheid aan beschikbare data.

De ontwikkeling van AI kende sinds de jaren 1950 pieken en dalen. De dalen worden 'AI-winters' genoemd, met verminderde interesse en afgenomen enthousiasme. Momenteel is sprake van een 'AI-zomer', dankzij nieuwe technologische ontwikkelingen en applicaties; toegenomen opslagcapaciteit, snelheid en rekenkracht; en big dataprojecten die de immense datasets aanleggen waarmee algoritmen getraind worden. Veelgebruikte technologieën zoals de zoekmachine van Google, de tijdslijn van Facebook of spraakassistenten Siri of Alexa, maken gebruik van AI. In de eerste industriële robot in de jaren 1960 zat nog geen AI, maar inmiddels zijn robots de *belichaming* van AI (Lin, Abney, & Bekey, 2014, p. 4). Big data en het *Internet of Things* (IoT) dragen bij aan AI-mogelijkheden (cf. datagestuurde AI); het zijn samenhangende evoluties. IoT is een verzamelnaam voor allerhande toestellen, naast laptops, tablets of smartphones, die aan het netwerk gekoppeld worden. Zo ontstaat een zelforganiserend netwerk van met elkaar verbonden en uniek identificeerbare objecten. Een goed voorbeeld zijn sensoren in de thuiszorg (CEG, 2020). Die sensoren worden thuis bij de patiënt bevestigd op bijvoorbeeld de slaapkamerdeur en het koffiezetapparaat en registreren activiteit en beweging. De algoritmen zoeken patronen in het dagelijks functioneren van de patiënt. Bij een afwijking van het patroon, zoals de slaapkamerdeur die dicht blijft of de koffie die niet rond een vast tijdstip wordt gemaakt, worden zorgverleners automatisch gealarmeerd.

Inmiddels bestrijkt AI een breed en interdisciplinair veld dat bevolkt wordt door onder andere computerwetenschappers, wiskundigen, linguïsten, neurowetenschappers en filosofen. Die laatsten buigen zich over vragen zoals 'wat is intelligentie?' en 'kan een machine op dezelfde manier intelligent handelen als de mens?'.

1.2 Sterke vs. zwakke AI – tegenstelling of continuüm?

Filosofen maken vaak het onderscheid tussen sterke en zwakke AI. Bij zwakke AI gaat het om het uitvoeren van specifieke taken, waarbij de mens een beperkte doelstelling oplegt aan de techniek. In dat geval handelen machines *alsof* ze intelligent zijn (simulatie). Siri, Alexa en Woebot, bijvoorbeeld, werken binnen een vooraf gedefinieerde set van functies. Het zijn voorbeelden van zwakke AI, en deze toepassingen kunnen niet zelf moreel advies geven.

Sterke AI zou wél grondig moreel advies kunnen geven, maar het is hoogst onzeker of sterke AI ooit mogelijk zal zijn. De meningen zijn althans sterk verdeeld.[4] Sterke AI betekent dat AI een vorm van zelfbewustzijn heeft en intelligent *is*. Dit spreekt natuurlijk het meest tot de verbeelding, dus veel sciencefictionverhalen gaan hierover. Bijvoorbeeld, in de roman *Machines Like Me* (2019, vertaald als *Machines zoals ik*) van de Britse schrijver Ian McEwan wonen robots bij mensen in huis. Alan Turing, de geestelijke vader van de computer, heeft een zeer intelligente robot ontwikkeld: een synthetische mens. Er worden slechts vijfentwintig exemplaren op de markt gebracht. Charlie, de hoofdpersoon, koopt er één. De robot van Charlie is erg gevoelig en schrijft zelfs haiku's. Maar met zijn onfeilbare geheugen, ingebouwde camera- en geluidsregistratie en consistent moreel kompas brengt de robot Charlie en zijn partner in de problemen. Voor de vijfentwintig robots die ontworpen zijn om consistent en efficiënt te handelen, is het niet eenvoudig om met mensen samen te leven. We zeggen vaak dat we iets gaan doen, maar handelen er vervolgens niet naar. Of we vertellen een leugentje om 'bestwil'. De menselijke feilbaarheid en complexiteit drijft sommige robots tot drastische maatregelen. *Machines Like Me* geeft een interessante inkijk in hoe het dagelijkse leven met een intelligente en moreel consistente robot zou kunnen zijn, maar het blijft bij een gedachte-experiment.

Voorbij de tweedeling zwakke vs. sterke AI

Een aantal prominente AI-onderzoekers, waaronder natuurkundige Max Tegmark, hanteert de tegenstelling tussen zwakke en sterke AI liever niet. De tweedeling onderschat volgens Tegmark waar AI-systemen nu reeds toe in staat zijn, omdat

4 Voor voorstanders, zie bijvoorbeeld het werk van Ray Kurzweil (over Singularity) en Eliezer Yudkowsky. Voor tegenstanders, zie bijvoorbeeld het werk van Joseph Weizenbaum en Hubert Dreyfus. Zie bijvoorbeeld ook Fjelland, R. (2020). Why General Artificial Intelligence Will Not Be Realized. *Humanities and Social Sciences Communications* 7(10). https://www.nature.com/articles/s41599-020-0494-4

we meer uit AI halen dan de vooraf gedefinieerde functies. Het bekendste voorbeeld is waarschijnlijk *AlphaGo*, een AI die in maart 2016 een menselijke tegenspeler, wereldkampioen Lee Sedol, versloeg bij het complexe Chinese bordspel Go. AlphaGo werd ontwikkeld door DeepMind, onderdeel van Alphabet Inc., het moederbedrijf van Google en valt in principe onder zwakke AI. De eerste stappen van AlphaGo bestonden uit het trainen van diepe kunstmatige neurale netwerken (deep learning) gebaseerd op miljoenen zetten van Go-spelers. De algoritmen van AlphaGo speelden het spel herhaaldelijk tegen een kopie van zichzelf (*selfplay*) en leerden daaruit, op een bottom-up manier. Door miljoenen wedstrijden te spelen, ontdekte AlphaGo zelf strategieën en werd het almaar beter. AlphaGo toonde verrassende inzichten en strategieën, die als een vorm van creativiteit beschouwd kunnen worden. Algoritmen leren uit interactie met de data en kunnen zich aan die nieuwe kennis aanpassen én aan veranderende contexten. Zo passen de algoritmen van Spotify zich bijvoorbeeld aan de veranderende muzieksmaak van gebruikers aan (du Sautoy, 2020, p. 83). Tijdens het spel met Sedol deed AlphaGo een zet waarvan kenners in eerste instantie dachten dat deze een vergissing moest zijn. Maar het bleek een briljante zet te zijn die getuigde van veel inzicht. Lange tijd werd gedacht dat je eenvoudigweg uit algoritmen haalt wat je erin stopt, maar AlphaGo verraste en zorgde ervoor dat mensen Go anders gingen spelen door de nieuwe zetten uit te proberen (du Sautoy, 2020, p. 283). Algoritmen kunnen dus méér geven dan je erin stopt.

Inzicht betekent in het geval van AlphaGo dat er een nieuwe benadering is, een nieuwe manier van spelen, waar mensen niet eerder aan dachten. Dit roept de vraag op: zou AI ons ook op een nieuwe manier naar bio-ethische kwesties kunnen laten kijken of, algemener, naar moraliteit? Laten we bij het begin beginnen: waar moet de technologie eigenlijk aan voldoen wil er sprake zijn van een *Artificial Moral Agent* (AMA), of – in het Nederlands – van een kunstmatige morele actor?[5]

1.3 Vier typen kunstmatige morele actoren

In het denken over moraliteit en AI wordt onderscheid gemaakt tussen vier typen kunstmatige morele actoren. De AMA's verschillen onderling wat betreft hoe zelfstandig ze zijn, en qua complexiteit en morele gevoeligheid (zie Moor 2006; 2007).

5 Ook is inzicht nodig in wat moraliteit is om die vervolgens in AI te kunnen inbouwen. Mogelijkheden om moraliteit en specifieke morele perspectieven in te bouwen in AI komen aan bod in hoofdstuk 3.

1.3.1 De ethische impact actor

Het eerste type is de 'ethische impact actor' (AMA-1). Daaronder vallen alle AI-systemen en robots die ethische impact hebben. In principe valt dus elke robot en elk AI-systeem (tenminste) in deze categorie, omdat de gevolgen ervan zowel voor- als nadelig kunnen zijn.

Toegespitst op morele oordeelsvorming en moreel advies (cf. Giubilini & Savulescu, 2018) vallen hier alle AI-systemen onder die data aanleveren om tot een beter onderbouwd moreel oordeel te komen.[6] In dit geval wordt dat oordeel nog door de mens gevormd; de AI levert enkel de data (en de analyses) die het oordeel (mee) onderbouwen.[7] In hoofdstuk 2 bespreek ik een casus in het kader van euthanasie om dit type AMA concreter te maken.

1.3.2 De impliciet ethische actor

De tweede categorie is de 'impliciet ethische actor' (AMA-2). Hierbij wordt automatisch rekening gehouden met specifieke ethische overwegingen. De ethische keuze, bijvoorbeeld het beschermen van privacy, zit namelijk al in het ontwerp. Voorbeelden zijn internetbrowsers zoals Firefox of zoekmachines zoals DuckDuckGo. Of toepassingen zoals 'Ziggo Safe Online' die pornografische inhoud blokkeren. Een 'impliciet onethische actor' bestaat overigens ook. Denk aan een *spam zombie*, bedoeld om een virus op je computer te verspreiden.

Op het vlak van *kunstmatige morele adviseurs* impliceert AMA-2 dat de AI een bepaalde ethische overweging, waarde of criterium *expliciet* meeneemt, om tot een moreel oordeel of advies te komen. Een voorbeeld is het rangschikken van de volgorde van vaccineren. AMA-2 zou ingesteld kunnen worden volgens een aantal ethische criteria en daar vervolgens rekening mee houden in het beoordelen wie als eerste een vaccin moet krijgen. In tegenstelling tot de eerste categorie waar AI simpelweg data aanlevert die de mens kunnen helpen om een oordeel te vormen, is dit type AMA specifieker. Het volgt de morele normen en criteria waarmee het ingesteld is, maar heeft verder geen 'actieve' inbreng (zie ook Nagenborg, 2007,

6 Giubilini & Savulescu (2018) gebruiken de afkorting AMA voor kunstmatige morele *adviseurs*. In dit preadvies heeft AMA zowel betrekking op kunstmatige morele actoren/agents als op kunstmatige morele adviseurs. Dit vanuit de gedachte dat de verschillende typen kunstmatige morele actoren ook steeds een andere rol (kunnen) spelen in morele oordeelsvorming en advies.

7 Deze data zijn niet automatisch moreel neutraal. Bij de selectie van data kunnen meer of minder impliciete morele oordelen en vooroordelen (*bias*) zoals bekend een rol spelen.

p. 131). Dat is anders bij de derde en vierde categorie AMA's: de 'expliciet ethische actor' (AMA-3) en de 'volledig ethische actor' (AMA-4) hebben wel een actieve inbreng op moreel vlak. Dat maakt deze toepassingen (nog) spannender.

1.3.3 De expliciet ethische actor

De expliciet ethische actor[8] kan moreel relevante informatie over verschillende situaties heen identificeren, analyseren en beslissingen nemen over wat er in die situaties moet gebeuren. In tegenstelling tot AMA-2 zou deze AI in staat zijn om morele normen en criteria zelf bij te stellen. AMA-3 is er nog niet, maar is volop in ontwikkeling, zoals de volautomatische zelfrijdende auto. Stel dat een zelfrijdende auto op het punt staat om te botsen en de *moreel beste* keuze gemaakt moet worden, bijvoorbeeld om zoveel mogelijk levens te redden.[9] Wie moet gered worden: de zwangere vrouw of de oudere man? Een computer kan in korte tijd zeer veel berekeningen maken en kan dus, in tegenstelling tot de mens die op zo'n moment gedwongen is om intuïtief te beslissen, snel een onderbouwde morele keuze maken, *gebaseerd* op de criteria die programmeurs hebben ingegeven in de algoritmen (zie ook Gabriels, 2021).

In tegenstelling tot de 'impliciet ethische actor' (AMA-2) houdt de expliciete actor meer rekening met de context. Op een top-down manier worden criteria en regels in de AI geprogrammeerd, maar bottom-up leert de zelfrijdende auto op basis van binnenkomende gegevens (bijvoorbeeld uit sensoren en camera's op de auto). AI zou de criteria die de programmeurs hebben ingegeven dus terzijde kunnen schuiven, op basis van nieuw opgedane inzichten door machine learning. Dit zou dan impliceren dat de auto in de praktijk 'bewust' kan afwijken van de regels die mensen erin stopten, op basis van berekeningen, en kan overgaan tot een bepaalde actie (beslissing). Een criterium kan bijvoorbeeld zijn dat de auto in urgente situaties altijd veilig langs de weg moet parkeren. Maar op basis van berekeningen zou de auto ook kunnen beslissen dat het veiliger is om in het midden van de baan te stoppen en zo van het criterium af te wijken.

8 Ook wel 'uitgesproken ethische actor' genoemd.

9 Dat zoveel mogelijk levens gered moeten worden, is een voorbeeld van een moreel (utilitaristisch) criterium dat van invloed is op ons oordeel over wat in dit voorbeeld 'juist' is om te doen. In hoofdstuk 3 ga ik kort in op de vraag of het inbouwen van moraliteit in verschillende typen AMA verenigbaar is met meerdere ethische perspectieven, of dat we ons daarbij (zouden) moeten committeren aan een bepaalde stroming.

1.3.4 De volledig ethische actor

De volledig ethische actor (AMA-4) zou kunnen handelen op een vergelijkbaar niveau als dat van de mens. Mensen zijn morele wezens die in staat zijn tot kritische reflectie op hun eigen gedrag, principes, oordelen, waarden en normen. Mensen zijn ook verantwoordelijk voor hun gedrag, zolang ze wilsbekwaam zijn en vrij handelen. Deze sterkste vorm van AMA zou ook moreel verantwoordelijk zijn en in staat tot (bewuste) reflectie op het eigen gedrag. Dit type zou zelf morele oordelen kunnen vellen over een breed scala aan situaties en contexten en zou deze oordelen ook kunnen rechtvaardigen. AMA-4 zou een bepaalde vorm van bescherming kunnen genieten, bijvoorbeeld dat die nooit zomaar, zonder gegronde reden, uitgeschakeld mag worden, omdat die op een dieper niveau kan denken en bewustzijn heeft.

Sciencefiction bracht al soortgelijke morele machines met (zelf)bewustzijn in beeld. De Zweedse fictieserie *Real Humans* geeft er een interessante invulling aan: *hubots* – een samentrekking van *human* en robot – leven samen met mensen en komen in opstand om meer rechten op te eisen. Mensen moeten niet alleen respect voor hen hebben, maar ook hun belangen serieus nemen. Dit type robot kan ook moreel verantwoordelijk gesteld worden bij een overtreding. Tekstbox 1 beschrijft aan welke voorwaarden daarvoor moet worden voldaan.

AMA's en morele verantwoordelijkheid

Onder bepaalde voorwaarden zou de volledig ethische actor moreel verantwoordelijk (en schuldig) kunnen worden geacht (zie van de Poel & Royakkers, 2011). Ten eerste moet AMA-4 een overtreding maken door bijvoorbeeld een norm te schenden. Ten tweede moet er een direct verband zijn tussen de overtreding en de gevolgen ervan. Ook iets niét doen, dus niet handelen of nalatigheid, kan een causale bijdrage leveren. Ten derde, de AMA kon redelijkerwijs weten en voorzien dat er negatieve gevolgen zijn van een bepaalde handeling (of door iets te verzuimen). Daarom worden betrokkenen geacht om alles, binnen hun mogelijkheden, te doen om de mogelijke gevolgen te kennen en in te schatten. En ten slotte moet er sprake zijn van vrij handelen en dus niet van handelen onder dwang.

Tekstbox 1: Onder bepaalde voorwaarden kan een AMA moreel verantwoordelijk worden gehouden

1.4.1 Stand van de techniek en toekomstperspectief

Waar staan we nu met de ontwikkeling van kunstmatige morele actoren/adviseurs? Volgens filosofen Wendell Wallach en Colin Allen zitten robots en AI vandaag de dag op de grens tussen operationele moraliteit – waarbij morele besluitvorming volledig in handen ligt van menselijke supervisie – en functionele moraliteit, het vermogen om zelf morele uitdagingen in te schatten en er een antwoord op te bieden (Allen & Wallach, 2014, p. 57; zie ook Wallach & Allen, 2009). Tussen AMA-2 en AMA-3 dus. Voor écht functionele moraliteit, zoals bij de volautomatische zelfrijdende auto (AMA-3), is de huidige technologie nog te beperkt, maar die beperkingen beschouwen Wallach en Allen op termijn als overkomelijk. Op dit moment kan een AMA nog lang niet vrij handelen en ook het begrip van context is ondermaats. Maar als zou blijken dat AMA-4 moreel beter (minder feilbaar) is dan wij, dan zou deze wellicht onze morele mentor kunnen worden en ons kunnen bijstaan met het beslechten van morele dilemma's.

1.4.2 De vier typen AMA's op het continuüm van zwakke naar sterke AI

Eerder noemde ik de tweedeling tussen zwakke en sterke AI. De tweedeling is nogal beperkt; er zou een continuüm of op zijn minst een tussencategorie moeten zijn die erkent dat je niet alleen meer uit AI kunt halen dan de vooraf gedefinieerde functies, maar dat een AMA ook 'bewust' van regels zou kunnen afwijken, zoals bij de volautomatische auto. De vier typen AMA's kunnen dan als volgt worden ingedeeld: AMA-1 en AMA-2 behoren tot zwakke AI, waarbij AMA-2 al minder zwak is dan AMA-1. AlphaGo hoort bij AMA-2, omdat die nog binnen een beperkte set van functies opereert, ook al haalden de onderzoekers er meer uit dan ze erin stopten. AMA-3 staat tussen zwakke en sterke AI in; deze tussencategorie is er nog niet maar is volop in ontwikkeling. AMA-4, een volledig ethische actor, zou een vorm van sterke AI zijn met zelfbewustzijn.

1.4.3 Wringende concepten

In de afgelopen jaren werden machines steeds 'autonomer'. Autonomie is hier een wat misleidend begrip: het betekent niét dat de technologie buiten menselijke controle valt. In deze context betekent autonomie het bereiken van een

doelstelling die niet op voorhand al volledig werd vastgelegd door programmeurs. Hoe minder menselijke interventie er is in de werking en hoe breder de reikwijdte van de actie, des te autonomer de machine, in deze betekenis. Robots met een volledig autonoom moreel handelingsvermogen, zoals de robots (AMA-4) in *Machines Like Me* en *Real Humans*, zijn volgens Wallach en Allen slechts in theorie mogelijk. Dat betekent dat gebruik van het concept autonomie in het kader van AI vooralsnog wringt, zeker ten aanzien van AMA 1 tot en met 3. In dit preadvies betekent autonomie toegepast op AMA's dus het bereiken van doelen met minder menselijke tussenkomst of supervisie.

Ook op termen als 'agent' en 'moral agent' is kritiek mogelijk. Zo leiden ze volgens sommige auteurs (o.a. Johnson, 2006; Nagenborg, 2007; Sharkey, 2017) tot te veel verwarring. Om te beginnen kan het begrip 'agent' tamelijk zwak maar ook vrij sterk worden opgevat. In een zwakke betekenis verwijst 'agent' naar software of een machine die een bepaalde handeling of taak uitvoert (Johnson & Miller, 2008), of naar *"a person or thing that produces a particular effect or change"*.[10] Zo'n effect kan ethische gevolgen hebben, zonder dat de machine zelf in staat is om ethisch te reflecteren. Een recent voorbeeld is het gebruik van algoritmen in de toeslagenaffaire. De algoritmen voerden een taak uit volgens de richtlijnen waarmee ze werden ingesteld, die vervolgens een specifiek effect had. Op een niet-transparante wijze, die bovendien in strijd was met de Algemene Verordening Gegevensbescherming (AVG)[11], profileerde het risico-classificatiemodel mensen met 'exotische' namen en dubbele nationaliteiten. In deze zwakke betekenis kunnen overigens ook katten en honden als 'agent' gedefinieerd worden, in dit geval zelfs agents met bewustzijn. Een 'agent' hoeft dus niet per se te kunnen redeneren en handelen op het niveau van de mens; iets wat een sterke interpretatie van het begrip 'agent' wel vereist. Maar als de eigenschappen van 'agents' zo ver uiteenlopen, is de term 'agent' dan wel in alle gevallen even geschikt?

We zien iets vergelijkbaars bij het begrip 'moral agent', dat ook meer of minder sterk kan worden opgevat. In de sterke betekenis moet sprake zijn van een vrije wil, vrijheid van handelen, rationaliteit en bewustzijn om van een 'morele actor' te kunnen spreken (Himma, 2009). In dit geval gaat het om een actor die aan zichzelf morele standaarden kan opleggen en al dan niet beslist om

10 Zie https://dictionary.cambridge.org/dictionary/english/agent Geraadpleegd op 12 mei 2021.

11 Zie https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf Geraadpleegd op 18 april 2021.

die te volgen (vrijheid van keuze), en die het verschil tussen goed en kwaad kent en in verschillende contexten een afweging kan maken. 'Moral agency' impliceert hier ook moreel verantwoordelijk (kunnen) zijn voor het betreffende gedrag. De kritiek zou dan kunnen zijn dat het weinig zin heeft om 'actoren', die niet vrij kunnen handelen en geen voornemens kunnen maken tot het stellen van een bepaald gedrag, enige morele verantwoordelijkheid toe te dichten. *"Computer systems do not have mental states, and even if they could be construed as having mental states, they do not have intendings to act, which arise from an agent's freedom"* (Johnson, 2006, p. 195). De mogelijkheid van een AMA hangt dan af van het gegeven of we ooit een AI-systeem kunnen bouwen met bewustzijn, dat emoties zoals pijn kan ervaren, kan reflecteren en afwegingen kan maken (AMA-4). In de zwakke betekenis gaat het slechts om AMA's die ethische impact kunnen hebben, zoals bij het eerder vermelde voorbeeld van de toeslagenaffaire. Critici (zie bv. van Wynsberghe & Robbins, 2018) stellen dat het beter is om niet van AMA's, maar bijvoorbeeld wel van een 'veilige machine' (als veiligheid het doel is) te spreken en om begrippen die we doorgaans toepassen op mensen (bv. 'moral agent') weg te houden van een AI-context.

Om een aantal redenen blijf ik de term AMA in dit preadvies hanteren. Ten eerste, gebruik van technologie heeft veelal betrekking op *meerdere* morele waarden tegelijk, zeker bij kunstmatige morele actoren. Het centraal stellen van één waarde in de benaming van een technologie (zoals 'veilige machine') onttrekt aan het zicht dat bij gebruik van een technologie vaak een afweging moet worden gemaakt tussen verschillende morele waarden, bijvoorbeeld tussen veiligheid en privacy. Bovendien geeft een term als 'veilige machine' automatisch prioriteit aan één van die waarden. Of dat de meest belangrijke waarde is, staat niet automatisch of op voorhand vast.[12] Ten tweede, gebruik van de term AMA kan voorlopig ook een ander belangrijk punt scherp in zicht houden, namelijk de ontwikkeling waarin we functies, eigenschappen en rollen in technologie proberen in te bouwen die tot de kern van ons mens-zijn behoren, in dit geval moraliteit en morele oordeelsvorming. Dat de technologie nog niet aan de standaarden voldoet die we daarvoor bij de mens hanteren, doet daar niet onmiddellijk aan af. Als AI een rol heeft in morele oordeelsvorming, geef ik er daarom de voorkeur aan om dat in de beschrijving van de technologie tot uiting te brengen, ook als de

12 Dat een technologie veilig is, kun je bovendien met een keurmerk wellicht afdoende duiden.

technologie in belangrijke opzichten (nog) tekortschiet. Ik erken wel dat de term tot verwarring kan leiden. In dit preadvies moet daarom de indeling in de vier typen steeds voor ogen gehouden worden.

1.5 Moreel oordelen: een samenspel van intuïtie en rede

Als we AI willen inzetten voor morele oordeelsvorming, moeten we onder andere inzage hebben in begrippen zoals morele intuïtie en moreel redeneren. Binnen disciplines zoals filosofie, morele psychologie en moraalwetenschappen bestaat echter geen eensgezindheid over hoe we moraliteit moeten definiëren en hoe de mens zelf tot morele oordelen komt. Moreel psycholoog Jonathan Haidt (2013, p. 314) omschrijft moraliteit bijvoorbeeld als *aangeleerde* sets van waarden, deugden, normen en praktijken en *aangeboren* psychologische mechanismen die samenwerken om eigenbelang te onderdrukken of reguleren en om coöperatieve samenlevingen mogelijk te maken. Moraliteit is, net zoals taal, een product van evolutie en vormde een essentiële bijdrage aan onze overleving (Allott, 1991). De bouwstenen van moraliteit zijn dus ouder dan de Homo sapiens. Een enge definitie van moraliteit is samenwerking. Een simpel voorbeeld uit de natuur is het delen van voedsel. Een organisme dat samenwerkt, heeft meer kans op overleven. De menselijke morele psychologie heeft zich aangepast als reactie op samenwerkingsproblemen. Morele oordelen zijn dan het product van evolutionaire druk voor coöperatief gedrag in zelfzuchtige organismen (Leben, 2019; zie ook Gabriels, 2021). Dit leidt tot twee interessante vragen. Ten eerste, in hoeverre kunnen we moraliteit in een machine bouwen of aanleren, die dan vervolgens, aan de hand van steeds nieuwe data, inzichten ontwikkelt, net zoals AlphaGo? Ten tweede, in hoeverre kunnen we menselijke morele evolutie nabootsen in een gesimuleerde omgeving? In hoofdstuk 3 kom ik op deze vragen terug. Zoals we daar zullen zien, worden algoritmen volop samenwerkingsstrategieën aangeleerd, onder meer door hen het bordspel *Diplomacy* te laten spelen.

Hoe vormt de mens eigenlijk een moreel oordeel? Volgens moreel psycholoog Jonathan Haidt sturen onze intuïties in eerste instantie het moreel redeneren. Bij het vellen van een moreel oordeel is de intuïtie leidend en dient de rede onze intuïtie. Haidt volgt daarin David Hume, die in de 18de eeuw stelde dat de rede de slaaf is van onze passies. Een voorbeeld: Haidt legde aan proefpersonen een taboescenario voor, over een broer en een zus die eenmalig seks met elkaar hebben. Enkele mogelijke bezwaren zijn in het scenario afgedekt: de broer en zus

gebruiken voorbehoedsmiddelen (anticonceptiepil en condoom) en de seks vindt plaats met wederzijdse toestemming. Wat bleek? Het merendeel van de respondenten keurde het scenario intuïtief af en zocht vervolgens naar argumenten om die intuïtie te staven. Als respondenten meer tijd krijgen om erover na te denken en als ze daarbij ook een onderbouwd argument krijgen aangereikt, zijn ze eerder geneigd om ermee in te stemmen dat de broer en zus eenmalig seks kunnen hebben. In een rustig gesprek waarin argumenten worden aangereikt kunnen intuïties én oordelen veranderen (Haidt, 2013, p. 55). In een open gesprek, waarbij wederzijds begrip en affectie getoond wordt, is er meer bereidheid om van gedachten te veranderen. De rede is dus misschien niet volledig de slaaf van onze passies.

Stel nu dat we ons baseren op deze inzichten bij het bouwen van een AMA. We moeten dan rekening houden met zowel intuïtie als redeneren, inclusief het wikken en wegen van argumenten om het oordeel eventueel bij te stellen. In principe moeten we dan eerst gedetailleerd weten en vervolgens beschrijven hoé mensen moreel oordelen, voordat we dat kunnen overbrengen naar een machine. Vervolgens moeten we beide kunnen vergelijken, wat allerminst eenvoudig is, om te kijken in hoeverre we bij de AMA van een moreel oordeel kunnen spreken en of dit in de lijn ligt van menselijke morele oordelen of toch iets totaal anders is.

Welke mogelijkheden biedt AI om morele intuïtie in te bouwen in techniek? Regelgestuurde AI lijkt om te beginnen geen optie. Mensen leren intuïtie en *knowhow* namelijk door oefening en ervaring (Dreyfus & Dreyfus, 1986), en menselijke intuïtie valt niet zomaar te formaliseren. Een menselijke expert deelt gedrag niet op in abstracte regels om ze vervolgens te kunnen volgen. Een beginner trekt nog conclusies met behulp van regels en feiten, zoals iemand die leert autorijden of een chirurg die leert opereren, maar een expert ziet intuïtief wat hij of zij moet doen zonder (expliciet) regels toe te passen, gebaseerd op ervaring (Dreyfus & Dreyfus, 1986, p. 385). Datagestuurde AI (machine learning en deep learning) lijkt wel mogelijkheden te bieden om AI een soort van morele 'intuïtie' te laten ontwikkelen, door data aan te leveren en algoritmen er patronen in te laten zoeken. Maar is intuïtie wel een passend begrip? Bij AI gaat het immers nog steeds om berekeningen. Net zoals bij AlphaGo zou je een algoritme met miljoenen scenario's (in dit geval over moraliteit) kunnen confronteren, daartegen laten 'spelen' (als vorm van selfplay) en laten leren uit fouten. AI zal een patroon vinden, bijvoorbeeld over goed en slecht, en kan zo een soort van intuïtie ontwikkelen.

Stel dat het vraagstuk van morele intuïtie inbouwen in techniek langs deze lijnen kan worden opgelost, dan zijn we er nog niet. Een 'goed' en overtuigend moreel oordeel gaat namelijk verder dan intuïtie: het is onderbouwd met kwaliteitsvolle argumenten. Als iemand oordeelt dat een specifieke handeling verkeerd is, moet hij of zij kunnen onderbouwen waarom. Zo wordt willekeur vermeden en kunnen anderen inschatten in hoeverre het oordeel bijvoorbeeld onderhevig is aan vooroordelen. AI zal dus ook moeten leren argumenteren. In het juridische domein wordt al langer onderzocht hoe AI ingezet kan worden om juristen bij te staan met het evalueren van juridische argumentatie (zie bv. Walton, 2016). Het gaat in dit geval vooral om het modelleren van juridische argumentatie. In Nederland onderzoeken filosofen in hoeverre een 'argumentatiemachine' drogredenen kan herkennen.[13] Er wordt dus zeer interessant onderzoek gedaan binnen *Natural Language Processing* (NLP), het domein binnen AI-onderzoek dat zich bezighoudt met taal- en spraaktechnologie. Het onderzoek staat echter nog in de kinderschoenen.

Om moreel te kunnen redeneren, zal AI bovendien aan enkele minimale vereisten moeten voldoen (gebaseerd op Himma, 2009). Zo zal die morele termen als 'goed', 'kwaad', 'plicht', en '(on)toelaatbaar' moeten begrijpen en kunnen toepassen. Verder zal AI fundamentele morele principes correct moeten begrijpen, zoals dat en *waarom* het verkeerd is om mensen opzettelijk schade toe te brengen, behalve als er een reden is die dit zou rechtvaardigen. Ook moet AI het vermogen hebben om de relevantie van morele regels in verschillende contexten te evalueren en toe te passen.

1.6 Normatieve vragen – een eerste opbrengst

De theoretische bespreking tot dusver levert ook een eerste verzameling normatieve vragen op. De eerste hangen samen met de discussie of we de term AMA wel moeten gebruiken. Voorstanders van AMA's, zoals in de vorm van zelfrijdend transport, stellen dat deze schade kunnen voorkomen en de veiligheid kunnen vergroten. Tegenstanders van AMA's (zie bv. Johnson & Miller, 2008; van Wynsberghe & Robbins, 2018) verzetten zich tegen de ontwikkeling ervan, zolang er

13 Zie https://humane-ai.nl/seed-funded-projects-2020-2021/towards-an-epistemological-and-ethical-explainable-ai/, zie ook https://www.volkskrant.nl/wetenschap/deze-filosofen-maken-een-argumentatiemachine-die-hopen-zij-drogredenen-uit-toespraken-kan-vissen~b77a1a75/ Geraadpleegd op 18 april 2021.

geen gegronde redenen zijn om ze te ontwikkelen. Of we voor of tegen de ontwikkeling van (specifieke) AMA's moeten zijn, wil ik hier nog even tussen haakjes zetten. Een eerste relevante vraag is die naar redenen om AMA's te ontwikkelen: wat is het doel, en is dat een legitiem doel? Daar hangt een tweede vraag mee samen, die ook raakt aan innovatiebeleid: moet er eerst een legitiem doel zijn, voordat een technologie ontwikkeld mag worden? En hoe bepalen we dan (en wie?) of het doel legitiem is?

Er zijn ook al enkele normatieve vragen te benoemen die samenhangen met specifieke typen AMA's. Zo zagen we dat AMA types 3 en 4 zélf tot een moreel oordeel kunnen komen. Dat roept de vraag op in hoeverre en in welke contexten het problematisch is als dat oordeel afwijkt van dat van mensen. In het verlengde hiervan: de expliciet ethische actor (AMA-3) en de volledig ethische actor (AMA-4) worden gekenmerkt door een zekere mate van 'vrijheid' en autonomie (minder menselijke supervisie en tussenkomst). Is deze vorm van 'vrijheid' wenselijk? Er is bij een volautomatische zelfrijdende auto (AMA-3) immers geen zekerheid over het gedrag van de auto, net zoals AlphaGo een zet deed waar mensen in eerste instantie niet aan dachten. Zouden specifieke voorwaarden moeten worden gesteld, omdat er hier, in tegenstelling tot bij AlphaGo, levens op het spel kunnen staan? Bijvoorbeeld dat de auto systematisch moet hebben bewezen dat die het beter kan dan de mens, en dat er hoe dan ook transparantie en inzicht moet zijn in hoe die tot een beslissing komt? Tot slot, AMA-4 zou moreel superieur (kunnen) zijn aan de mens. Ook dat roept een aantal interessante vragen op, zoals hoe wij, mensen, een moreel superieure AMA (zouden) kunnen evalueren. En in hoeverre zouden mensen aanvaarden dat er een kunstmatige morele actor én adviseur is die op een *hoger* moreel niveau kan denken? Zou de mens die adviezen naast zich (kunnen) neerleggen? Of in opstand komen, zoals in *Real Humans*, waarbij mensen van vlees en bloed zich verenigen in een actiegroep van échte mensen?

Een volgende uitdaging is de vraag met welke data en met welk moreel perspectief AI moet worden getraind. Het moreel juiste om te doen, onder alle omstandigheden, is datgene waarvoor *de beste redenen* (argumenten) zijn om het te doen, en daarbij evenveel gewicht te geven aan de belangen van elk individu dat zal worden beïnvloed door wat men doet (Rachels, 2007, p. 11, p. 14). Nog los van de vraag of AI in staat zal zijn om dit te kunnen, is er echter in de ethiek geen consensus over wat die 'beste redenen' zijn. Ik kom daar in hoofdstuk 3 op terug.

Het blijft tot slot uitermate belangrijk om kritisch te kijken naar hoe AI patronen ontdekt. Niet elk patroon is immers wenselijk: AI kan ook voorkeuren op basis van populariteit ontwikkelen. Een andere uitdaging is, zoals we later nog zullen zien, om de contextafhankelijkheid van morele regels en gedrag te leren.

In het volgende hoofdstuk bespreek ik eerst vier casussen om de verschillende typen AMA verder te verhelderen, en te kunnen reflecteren op welke (verdere) technische, conceptuele en morele vragen de typen AMA's meebrengen.

2 Vier casussen

In dit hoofdstuk bespreek ik vier casussen. De casussen dienen primair als voorbeelden om de vier types van Artificial Moral Agents (AMA's) nader uit te werken. De cases zijn geselecteerd op basis van twee criteria. Ten eerste moeten ze een inkijk bieden in de toenemende autonomie, complexiteit en morele gevoeligheid van de vier typen AMA's. Elke casus is daarom geënt op een ander type AMA. Ten tweede moeten ze helpen om te illustreren welke technische, ethische en conceptuele uitdagingen bij het 'inbouwen' van moraliteit komen kijken.[14] We zullen zien dat er een groot verschil is tussen een AMA die louter data aanlevert, maar zelf niet oordeelt, en een AMA die moet aangeven wat het meest wenselijke is (en waarom) en volgens welke normatieve ethische theorie(ën) die redeneert.

2.1 Casus 1: Gebrek aan informatie bij euthanasie

Zoals we eerder zagen bij algoritmen die elektronische patiëntendossiers scannen, kan de inzet van AI efficiënt zijn in situaties waar onder tijdsdruk een beslissing genomen moet worden. Een voorbeeld is besluitvorming over euthanasie. Euthanasie is in Nederland in het strafrecht opgenomen. Alleen artsen mogen onder voorwaarden op verzoek van een patiënt het leven van de patiënt (helpen) beëindigen. Dat is strafbaar, *tenzij* aan een aantal zorgvuldigheidseisen wordt voldaan. Naast een vrijwillig en weloverwogen verzoek van de persoon zelf, moet onder andere sprake zijn van uitzichtloos en ondraaglijk lijden.[15] Het is voor een arts niet altijd eenvoudig om te bepalen in hoeverre het lijden uitzichtloos is, of er nog behandelingsmogelijkheden zijn, en of het lijden ondraaglijk is. Ik concen-

14 Wat er komt kijken bij het 'inbouwen' van moraliteit in AI, staat centraal in hoofdstuk 3.

15 Zie https://www.rijksoverheid.nl/onderwerpen/levenseinde-en-euthanasie/zorgvuldigheidseisen Geraadpleegd op 18 april 2021.

treer me op het laatste: het bepalen of het lijden voor een patiënt ondraaglijk is.

In theorie zou AI daar een rol bij kunnen spelen. Extra data kunnen mogelijk meer inzicht bieden. AI-systemen zouden 24/7 over een langere periode data kunnen genereren. Dit kan onder meer via sensoren op de huid, die ook hartslag en bloeddruk meten, en sensoren in huis, die de beweging en activiteiten registreren. Om die rol te kunnen spelen zijn onder meer duidelijke parameters nodig voor het meten van welzijn en lijden. Het lijden van een patiënt is subjectief en houdt verband met wat voor een patiënt het meest essentieel is in het leven. Dat het lijden voor een patiënt ondraaglijk is, moet voor de arts 'invoelbaar' zijn. AI zou een rol kunnen spelen in het verkrijgen van inzicht in wat het lijden voor de patiënt ondraaglijk maakt, door in de vorm van langdurige *lifelogging* gegevens te verzamelen over wat de patiënt essentieel vindt voor een goed en waardevol leven. Stel dat essentiële elementen voor welzijn en geluk zijn weggevallen, dan zou dit onmiskenbaar kunnen blijken uit de jarenlange metingen. De data zouden kunnen bijdragen aan inzage in wat, vanuit het persoonlijke perspectief van de patiënt, ondraaglijk is.

Deze casus valt onder AMA type 1, de 'ethische impact actor'. AMA-1 beperkt zich tot het genereren en analyseren van data en doet een voorspelling. Dit type AMA is een hulpmiddel dat, bijvoorbeeld in de vorm van een app, ondersteuning zou kunnen bieden om mogelijk tot een beter geïnformeerde beslissing te komen. In de voorbeeldcasus over of wel of niet aan specifieke zorgvuldigheidseisen voor het verlenen van euthanasie is voldaan. De data kunnen het ethische proces en de besluitvorming weliswaar beïnvloeden, maar het onderbouwde oordeel en besluit liggen bij de arts.

Tussentijdse reflectie

Onder bepaalde voorwaarden zou AI in deze context naar mijn idee meerwaarde kunnen hebben. De technologie moet vanzelfsprekend betrouwbaar en kwaliteitsvol zijn. Gezien de aard van de metingen en data moeten data-ethiek en patiëntenrechten de nodige aandacht krijgen. Zo moet het meten menswaardig gebeuren en mag het niet worden opgedrongen. Het bepalen van parameters zal in ieder geval door mensen moeten gebeuren; AI is daar niet toe in staat. Ook het bepalen van drempelwaarden (bijvoorbeeld vanaf wanneer er voldoende data zijn om tot een onderbouwd oordeel te komen), is een taak voor mensen. De *interpretatie* van de data mag niet genegeerd worden en moet door verschillende

personen gebeuren, om de kans te beperken dat er zaken over het hoofd gezien worden en om al te simplistische gevolgtrekkingen te vermijden. Ook de patiënt moet hierbij betrokken worden.

Ook als aan genoemde voorwaarden wordt voldaan, speelt er een aantal ethische kwesties. Zo brengt langdurige lifelogging een verregaande vorm van privacyverlies met zich mee. Bovendien is niet zeker dat meer data altijd onomstotelijk tot meer duidelijkheid leiden. Belangrijke vraag is verder of gebruik van AI voor dit doel invloed heeft op de manier waarop wordt bepaald of het lijden van een patiënt voor hem of haar ondraaglijk is. Voorkomen zou moeten worden dat door gebruik van AI om meer inzicht te krijgen in wat het lijden subjectief ondraaglijk maakt (ondraaglijk voor *deze* patiënt, vanuit zijn of haar *eigen* perspectief), de facto een meer objectieve toets wordt aangelegd, waarbij 'invoelbaarheid' van het lijden afhankelijker wordt van verifieerbare en objectiveerbare data.

Vervolg op de casus

Een AMA zou ook op een ander punt een rol kunnen spelen in de context van euthanasie, namelijk in het geval waarin een patiënt niet meer in staat is om aan te geven wat hij of zij wil, en er geen (duidelijke) wilsverklaring is.[16]

In theorie zou AI kunnen helpen om de wil van de patiënt op de één of andere manier te achterhalen of te reconstrueren.[17] Zo zouden we met onder andere *wearables* en apps jarenlang al onze activiteiten en beslissingen kunnen vastleggen als een vorm van lifelogging. Een goed voorbeeld daarvan komt tot uiting in de VPRO-documentaire *De Techmens*: die brengt internetpionier Gordon Bell in beeld die sinds 1998 zijn leven met allerhande technologieën in data vastlegt.[18] Hij draagt onder meer een camera die zijn dagelijkse activiteiten registreert. Het is denkbaar om ons leven en de keuzes die we maken zo gedetailleerd in data vast te

16 Het voorbeeld is geïnspireerd door Biller-Andorno & Biller (2019). Zij presenteren een gedachte-experiment waarin een patiënt een zware beroerte heeft gehad en het medisch team moet kiezen om de patiënt te behandelen of de behandeling te staken. Behandelen zou ernstige handicap tot gevolg kunnen hebben, als de patiënt de behandeling al overleeft. Het staken van de behandeling zou tot de dood leiden, door de patiënt niet langer kunstmatig te beademen. De familie wordt geconsulteerd om te achterhalen wat de patiënt gewild zou hebben, maar de patiënt sprak haar wens tegen geen enkel familielid uit. Op aanraden van een familielid wordt dan gebruik gemaakt van een app die de wens van de patiënt voorspelt, op basis van een aantal gegevens die in het systeem moeten worden ingegeven.

17 Ik neem daarbij aan dat verder aan alle zorgvuldigheidseisen die de wet stelt, is voldaan.

18 Zie https://www.vpro.nl/programmas/tegenlicht/lees/biografieen/b/gordon-bell.html Geraadpleegd op 12 mei 2021.

leggen, dat ook voor situaties waarin we niet zelf expliciet en/of ondubbelzinnig een keuze hebben gemaakt, voorspeld of gereconstrueerd kan worden wat we in dat geval zouden willen.

Reflectie

Het vervolg op de casus roept verschillende vragen op, niet in het minst over privacy. Sommige van deze vragen kunnen ondervangen worden door te borgen dat aan bepaalde voorwaarden wordt voldaan.

Om te beginnen, als op basis van informatie die jarenlang onafgebroken verzameld wordt, een medische beslissing genomen moet worden, dan moeten derden (bijvoorbeeld artsen) er toegang toe kunnen hebben. Aangezien het om zeer persoonlijke en gevoelige gegevens gaat, moeten die goed beveiligd zijn. Waar worden de data opgeslagen? Die servers moeten in ieder geval regelmatig geüpdatet worden om voldoende weerbaar te zijn tegen cyberaanvallen.

Een andere vraag is vanaf wanneer (welke leeftijd) je dan het beste kunt/moet beginnen met data verzamelen, zodat het algoritme met een hoge accuraatheid kan voorspellen wat jouw wil is? Voor veel mensen gaat zo'n vorm van lifelogging overigens mogelijk te ver; laat staan dat ze anderen er zomaar toegang toe geven.

In de huidige praktijk kunnen familieleden een rol spelen bij het verkrijgen van meer duidelijkheid, bijvoorbeeld over wat de patiënt precies bedoeld heeft in zijn of haar wilsverklaring. Maar familieleden kunnen net als iedereen vooroordelen hebben en hun waardepatronen kunnen verschillen van die van de patiënt in kwestie. In die zin zou een AI-systeem mogelijk een 'juistere' interpretatie kunnen bieden, gebaseerd op persoonlijke gegevens en eerdere keuzes van de patiënt. Daar staat echter tegenover dat het onduidelijk is in hoeverre de technologie in staat zal zijn om een voldoende *betrouwbare* voorspelling te maken. Mensen kunnen gaandeweg (of na een levenservaring met grote impact) van standpunt veranderen en die verandering moet dan in de data naar voren komen, zodat er geen verkeerde interpretatie ontstaat. Sowieso gaat het altijd om het voorspellen van een keuze op basis van gegevens uit het verleden. Op dit moment is het onzeker hoe goed voorspellingen over zulke complexe onderwerpen kunnen worden en het blijft sowieso bij een kansberekening.

Maar dan zijn we wat het *vervolg* op de casus betreft terug bij af. Gebruik van AI voor het interpreteren of reconstrueren van de wil van een patiënt in

het kader van euthanasie, is mijns inziens op dit moment niet wenselijk. Ook als de accuraatheid van de technologie sterk verbeterd kan worden, blijft een fundamenteel vraagstuk over. In elk geval voor die situaties waarin een patiënt niet meer in staat is om zijn of haar wil te bepalen, en een wilsverklaring *ontbreekt*. Volgens artikel 2, tweede lid van de Wet toetsing levensbeëindiging op verzoek en hulp bij zelfdoding (Wtl) kan een schriftelijke wilsverklaring in de plaats treden van een actueel verzoek, als de patiënt niet meer in staat is om zijn of haar wil te bepalen.[19] Gebruik van AI om bij het volledig *ontbreken* van een wilsverklaring de wil van een patiënt te reconstrueren zou naar alle waarschijnlijkheid op gespannen voet staan met de wet.[20]

Onder voorwaarden zou gebruik van AI als ondersteuning om de ondraaglijkheid van het lijden van een patiënt te bepalen mijns inziens wel verdedigbaar kunnen zijn, al is het privacyvraagstuk hier wel groot.

2.2 Casus 2: Coördinatie en distributie van hulp en goederen bij ramp

Na de doortocht van orkaan Katrina in New Orleans in 2005 was de coördinatie en distributie van goederen slecht geregeld. Dit zou op termijn uitbesteed kunnen worden aan AI-systemen. Een algoritme is beter in het analyseren van grote hoeveelheden data en kan aangeven (rangschikken) wie eerst hulp moet ontvangen. Misschien kunnen ze dit op een bekwamere en rechtvaardigere manier doen dan mensen. De huidige aanpak van coronavaccinatie leidde bijvoorbeeld tot schandalen, omdat mensen voordrongen op de wachtlijst. In de Belgische stad Sint-Truiden werden de burgemeester en een aantal mensen uit haar entourage al gevaccineerd op een moment dat de 85-plussers aan de beurt waren.[21]

Reflectie

Het op termijn aan AI uitbesteden van zulke beslissingen (het bepalen van een volgorde) zou kunnen leiden tot een correcter en ethischer verloop. Als het mogelijk is om menselijke oordelen hier te vervangen door algoritmen en deze beter blijken te zijn, dan lijkt de toepassing mij ethisch gerechtvaardigd. Dit is een

19 https://wetten.overheid.nl/BWBR0012410/2020-03-19 Geraadpleegd op 24 mei 2021.

20 Gebruik van AI om bij het volledig ontbreken van een wilsverklaring de wil van een patiënt te reconstrueren, zou ook al een voorbeeld zijn van een AMA-2, de impliciet ethische actor.

21 Zie bijvoorbeeld https://www.hbvl.be/cnt/dmf20210511_97699061 Geraadpleegd op 12 mei 2021.

heel specifieke taak; de AMA, in dit geval de 'impliciet ethische actor' (AMA-2), heeft hier geen hoge graad van autonomie voor nodig. De systemen hoeven niet perfect te zijn om competent te zijn en kunnen steeds bijgesteld en verbeterd worden door mensen die de supervisie behouden (*human-in-the-loop*). AI kan een lijst (rangschikking) aanleveren waar meerdere mensen naar kijken, zodat het eindoordeel niet bij één iemand komt te liggen.

Het idee om coördinatie en distributie van goederen bij (medische) rampen (deels) uit te besteden aan AI, werd al eerder geopperd (zie Moor, 2007; Leben, 2019). AMA-2 houdt rekening met een aantal criteria, zoals dat van 'hoogste nood', maar heeft verder geen eigen inbreng. Hiervoor is in eerste instantie dus een lijst nodig met duidelijke instructies en criteria om de algoritmen vervolgens mee te trainen en te programmeren.

Logistiek is een technisch probleem waar je AI voor kan inzetten. Tegelijkertijd is er ook een normatieve vraag: wie krijgt, op basis van een aantal duidelijk geselecteerde criteria, voorrang? Dat zijn binnen de ethiek geen nieuwe vragen, maar hoe worden die criteria bepaald en wie beslist daarover? Deze casus maakt gebruik van rekenmodellen, die al op vergelijkbare manieren in veel domeinen ingezet worden, bijvoorbeeld om te berekenen hoe een virus zich zal verspreiden en wat er gebeurt als je *x* doet. Bij deze casus zouden (gevalideerde) modellen ontwikkeld kunnen worden die bij een ramp worden ingezet. Het doel is om de modellen te laten berekenen en classificeren, op basis van steeds nieuwe input. Data kunnen op diverse manieren verzameld en aangeleverd worden, in het geval van een natuurramp bijvoorbeeld ook door drones, uitgerust met sensoren en camera's.

Het zal in ieder geval een uitdaging zijn om de algoritmen met de juiste datasets te trainen die in een noodsituatie in real-time gevoed worden met steeds nieuwe gegevens. Bovendien moeten ook de sensoren, die informatie doorsturen, optimaal blijven functioneren onder extreme omstandigheden. Deze casus gaat uiteindelijk om een besluitvormingsproces binnen een beperkte opzet (zwakke AI), dat AI sneller, efficiënter en consistenter dan mensen zou kunnen berekenen en afleveren. Een valkuil is natuurlijk dat AI maar zo goed is als de data waar de algoritmen mee gevoed worden.

2.3 Casus 3: Een AI-gesprekspartner

Deze volgende casus is een vorm van AMA type 3, de expliciet ethische actor, en behelst een 'AI-gesprekspartner' die onze morele argumenten evalueert en helpt

met het oplossen van morele dilemma's. Ook dit idee werd al eerder geopperd door filosofen. Deze casus is gebaseerd op het idee van een socratische assistent (Lara en Deckers, 2020; zie ook Gabriels & Volkman, 2021). Net als de antieke filosoof Socrates kan de AI-assistent een aantal vragen stellen en vervolgens tekortkomingen in de argumentatie blootleggen om tot beter geïnformeerde morele beslissingen te komen.

In tegenstelling tot de vorige 2 casussen (AMA-1 en 2) zijn we nu aanbeland bij een categorie AMA die nog niet bestaat en die bovendien een grotere, meer vrije en 'autonome' rol zou hebben bij de morele oordeelsvorming. Hoewel het uiteindelijk nog steeds de mens is die moet handelen en beslissen, speelt de technologie hier ook een eigen rol in de morele oordeelsvorming, en dient als een morele mentor of adviseur van de mens.[22]

Het doel van de AI-gesprekspartner is om mensen beter moreel te laten redeneren. Zo'n AI-systeem daagt mensen intellectueel uit en wijst op denkfouten. De AMA is expliciet ethisch omdat die mensen uitdaagt om de zaken vanuit een ander perspectief te bekijken en om ons denkvermogen te trainen zonder het te sturen. De AMA kan bijvoorbeeld gebrekkige argumentatie onthullen zonder dat mensen gestuurd of gemanipuleerd worden. In tegenstelling tot een menselijke gesprekspartner is de AI-variant steeds beschikbaar en zou je die als een app kunnen downloaden op je smartphone.

Reflectie

Dit type AMA brengt naast de technische uitdagingen ook theoretische en normatieve uitdagingen mee, zoals de vraag met welke ethische theorie we de algoritmen moeten trainen (zie hoofdstuk 3). Deze AMA heeft in ieder geval een diep inzicht in menselijke morele oordelen en argumentatie nodig. De AI-gesprekspartner moet immers kunnen helpen met complexe morele dilemma's, zoals die bijvoorbeeld besproken worden in (bio-)ethische commissies.

In hoeverre kan de software ons sneller en efficiënter moreel advies geven dan ons brein ooit zou kunnen doen? Voordat deze kunstmatige morele adviseur een beslissing aanbeveelt, kan die data uit de omgeving en uit datasets verzamelen en analyseren, en ons vervolgens adviseren wat moreel het beste is om te doen, in overeenstemming met de morele criteria die we instelden in het

22 De casus ligt ook in de lijn van Giubilini en Savulescu (2018) hun kunstmatige morele adviseur en zit in de tussencategorie tussen sterke en zwakke AI.

systeem. Een voorbeeld van zo'n criterium is een morele waarde, zoals eerlijkheid of duurzaamheid. Onze voorkeuren kan het systeem natuurlijk ook kennen via eerder vermelde vormen van lifelogging. De app (AI-gesprekspartner) kan die gegevens namelijk uitwisselen met de data die wearables en apps en andere slimme IoT-apparatuur van ons registreren. Dan moeten we de criteria niet zelf invoeren in het systeem: de algoritmen leiden onze morele voorkeuren af uit onze gesprekken en handelingen die geregistreerd worden. De app kan de gebruiker zo zeer goed kennen en inschatten. Op basis van al die informatie kan het systeem 'meedenken', beslissingen eventueel toetsen aan wetgeving, doorzoeken wat in andere vergelijkbare situaties werd aangeraden, enzovoort. Bovendien kunnen we de AI-gesprekspartner ook vragen om verschillende scenario's te berekenen, om te bekijken wat er telkens kan gebeuren.

Klincewicz (2016) presenteert het idee van een 'artificial moral reasoning engine'. Indien goed ontworpen, zou een dergelijk systeem gebruikers door middel van argumenten kunnen overtuigen om wel of niet op een bepaalde manier te handelen. De AI-tool zou worden getraind met machine learning om moreel te redeneren. Het kan normatieve suggesties geven over hoe je je in specifieke situaties moet gedragen. Het zou zelfs vragen als 'moet ik dit melden bij de autoriteiten?' kunnen beantwoorden en daar argumenten voor kunnen geven en gaat zo dus verder dan het idee van Lara en Deckers. Ook al heeft dit type AMA geen (zelf)bewustzijn en valt het om die reden onder type 3, het kan toch mensen overtuigen en normatieve adviezen onderbouwen en rechtvaardigen.

Als het systeem dat allemaal kan, in hoeverre dreigt dan morele luiheid in die zin dat we beslissingen systematisch delegeren omdat we ervan uitgaan dat zo'n AMA en een grotere dataset steevast tot betere beslissingen leiden? In hoeverre leiden meer data altijd tot betere morele oordelen en betere argumenten? En welke data wel en welke niet? Is het bijvoorbeeld relevant om ook te gaan vergelijken wat andere mensen deden in vergelijkbare situaties? In hoeverre moet hier rekening mee gehouden worden? Het gedrag van de meerderheid zegt immers niet noodzakelijk iets over hoe wenselijk het gedrag *moreel gezien* is.

Een andere valkuil is dat de mens steeds de eigen toetssteen blijft. Het blijft aan gebruikers om te bepalen: vind ik dit wel of geen wenselijke beslissing? Je kan het advies van de AI-gesprekspartner immers altijd naast je neerleggen.

Deze casus brengt technische, conceptuele en ethische uitdagingen met zich mee. AI heeft, in ieder geval op dit moment, een ondermaats begrip van

context en ook taal vormt nog een zeer grote uitdaging. Maar bij fundamentele doorbraken op het vlak van taal, begrip en argumentatieleer door AI zou deze casus binnen bepaalde contexten wenselijk kunnen zijn. De AI-gesprekspartner zou de kwaliteit van morele oordelen en argumentatie van mensen kunnen verbeteren en bovendien 24/7 beschikbaar zijn voor een gesprek en feedback.

Een volgende uitdaging is dan dat ethiek verwijst naar een specifieke filosofische discipline die een systematische en kritische reflectie op moraliteit beoogt. AI zou dan eerst dit niveau moeten bereiken.

2.4 Casus 4: De robotopvoeder

Het vierde type AMA is de volledig ethische actor en een vorm van sterke AI met zelfbewustzijn. Net zoals de robots of synthetische mensen in *Machines Like Me*, zou het hier gaan om een AMA met consistent moreel kompas, die bovendien toegang tot alle beschikbare informatie ter wereld heeft, inclusief wetgeving. AMA-4 begrijpt taal (en de context ervan), heeft een diepgaand inzicht in moraliteit en is autonoom en 'wilsbekwaam' op een vergelijkbaar niveau als dat van mensen. Rachels (2007, p. 14) omschrijft de gewetensvolle morele actor als iemand die onpartijdig betrokken is bij de belangen van iedereen die wordt geraakt door wat hij of zij doet. Iemand die zorgvuldig de feiten uitzoekt en hun implicaties bestudeert. De gewetensvolle morele actor is bereid om naar de rede te luisteren, ook als dit betekent dat eerdere oordelen misschien moeten worden herzien. En, tot slot, de gewetensvolle morele actor handelt consistent volgens de resultaten van dit overleg en onderzoek. Deze AMA zou een gewetensvolle kunstmatige morele actor zijn die rechtlijniger kan handelen dan de mens. Bovendien kan die door camera- en geluidsregistratie en sensoren voortdurend extra informatie uit de context opnemen en informatie uitwisselen met andere vormen van AMA-4.

In hoeverre zouden we zo'n AMA (type 4) kunnen inzetten in een zorgcontext? De ethische implicaties van zorgrobots, bijvoorbeeld in de context van ouderen- en kinderzorg, krijgen veel aandacht in de academische literatuur (zie bv. Sharkey & Sharkey, 2010; 2014; Coeckelbergh, 2016; Sharkey, 2016; 2017; Royakkers, 2021).[23] Die artikelen gaan meestal niet over een hoogontwikkelde

23 Voorbeelden van relevante ethische vragen bij gebruik van een AMA-4 in de gezondheidszorg zijn: Wat betekent 'in het belang van de patiënt' in verschillende zorgcontexten en hoe moeten we AMA's daarin opleiden? Mag de AMA meebewegen met de interesses van de patiënt of moet die aan strikte deontologische voorwaarden voldoen?

vorm van AMA zoals in deze casus. Films bieden wel een aantal interessante gedachte-experimenten. In de fictiefilm *Robot & Frank*, bijvoorbeeld, krijgt Frank, een beginnend dementerende man, een zorgrobot in huis. De robot heeft veel talenten, zoals het bereiden van verfijnde maaltijden, zodat Frank goed en evenwichtig blijft eten. Frank is echter een dief die er nog regelmatig op uit gaat om in te breken. Dit leidt tot een interessante ethische kwestie, omdat de robot mee op pad gaat. Bovendien beschikt de robot over handige eigenschappen om het inbreken vlotter te doen verlopen. De robot overtreedt hier weliswaar ethische en juridische grenzen, maar in het kader van een zorgcontext 'beslist' die dat het beter is om de patiënt bij te staan in zijn dagelijkse activiteiten en hem daarbij te ondersteunen. Bovendien is het creëren van een goede, vriendschappelijke verstandhouding ook in het belang van de patiënt en diens levenskwaliteit. Mocht er iets mislopen, dan is de zorgrobot er steeds bij om in te grijpen.

Het is hoogst speculatief of deze AMA er ooit komt. Maar stel dat het kan, en dat de AMA moreel superieur is aan de mens: kunnen, moeten of mogen we onze kinderen er dan aan toevertrouwen? Het opvoeden, inclusief de morele socialisatie van het kind, gebeurt dan allicht consistenter. De robotopvoeder zou in een bepaald opzicht een verhoogde vorm van veiligheid en bescherming kunnen bieden, omdat die vol met camera's en sensoren hangt. Een oogje in het zeil houden, is dus geen probleem. Bovendien hoeft een robot niet te slapen, dus kan die, als het wenselijk is, constant toezicht bieden. Een verhaaltje voorlezen voor het slapengaan, is ook geen probleem: de robot heeft toegang tot de hele wereldliteratuur. Bovendien kan de robot ook steeds alle artikelen en boeken over opvoeding scannen. Deze AMA biedt verder diverse leerfuncties, zoals woordjes leren in verschillende talen. Als het kind onrustig is, kan de robot ook rustgevende muziek afspelen. En omdat de robot steeds onmiddellijk toegang heeft tot het hele internet, kan die in theorie steeds een pasklaar antwoord geven op de vele waarom-vragen van kinderen. Een aantal bezorgdheden die we bij menselijke opvoeders kunnen hebben, zoals misbruik of mishandeling, speelt hier allicht minder.

Reflectie

Aan welke eisen en capaciteiten moet een robotopvoeder voldoen? De technologie zou om te beginnen véél verder ontwikkeld moeten zijn op het vlak van taal- en emotieherkenning en fijne motoriek. Dat laatste is belangrijk om voor

eten te kunnen zorgen (of dat in ieder geval op te kunnen warmen) en kinderen te wassen en verschonen.

De robot moet ook in verschillende contexten kunnen inschatten wat goed is voor het kind. Stel dat het kind hard huilt, omdat het geen spuitje wil of omdat het geen groenten wil eten, dan mag de AMA niet zomaar toegeven, tenminste niet als het als wenselijk wordt gezien om de inenting te krijgen en omdat het gezonder is om wel groenten te eten. Maar de robotopvoeder kan zich niet verplaatsen in wat het betekent en hoe het voelt om een spuit te krijgen. De robot kan zich niet beroepen op eigen ervaringen en hoeft zelf niet te eten. In een andere context kan het juist wel goed zijn om toe te geven als het kind hard huilt, bijvoorbeeld wanneer het kind hevige pijn heeft (die de robot weliswaar nooit zelf zal kunnen voelen). De robot zal hoe dan ook een diepgaand inzicht in verschillende contexten moeten ontwikkelen.

Aangezien de robot (mede)verantwoordelijk zou zijn voor een kwetsbaar kind, moet er een hoge mate van vertrouwen zijn. De beslissingen die de robot (of diens algoritmen) neemt, moeten bovendien transparant zijn, zodat er eventueel ingegrepen kan worden als het misgaat met de technologie.

Een andere belangrijke vraag is: welk normatief-ethisch perspectief is het meest wenselijk voor een robotopvoeder? Het betreft immers een volledig ethische actor. Om het welzijn van het kind steeds centraal te stellen, biedt consequentialisme een interessante mogelijkheid. Maar wiens invulling van consequentialisme krijgt de bovenhand: die van Jeremy Bentham of van John Stuart Mill?[24] Mogelijk is de focus van Mill op kwaliteit en hogere vormen van genot (zoals cognitieve, intellectuele ontwikkeling) meer geschikt in een opvoedingscontext dan de insteek van Bentham, waarin verschillende vormen van genot niet hiërarchisch worden ingedeeld. Maar als er bij deze toepassing van de AMA-4 een groot risico is voor het individuele kind, dan is het misschien beter om de AMA te trainen volgens een striktere Kantiaanse deontologie. De robotopvoeder kan dan op basis van de categorische imperatief[25] opvoedkundige beslissingen maken.

24 Bentham en Mill worden beschouwd als grondleggers van het consequentialisme/utilitarisme.

25 In zijn moraalfilosofie onderscheidt Immanuel Kant hypothetische imperatieven van de categorische imperatief. Hypothetische imperatieven zijn algemeen en schrijven voor hoe iemand moet handelen. Het gebod geldt echter enkel voor degene die het doel heeft gesteld. Een voorbeeld: wie advocaat wil worden, moet rechten studeren. De hypothetische imperatief heeft, in tegenstelling tot de categorische imperatief, geen betrekking op het morele domein. Ethiek gaat immers niet om wat je wilt doen, maar om wat je moét doen. De categorische imperatief is algemeen en onvoorwaardelijk. De categorische imperatief kan op

In hoeverre zou een consistente opvoeding aan de hand van de categorische imperatief leiden tot moreel consistentere kinderen? Het kind zal misschien minder makkelijk wegkomen met 'leugentjes om bestwil'. Een goede robotopvoeder zal ook moeten kunnen vergeven. De AMA zou getraind kunnen worden met *Over pedagogiek*, waarin Kant zijn ideeën over opvoeding uiteenzet. Het doel van opvoeden is voor hem kinderen begeleiden tot zelfverantwoordelijke *wereldburgers*, vol zelfvertrouwen als basis van persoonlijk geluk.

Deze casus roept veel ethische vragen op. Hieronder bied ik een selectie. Om te beginnen speelt privacy opnieuw een grote rol: een robotopvoeder zou de hele ontwikkeling van het kind in data, waaronder video's, kunnen vastleggen. Wie heeft toegang tot die data? En krijgt het kind, vanaf de volwassen leeftijd, het recht om te beslissen over wat er met die data moet gebeuren?

En in hoeverre is iemand met de beste kennis over opvoeding ook in de praktijk de beste opvoeder? Kan een robot, die zelf geen geluk of emoties kan ervaren, een kind goed opvoeden? Zorg draait immers niet in de eerste plaats om 'efficiëntie' maar om betrokkenheid en warmte.

Een daaraan gerelateerd ethisch vraagstuk draait om hechting en het belang van begrip en wederkerigheid: het kind zou empathie en betrokkenheid kunnen voelen, maar die zal niet (werkelijk) wederkerig zijn. Dit raakt aan misleiding. Het kind zou antropomorfe eigenschappen op de robot kunnen projecteren die die niet (werkelijk) heeft. Dat zien we vandaag de dag al: studies toonden al vaker aan dat kinderen allerlei menselijke eigenschappen projecteren op robots. Het kind zou zich hechten aan een artefact dat zelf geen emotionele band kan opbouwen. Dat zou een hypotheek kunnen leggen op de morele, psychologische en emotionele ontwikkeling van het kind, en het kind kunnen schaden. Moraliteit en zorg gaan uiteindelijk ook om kunnen meeleven met anderen. Er is een kwetsbaarheid die robots nooit zullen begrijpen, zoals lichamelijke pijn.

De AMA zou verder in staat moeten zijn om geduld, begrip en vergeving naar opgroeiende kinderen te tonen. Consistentie kan hierbij juist een obstakel zijn, om dit chaotische proces van vallen en opstaan te begeleiden. Het onfeilbare geheugen van de AMA kan ook een belemmering voor het kind betekenen. Veel

verschillende manieren geformuleerd worden. Bijvoorbeeld dat je mensen steeds als doel en nooit louter als middel mag behandelen en dat je steeds moet handelen volgens de morele regel die een universele wet kan zijn. De regel 'liegen mag' zou bijvoorbeeld geen universele wet kunnen zijn.

academici hebben al hun zorgen geuit over het meedogenloze geheugen van internettechnologie (zie bv. Allen, 2008; Mayer-Schönberger, 2009; Gabriels & Lanzing, 2020). De geschiedenis evolueerde recent van een 'vuilnisbak' naar een 'diepvriezer' waar gegevens spreekwoordelijk 'vers' gehouden worden. Bovendien weet je niet op welke servers er data, foto's of filmpjes van jou worden bijgehouden die later 'ontdooid' kunnen worden. Voorbeelden daarvan zijn naaktfoto's en wraakporno, die jaren later nog online kunnen opduiken. Onderzoekers (zie bv. Dodge & Kitchin, 2007) pleiten ervoor om een 'ethiek van vergeten' in het ontwerp te programmeren, door het feilloze geheugen van technologie opzettelijk minder perfect te maken of door gebruikers in elk geval die optie te geven. Dat kan door bepaalde gegevens te wissen, wazig te maken of door ruis toe te voegen aan de data. Zo een opzettelijke vorm van 'geheugenverlies', een soort van 'recht op ruis', kan mensen bevrijden van een last, door hen meer van het verleden los te weken. Friedrich Nietzsche schreef al hoe de mens "niet kan leren vergeten en altijd maar aan het verleden vastzit. Hoe ver, hoe snel hij ook loopt, de keten loopt mee" (1983/2006, p. 11). Hoe ouder je wordt, hoe zwaarder de bagage kan wegen. Maar als je niet vergeet, "is het echter onmogelijk ook maar te leven", stelt Nietzsche (1983/2006, p. 13). (Mede) opgevoed worden door een entiteit met een onfeilbaar geheugen kan onwenselijk zijn en zou de ontwikkeling van het kind kunnen tegenwerken. Ook kinderen hebben recht op privacy; in het Verdrag inzake de Rechten van het Kind werd privacy opgenomen. In dat verband kan ook de vraag worden gesteld hoe een robotopvoeder zich verhoudt tot het recht op een open toekomst van een kind.

Als deze vorm van sterke AI (AMA-4) mogelijk zou zijn, dan betekent dit zonder meer een disruptief kantelpunt. In dat geval krijgt de mens een entiteit naast zich die beter zou kunnen zijn in morele oordeels- en besluitvorming. Zal de mens dit aanvaarden? Hoe zal die samenwerking verlopen? In deze AMA en casus komen de voorgaande casussen samen: de robotopvoeder kan voortdurend data van het kind verzamelen en in realtime analyseren en fouten en zwaktes in morele argumentatie opmerken en bijsturen. Dat betekent ook dat de technische, conceptuele en ethische uitdagingen optellen naarmate we opklimmen van AMA-1 (de ethische impact actor) naar AMA-4 (de volledig ethische actor).

De robotopvoeder zou naar mijn idee alleen wenselijk zijn als die het beter zou doen en systematisch het welzijn van anderen en de maatschappij zou vergroten. Maar er zijn redenen om aan te nemen dat dat niet het geval zou zijn. Een

niet-mens die een mens moet grootbrengen en begeleiden naar volwassenheid, zou de-humaniserend zijn en bovendien een doorgedreven vorm van rationalisering en technologisch paternalisme (Royakkers, 2021, p. 8).

Tot slot, afgelopen april (2021) presenteerde de Europese Commissie haar eerste voorstel tot een Europese AI-wetgeving. Daarin komt duidelijk naar voren dat hoe groter het risico is bij de AI-toepassing, hoe strenger de regels moeten zijn. Ook al ben ik in dit preadvies primair geïnteresseerd in de ethische vraagstelling, toch lijkt het evident om ook juridisch zeer strenge voorwaarden aan dit type AMA te stellen, omdat het hier niet louter om een AI-tool of app gaat die data aanlevert of ons adviseert, maar om een synthetische mens die een zorgtaak op zich kan nemen (gesteld dat dat mogelijk zou zijn).

3 Hoe kunnen we een kunstmatige morele actor programmeren?

3.1 Inleiding

Als AI ons kan helpen met morele oordeelsvorming, moeten we de systemen opleiden. Maar er is niet zomaar een blauwdruk – als in: een onfeilbare mens – voorhanden waar we een kunstmatige morele actor (AMA) eenvoudigweg mee kunnen trainen. Er is evenmin consensus over welke ethische theorie de basis zou moeten vormen voor de benodigde ethische standaarden.

Hoewel op zich niet nieuw, is voor het programmeren van moraliteit in AI zeer relevant dat morele theorieën nogal sterk verschillen in hun focus. Neem de drie normatieve theorieën consequentialisme, deontologie en deugdenethiek. Bij consequentialisme wordt de juistheid van een handeling volledig bepaald door welke handeling de beste gevolgen heeft. In het utilisme – een vorm van consequentialisme – geldt bijvoorbeeld dat een handeling moreel juist is (en verplicht) als deze het geluk van zoveel mogelijk mensen verhoogt en het lijden van zoveel mogelijk mensen beperkt. Bij deontologie ligt de juistheid van de daad, of de ethische evaluatie ervan, in het opvolgen van een morele regel (bijvoorbeeld 'niet liegen'). Bij deugdenethiek ligt de klemtoon op het karakter van de handelende persoon: die moet zo handelen opdat die zijn/haar goede karakter niet schaadt. Deugdenethiek beschrijft welke deugden of goede karaktereigenschappen je moet ontwikkelen om moreel te zijn. Deugden zijn kwaliteiten om het goede leven te bereiken, en je kan ze elke dag oefenen en verbeteren in concrete situaties. Een goed of deugdelijk karakter komt dus niet

vanzelf: je moet je daarin oefenen, jezelf bijsturen en je mogelijkheden benutten. Een deugdzaam karakter zou vervolgens tot moreel juist gedrag leiden.

De theorie, en specifieker, de definitie van moraliteit die je aanhoudt en waar je AI mee traint, zal mede de output bepalen, in dit geval het morele oordeel. Moet het systeem dan vooral gehoorzamen aan morele regels of moet het zo adviseren en/of handelen dat het geluk van het grootste aantal mensen wordt vergroot? Of beide? Wellicht is een combinatie van ethische theorieën mogelijk, maar wat als die combinatie tot tegengestelde uitkomsten leidt? Welk oordeel en welke handeling heeft dan de voorkeur?

In dit hoofdstuk staat de vraag centraal hoe we AMA's moraliteit kunnen aanleren, om ervoor te zorgen dat ze zich ethisch wenselijk 'gedragen' (gebaseerd op Allen, Varner, & Zinser, 2000; Allen, Smit, & Wallach, 2005; Wallach & Allen, 2009; Allen & Wallach, 2014; Leben, 2019; Gabriels, 2019; 2020; 2021; Cervantes et al., 2020)? Ik bespreek drie manieren om een AMA te programmeren: top-down, bottom-up en hybride.

3.2 Top-down programmeren

Een eerste manier om een kunstmatige morele actor te programmeren is top-down, met een bestaande morele regel of normatief-ethische theorie. Laten we een simpel voorbeeld nemen, ter illustratie. Zoals de eeuwenoude gulden regel, namelijk 'wat gij niet wilt dat u geschiedt, doe dat ook een ander niet'. Hoewel, simpel...? Een voor mensen relatief eenvoudige morele regel als deze is namelijk allesbehalve eenvoudig over te brengen aan een AMA. Het AI-systeem zou dan namelijk in staat moeten zijn om te bepalen wat het zelf wel en niet wil, in diverse hypothetische contexten. Het moet op zijn minst het vermogen hebben om de gevolgen van de eigen handelingen op anderen te bepalen (berekenen), om in te schatten in hoeverre het zou wensen om zelf door anderen zo behandeld te worden. En daarbij moet het ook nog rekening houden met verschillen in individuele opvattingen en voorkeuren.

Laten we eens kijken tegen welke vragen we aanlopen als we een aantal bekende ethische benaderingen top-down in AI zouden willen programmeren.

Deontologie

Zoals eerder vermeld ligt bij deontologie de morele juistheid van een handeling in het opvolgen van een morele regel. Voor Immanuel Kant, die beschouwd

wordt als grondlegger van deontologie, is morele evaluatie primair gericht op de wil of de intentie. Voor Kant staat handelen uit goede wil (de intentie achter de handeling) gelijk met handelen uit plicht. Dus moraliteit ligt in de *intentie* om het intrinsiek goede te *willen*, met de goede wil als basis. Een AMA die een intentie kan (her)kennen is dan onvoldoende om van een morele handeling te kunnen spreken: de AMA, zoals de robotopvoeder (AMA-4), moet ook de juiste intentie hebben. Voor moreel juist gedrag moet de robotopvoeder niet alleen (kunnen) handelen in overeenstemming met de plicht, maar ook *omwille* van de plicht. Mogelijk kan een minder complexe AMA (AMA-3) een zinvolle rol spelen in morele oordeelsvorming, door te toetsen of een handeling aan de eisen voldoet die deontologie daaraan stelt, en de mens daarover te adviseren. Maar een volledig ethische actor (AMA-4), geprogrammeerd met (Kantiaanse) deontologie, is buitengewoon complex.

Consequentialisme

Misschien is consequentialisme eenvoudiger. Neem 'Jeremy'. Jeremy is een consequentialistisch AI-systeem, vernoemd naar Jeremy Bentham. Jeremy gebruikt de zogeheten 'hedonistische calculus' van Bentham om morele dilemma's op te lossen (Cervantes et al., 2020). In deze benadering is *die* handeling moreel juist (en verplicht) die waarschijnlijk het grootste netto-plezier of geluk oplevert, rekening houdend met alle betrokkenen. Welke handeling juist is, wordt bepaald door te kijken naar: wie allemaal betrokken zijn, wat voor élke persoon de intensiteit van het plezier of het lijden is,[26] en hoe lang het plezier of het lijden duurt. Ook wordt voor elke mogelijke handeling bepaald hoe *waarschijnlijk* het is dat dit plezier of lijden zal optreden. Voor elke betrokkene en per mogelijke handeling berekent het algoritme de intensiteit, de duur en de kans op het verkrijgen van netto-plezier. Dat wordt per handeling bij elkaar opgeteld, en daar rolt ten slotte uit welke handeling het meeste totale netto-plezier oplevert.[27] Dit blijft een kansberekening, maar consequentialisme leent zich er op zich goed voor om op al deze zaken een getal (numerieke waarde) te plakken.

Jeremy is weinig autonoom: het is veeleer een rekenmachine, want mensen moeten die gegevens (input) invoeren. Meer geavanceerde systemen bestaan

26 Op een schaal van 2 tot -2.

27 Als twee of meer handelingen waarschijnlijk het grootste netto-plezier opleveren, beschouwt Jeremy deze handelingen moreel gezien als even juist.

nog niet en botsen al snel op het volgende probleem: om consequentialisme in een machine te programmeren, moeten de effecten van een handeling op elk lid van de morele gemeenschap een numerieke waarde krijgen. Het is echter onmogelijk om dit in realtime voor elke handeling in de wereld te doen, vooral omdat de gevolgen van elke handeling weer tot nieuwe gevolgen leiden. Op zich kan je dit rekenprobleem beperken door een bovengrens in te stellen, waarboven geen verdere inschatting van gevolgen meer nodig is. Ook dat is echter onwaarschijnlijk complex en bovendien kan er net boven die grens misschien ongelooflijk veel lijden en pijn veroorzaakt worden, wat we toch als een moreel verwerpelijke daad zouden beschouwen.

Ook al kan en hoeft de AMA wellicht niet gedetailleerd *alle* mogelijke gevolgen te voorspellen, het moet wel zelf een inschatting kunnen maken van een grote hoeveelheid gevolgen op alle betrokkenen. Daar heeft de AI een diep inzicht voor nodig, waarvan het de vraag is of dit realistisch is. Een kritische tegenvraag is natuurlijk dat menselijke actoren al die effecten ook niet gedetailleerd kunnen inschatten. Is het dan redelijk om dit van een AMA te eisen?

Confucianisme

Er zijn ook prototypes die getraind worden met confucianisme. De virtuele Confucius is een interactieve chatbot die gebruikers laat kennismaken met de filosofie van Confucius (Cheok, Edirisinghe, & Karunanayaka, 2017), waarin menselijke relaties worden gezien als het fundament van de samenleving. 'Confucius' is een spel dat filosofische antwoorden geeft op vragen van gebruikers. Wat Confucius zou adviseren, stemt niet automatisch overeen met het advies van Jeremy, ook omdat in het Confucianisme de klemtoon sterker ligt op de mens als relationeel wezen.

Een mix van morele benaderingen

Andere prototypes maken gebruik van een mix van normatief-ethische theorieën. MoralDM, bijvoorbeeld, combineert deontologie met consequentialisme (zie Cervantes et al., 2020). In eenvoudig Engels krijgt het systeem een moreel dilemma voorgelegd. Eerst zoekt het uit of er morele regels bij betrokken zijn waar het zeker rekening mee moet houden, zoals het verbod op liegen. Als MoralDM vaststelt dat zulke regels moeten gelden, dan opereert het in de deontologische modus en geeft het de voorkeur aan passiviteit boven actie. Als dat niet het geval is, verkiest het AI-systeem een consequentialistische actie die de beste uitkomsten garan-

deert. Deze volgorde geeft aan dat MoralDM in de kern een deontologisch systeem is. AMA's trainen met verschillende ethische theorieën, in plaats van op voorhand één theorie te kiezen, zou een interessante manier kunnen zijn om te onderzoeken op welke manier de algoritmen leren, hoe ze verschillende morele scenario's kunnen 'doordenken' en tegen welke problemen ze vervolgens aanlopen.

3.3 Bottom-up programmeren

Er wordt ook geëxperimenteerd met bottom-up benaderingen om kunstmatige morele actoren te programmeren. Daarbij wordt de AMA gemodelleerd naar moreel gedrag van mensen. Je vertrekt hier dus niet vanuit een specifieke ethische theorie, maar vanuit een omgeving (context) waarin de AMA gedrag aanleert. Een voorbeeld is een associatief leerproces, waarbij het systeem leert uit interacties met de omgeving. Dat is vergelijkbaar met hoe kinderen moreel opgevoed worden, door bijvoorbeeld ongepast gedrag te bestraffen, zonder daarbij een expliciete theorie mee te geven of aan te houden (Allen, Smit, & Wallach, 2005, p. 151).

Mensen kunnen in dagelijkse interacties van elkaar leren en hun gedrag en oordelen vervolgens aanpassen. Specifiek voor kinderen is het gezin een belangrijke setting voor morele socialisatie (Ponzetti, 2010). Ouders geven het morele redeneren en de morele oordelen van hun kind mede vorm (Brody & Shaffer, 1982; Walker & Hennig, 1999). Ouder-kind interacties staan bol van impliciete en expliciete morele regels en instructies over gepast gedrag, waarden, normen of etiquette (Ochs & Kremer-Sadlik, 2007). Op basis van deze ervaringen leren en internaliseren kinderen morele concepten (Smetana, 1999).

Programmeurs, eventueel bijgestaan door ethici, zouden AI op een soortgelijke manier kunnen opvoeden. Een interessant experiment in de (verre) toekomst zou kunnen zijn om de AMA letterlijk in een gezin te plaatsen en die zo moreel te socialiseren. Dus in plaats van het gedachte-experiment met de robotopvoeder (casus 4 uit het vorige hoofdstuk) die een kind opvoedt, zou het kind samen met een robot in huis opgroeien en opgevoed worden door dezelfde ouders. Op die manier leert de robot in een complexe menselijke omgeving, en wordt de AI ook blootgesteld aan allerlei dagelijkse morele situaties en interacties.

Manieren om bottom-up leren te stimuleren

Dat bottom-up moreel leren door AI kan in theorie op verschillende manieren wor-

den gestimuleerd. 'Reinforcement learning' is bijvoorbeeld een AI-techniek die het mogelijk maakt om de AMA te trainen met een signaal dat kan variëren van sterk positief ('beloning') tot sterk negatief ('bestraffing'). De AMA moet vervolgens zelf leren welke beslissing, uit een mogelijk lange keten van beslissingen, verantwoordelijk was voor deze feedback en versterkt moet worden ('reinforcement') of juist afgeremd (Boden, 2016). Net zoals je bij het spelen van videogames een score krijgt die soms afhangt van beslissingen die je veel eerder in het spel nam.

Reinforcement learning brengt ook problemen met zich mee. Je moet het systeem, dat volop leert, op een kwaliteitsvolle manier feedback kunnen geven. Wat dat betreft zou een binair systeem – 'aanvaardbaar versus onaanvaardbaar' – tekortschieten. Dat probleem zou wellicht opgelost kunnen worden door gradaties van beloning en bestraffing toe te passen. Maar, belangrijker, goede morele training geeft ook context mee: *waarom* iets goed of fout is. Alleen maar aangeven *dat* iets goed of fout is, volstaat daarom niet. Morele ontwikkeling is een leerproces waarbij je opvattingen toetst aan gebeurtenissen en omstandigheden. Het gaat uiteindelijk om het internaliseren van regels en niet om ze simpelweg uit te voeren, bijvoorbeeld omdat je een bestraffing vreest als je de regel niet volgt, zonder dat je weet waar de regel voor staat. De AMA zou slechts een primitieve vorm van moraliteit bevatten, zolang die geen vorm van zelfbewustzijn (AMA-4) heeft.

Software die momenteel wordt ontwikkeld voor morele besluitvorming door een AMA kan gebruik maken van een waaier aan strategieën, waaronder *case-based* redenering of casuïstiek. Daarbij worden morele casestudies telkens apart behandeld en geëvalueerd zodat het systeem leert welke specifieke 'oplossing' of regel een concrete situatie vereist. Ook dit valt onder de bottom-up benadering.

Een laatste voorbeeld is het simuleren van onze morele evolutie. Dat staat centraal binnen *ALife*, een afkorting van Artificial Life. In een gesimuleerde omgeving leert de AMA aan de hand van kennis over sociobiologie en spelltheorie, zoals het bekende *prisoners' dilemma* (zie bijvoorbeeld Danielson, 1992). Het eerder vermelde Go is een zogeheten 'zero-sum game' (nulsomspel). Daarbij wint de ene partij evenveel als dat de andere partij verliest, waardoor de netto-opbrengst van het spel nul is. Maar er zijn ook spellen die sterker inzetten op samenwerking en onderhandeling met medespelers, zoals het bordspel *Diplomacy*, dat met maximaal 7 partijen gespeeld kan worden. In een zogeheten multi-agent omgeving leren algoritmen dan Diplomacy spelen, met behulp van reinforcement learning. Herhaaldelijke testen laten zien dat de kunstmatige acto-

ren steeds beter presteren (Anthony et al., 2020).

Menselijke moraliteit evolueerde in de praktijk natuurlijk volgens scenario's die véél complexer zijn dan deze (simplistische) spellen en simulaties in kunstmatige omgevingen. Bovendien bots je hier op soortgelijke problemen als bij associatieve leerprocessen: een *ontwikkelde* morele actor moet in staat zijn om moraliteit op een abstract en theoretisch niveau te begrijpen en erop te reflecteren, dus hier gaat het nog steeds om een erg basale vorm van moraliteit.

3.4 Hybride programmeren

Een andere optie om moraliteit in een AMA te programmeren combineert de top-down en bottom-upbenadering. Dat maakt deze benadering hybride.

Er zijn al prototypes van AI-adviseurs binnen de medische ethiek, die hybride geprogrammeerd zijn. Een voorbeeld is MedEthEx (zie Anderson, Anderson, & Armen, 2005; 2006; Cervantes et al., 2020, pp. 522-523). Het systeem is getraind door bio-ethici en bestaat uit drie componenten. Allereerst, een interface die de plichten helpt selecteren die relevant zijn in een casus. Ten tweede, een adviseursmodule die de juiste *handeling* per casus selecteert op basis van aangeleerde kennis. En ten slotte een leermodule die relevante principes abstraheert uit de specifieke casussen.

MedEthEx wordt getraind met een combinatie van casuïstiek (bottom-up) en een theorie van plichten (top-down geprogrammeerd). Bij MedEthEx wordt gebruik gemaakt van de theorie van Beauchamp en Childress om biomedische ethische dilemma's op te lossen. Deze omvat het principe van respect voor autonomie (*respect for autonomy*), het principe van niet-schaden (*nonmaleficence*), het principe van weldoen (*beneficence)* en het principe van rechtvaardigheid (*justice*).[28] Dit zijn zogeheten *prima facie* plichten. Dat zijn verplichtingen waaraan mensen zouden moeten proberen te voldoen, maar die bij gelegenheid kunnen worden overschreven door een andere, sterkere plicht.

Hoe werkt dat concreet? MedEthEx berekent en kent waarden toe aan de vier principes, en geeft daarmee aan in hoeverre een handeling aan deze principes voldoet, of deze juist schendt. Die waarde varieert van 2 tot – 2. Als een principe niet van toepassing wordt geacht bij een specifieke handeling, krijgt deze de waarde 0. Als een handeling een principe schendt, dan krijgt het principe voor die

28 Vanwege de centrale rol van plichten, wordt deze benadering ook wel een vorm van 'principlisme' genoemd.

handeling een negatieve waarde. -1 is dan een kleinere schending dan -2. Als een handeling aan een principe voldoet, dan krijgt het principe voor die handeling een positieve score. +2 betekent dat een handeling meer aan het principe voldoet dan bij een score van +1.

Ik geef een voorbeeld (gebaseerd op Cervantes et al., 2020, p. 523). Stel dat een patiënt een injectie weigert op basis van een irrationele angst. Een arts heeft dan twee handelingsopties: hij of zij kan proberen om de patiënt van gedachten te laten veranderen of de arts kan de keuze van de patiënt direct accepteren. Voor elke handelingsoptie moeten handmatig getallen worden ingevoerd in MedEthEx, om aan te geven hoe de handelingsoptie scoort op de relevante principes. De handelingsoptie waarbij de arts de patiënt van gedachten probeert te laten veranderen, zou bijvoorbeeld als volgt kunnen scoren: respect voor autonomie: -1; niet-schaden: +2; weldoen: +2. Voor de handelingsoptie waarbij de arts de keuze van de patiënt direct accepteert, moeten opnieuw cijfers worden ingevoerd. Bijvoorbeeld respect voor autonomie: +1; niet-schaden: -2; weldoen: -2. Volgens het model is het beste ethische advies dan om te proberen om de patiënt van mening te doen veranderen, omdat die handeling hogere pluswaarden heeft. Net zoals bij Jeremy impliceert dit dat je een numerieke waarde moet geven aan elke parameter om vervolgens de meest wenselijke uitkomst te berekenen.

Een ander voorbeeld van een hybride benadering zou zijn om deugden rechtstreeks in de machine te programmeren (top-down) en van daaruit de machine verder te laten leren (bottom-up). Om te onderzoeken of een AMA de deugd eerlijkheid bevat, heb je dan een algoritme nodig dat eenduidig bepaalt of handelingen eerlijk worden uitgevoerd. Uitdaging is dat deugden niet eenvoudig te vertalen zijn naar concrete 'regels'. Een optie is om dit te combineren met de bottom-up opvoedbenadering, omdat rolmodellen een essentiële taak hebben binnen deugdenethiek. In plaats van de deugd te programmeren, leert de AMA de deugd gaandeweg aan, als een vorm van praktische wijsheid. Maar ook dit is hoogst speculatief en zonder diepgaand inzicht blijft de AMA een primitieve vorm van moraliteit hebben.

3.5 Sociaal contract

In zijn boek *Ethics for Robots* (2019) probeert ook filosoof Derek Leben een methode te ontwikkelen om moraliteit in AMA's te programmeren. Leben vat moraliteit op als samenwerking (coöperatie). Zijn theorie is gebaseerd op morele psychologie en

efficiënte strategieën uit de speltheorie. Hij gebruikt het sociale contract (contractualisme) als model om moraliteit in algoritmen te programmeren.

Het sociaal contract gaat over gedeelde afspraken tussen leden van een gemeenschap. In John Rawls zijn sociale contract moeten rationele, autonome en zelfzuchtige personen de regels van de samenleving kiezen. Daarbij weten ze niet welke rol ze zelf in de samenleving zouden spelen: ze moeten verdelende principes vaststellen achter een zogenaamde 'sluier van onwetendheid'. Volgens Rawls zal elke persoon vervolgens proberen om principes vast te stellen voor het besturen van de samenleving en de distributie van goederen die het beste voor henzelf zijn. Dit betekent dat ze zullen proberen het minimumniveau van basisgoederen, zoals rijkdom, primaire hulpbronnen, gezondheid, enzovoort, die ze zullen ontvangen, te maximaliseren.[29] Omdat mensen achter de sluier van onwetendheid niet weten in welke positie ze terecht zullen komen, zo is de gedachte, zullen ze proberen de beste worst-case situatie te creëren waarin ze terecht zouden kunnen komen. Sociale en economische ongelijkheden – verschillen tussen mensen in hoeveel geld ze hebben, hun status, macht, enzovoort – zijn in deze benadering alleen toegestaan als ze ten goede komen aan de mensen die het slechtst af zijn.[30]

Derek Leben neemt de schaakcomputer als model voor het ontwerpen van sociale contracttheorie in algoritmen. Zijn zogenaamde 'ethics engine' vertegenwoordigt 'self-interested agents' (zelfzuchtige actoren) die primaire goederen zoals gezondheid moeten verdelen. Hij bespreekt onder meer het idee dat robots weigeren sigaretten aan een klant te verkopen; in die zin zouden ze moreel superieur kunnen worden aan mensen. Zijn 'ethics engine' omvat ook morele verplichtingen jegens niet-menselijke entiteiten, zoals dieren en het milieu. De contexten waarin de 'ethics engine' mogelijk wordt gebruikt, houdt hij bewust beperkt om transparantie te garanderen. Leben stelt bijvoorbeeld expliciet dat die niet in een oorlogscontext mag worden gebruikt, omdat dit tot verregaande discussies zou leiden over wie verantwoordelijk is als de 'ethics engine' overgaat tot beslissingen waarbij mensen gedood worden.

3.6 Kort besluit

Uit de bespreking tot nu toe komt duidelijk naar voren dat we nog ver verwij-

29 Dit wordt ook wel een 'maximin' strategie genoemd.

30 Dit is het zogeheten 'Difference principle'.

derd zijn van een AMA die enigszins in de buurt komt van de complexiteit van het menselijke oordeelsvermogen. Als er al prototypes zijn, waarbij moraliteit in AI wordt geprogrammeerd, dan staan die nog in de kinderschoenen en zijn ze gebaseerd op zeer vereenvoudigde modellen. De testen vinden bovendien plaats in gecontroleerde omgevingen, die niet vergeleken kunnen worden met échte complexe, morele situaties. Veel werk ligt bij de mens, om de benodigde gegevens in de systemen in te voeren, om doelstellingen te bepalen, om feedback te geven en supervisie te bieden. Een duidelijke meerwaarde van de systemen blijft voorlopig uit. Door een beperkt begrip van taal is het voorlopig niet mogelijk om algoritmen blind te laten zoeken in alle bestaande teksten over ethiek (in alle denkbare talen) om vervolgens tot diepere inzichten te kunnen komen. Kortom, er liggen nog veel uitdagingen voor ons.

4 Advies

4.1 Inleiding

In dit hoofdstuk schets ik een aantal aanbevelingen, vragen en bedenkingen waar we rekening mee moeten houden als we AI (willen) inzetten bij morele oordeelsvorming. In de inleiding vermeldde ik al hoe complex de centrale vragen van dit preadvies zijn. Die complexiteit is zowel technisch, ethisch als conceptueel van aard. Dat kwam al naar voren in de uiteenzetting van de vier types kunstmatige morele actoren (AMA's) (hoofdstuk 1), in de bespreking van casuïstiek om de AMA's te illustreren (hoofdstuk 2), en in de reflectie op hoe we moraliteit in AMA's in zouden kunnen bouwen (hoofdstuk 3). Die complexiteit keert ook in dit hoofdstuk terug.

Per type kunstmatige morele actor verandert de rol en bijdrage van AI in morele oordeelsvorming en neemt de complexiteit toe. Bij de ethische impact actor (AMA-1) en de impliciet ethische actor (AMA-2) zijn de functies en opzet nog relatief beperkt, omdat de technologie berekeningen en analyses (modellen) aanlevert die mensen kunnen helpen om tot een meer onderbouwd oordeel te komen. Toch liepen we in de bespreking van de eerste twee casussen, geënt op AMA-1 en 2 (hoofdstuk 2), al vrij snel tegen moeilijke vragen aan. Hoewel de vragen bij AMA-1 en de besproken casus (euthanasie) niet allemaal nieuw zijn – zoals het bepalen van parameters om welzijn en lijden te meten, en het privacy-vraagstuk – er is geen eensgezindheid over. Dat werkt door als we ons een beeld proberen te vormen hoe we het gebruik van een dergelijke AMA (en alle volgende types) voor morele oordeelsvorming kunnen evalueren.

Dit gebrek aan consensus keert terug in dit hoofdstuk, dat bestaat uit vijf delen. In het eerste deel ga ik (opnieuw) dieper in op de vraag met welke data en theorieën we de verschillende types AMA moeten trainen. Die keuze, gemaakt

door mensen, is sowieso bepalend voor de output, maar is vooral cruciaal bij AMA type 3 (de expliciet ethische actor) en type 4 (de volledig ethische actor). Daar liggen de standaarden hoger, en moéten ze ook hoger liggen, gezien de verregaande impact en gevolgen die het gebruik van deze typen AMA kunnen hebben. Maar wie beslist? En moeten we op alles getallen gaan plakken of juist een aantal (morele) waarden 'vrijwaren' van kwantificering? Vervolgens bespreek ik of meer snelheid, consistentie en efficiëntie bij morele oordeelsvorming in elke context wenselijk is. Het derde deel gaat over de vraag wat het betekent als een AMA op het gebied van morele oordeelsvorming 'beter' presteert dan de mens en wat de gevolgen daarvan kunnen zijn. In het vierde deel komt de verantwoordelijkheid van de maker aan bod. Ik focus ook op de risico's op 'morele ontkoppeling' (*moral disengagement*) en morele luiheid bij oordeelsvorming door AMA's. Ten slotte ga ik in op het belang van consensus én duidelijk omschreven begrippen.

4.2 Met welke data en theorie moeten we een AMA trainen? (vervolg)

AMA's worden volop getraind met ethische theorieën (cf. hoofdstuk 3). Ik besprak eerder het probleem dat als we een AMA top-down programmeren, we ons daarmee vastleggen op een specifieke ethische theorie. Voor AMA-1 en 2 is er wellicht een praktische oplossing denkbaar voor dit probleem. We zouden er bijvoorbeeld voor kunnen kiezen om de AMA's parallel op basis van *verschillende* ethische theorieën te laten bepalen welk moreel oordeel in specifieke situaties de voorkeur heeft. We hoeven ons dan niet vast te leggen op één specifieke ethische theorie, maar verkrijgen inzicht in in hoeverre verschillende ethische theorieën in concrete situaties tot andere uitkomsten leiden, of dat er enige overlap mogelijk is. De aanbevelingen door de AMA's leveren dan mogelijk een waaier aan opties op, waartussen de mens moet kiezen. Vermoedelijk zal dat veel discussie geven. Maar de verschillende berekeningen naast elkaar kunnen misschien toch helpen om de ethische argumentatie inzichtelijker te maken, en om ervoor te zorgen dat de uiteindelijke afweging zo goed mogelijk is onderbouwd.

Mensen zijn verre van perfect, maar kunnen wel reflecteren, leren uit verkeerde beslissingen en vervolgens een diepgaander moreel begrip ontwikkelen. Van een ethicus die lid is van een ethische commissie verwachten we geen perfectie. Wel verwachten we onder andere onpartijdigheid, diepgaande reflectie, dat hij of zij zich grondig inleest en laat informeren, argumenten geeft om zijn of

haar oordeel te staven, enzovoort. In principe zou de AMA dit dan allemaal beter moeten kunnen. Of zou een AMA ook imperfect mogen zijn?

AI gaat om een optimalisatieproces, dus je ontkomt niet aan berekeningen. Niet bij AMA-4 (de volledig ethische actor), maar ook niet bij AMA-1 (de ethische impact actor). Zoals we zagen, moet daarvoor op alles een getal worden geplakt, wat bijzonder complex én delicaat is. De basisvraag is dan steeds: wat willen we precies optimaliseren? En waarom? In hoeverre zijn berekeningen eigenlijk in staat om recht te doen aan alle aspecten die volgens ethische theorieën moreel gezien relevant kunnen zijn? Zijn er waarden en morele aspecten die niet kwantificeerbaar zijn? Zo ja, moet er dan bijvoorbeeld een duidelijke richtlijn komen die deze morele waarden en aspecten uitsluit van AI-toepassingen?[31]

Bij de inzet van AI bij medische oordeelsvorming weten we vaak vooraf al wat we willen, bijvoorbeeld een ziekte opsporen, dan wel vroege tekenen of gezondheidsrisico's. Er zijn veel data beschikbaar waaruit algoritmen kunnen leren en die gegevens zijn – schat ik in – minder complex dan data over moraliteit en menselijke intuïties over wat in welke context goed of slecht is. Bij morele oordeelsvorming weten we in veel gevallen niet wat mensen willen. Ja, we weten dat zij een antwoord willen op de vraag wat moreel gezien juist of het beste is om te doen, maar dat is nog niet erg specifiek. Ook is er waarschijnlijk niet één antwoord waar iedereen zonder meer mee akkoord gaat. Hoe moeten we omgaan met de verschillende normatief-ethische benaderingen? Kunnen we ooit bepalen welke theorie de meest wenselijke is? Zo niet: wie beslist? Wiens oordeel of perspectief gaat voor?

Een interessante mogelijkheid is het zoeken naar een 'wide reflective equilibrium' tussen al die verschillende oordelen en perspectieven. Om het maatschappelijk bewustzijn over morele keuzes door zelfrijdend transport te vergroten, ontwikkelde het *Massachusetts Institute of Technology* (MIT) een website waarop morele dilemma's met een zelfrijdende auto visueel worden voorgesteld.[32] Je kan diverse scenario's beoordelen en ook zelf scenario's ontwerpen. Onderzoek over de *MIT Moral Machine* legde culturele verschillen bloot, onder meer tussen westerse en oosterse culturen (zie Awad et al., 2018). Zo bleken Japanners gemiddeld

31 In het verlengde hiervan is ook maar de vraag in hoeverre je moraliteit kunt reduceren tot een beslissingsboom, zeker als je rekening houdt met menselijke intuïties over wat goed en fout is. Het is onhaalbaar om elk mogelijk denkbaar moreel dilemma in kaart te brengen en op voorhand een rechtvaardiging voor de best verdedigbare keuze op te stellen.

32 https://www.moralmachine.net/ Geraadpleegd op 11 mei 2021.

meer geneigd om de levens van oudere mensen te sparen dan inwoners van West-Europese landen, die gemiddeld de voorkeur geven aan jongere mensen. Het onderzoek geeft inzage in het morele denken van verschillende culturen. Een voorkeur die verschillende culturen deelden, was het sparen van mensen die zich aan de wet houden in tegenstelling tot voetgangers die oversteken terwijl het verkeerslicht op rood staat. Een andere cultuuroverstijgende voorkeur is dat het leven van een mens boven het leven van een dier gaat.

Op basis van deze MIT-studie stellen Savulescu, Gyngell, & Kahane (2021) een benadering voor die zij 'Collective Reflective Equilibrium in Practice' (CREP) noemen. De voorkeuren van het brede publiek dienen dan als input voor een deliberatieproces dat zoekt naar coherentie tussen wat in concrete situaties moreel gezien de voorkeur heeft, enerzijds, en de ethische principes die in die situaties van toepassing worden geacht. Dit zou een manier kunnen zijn om rekening te houden met menselijke voorkeuren, hoewel de auteurs ook erkennen dat dit niet noodzakelijk ethisch de meest wenselijke methode is. De meest populaire of meest gedeelde voorkeur zegt immers niets over hoe moreel wenselijk die is. Met een verwijzing naar John Stuart Mill zeggen de auteurs daarover: *"We should not return to the 'tyranny of the majority' or, in contemporary terms, the 'tyranny of big data'"* (p. 12). In het geval van de MIT morele scenario's ging het bovendien om dilemma's die in het echte leven nooit zo duidelijk zijn: je moet bijvoorbeeld kiezen tussen het redden van een dokter, een CEO of een crimineel, een keuze die in verkeerssituaties zoals wij die kennen niet voorkomt.

Het is moeilijk om te voorspellen in hoeverre de voorkeuren en keuzes van mensen veranderen zodra zelfrijdende auto's deel uitmaken van hun dagelijks leven. Het 'reflectieve evenwicht' dat via de genoemde methode bereikt wordt, kan wat dat betreft tijdelijk zijn. De afgelopen decennia zagen we veel voorbeelden van 'technomoral change' (morele opvattingen die wijzigen door technologie), zoals bij in-vitrofertilisatie (ivf). Waar de ouders van de eerste 'reageerbuisbaby' Louise Brown nog overladen werden met haatreacties, zouden veel mensen het vandaag de dag ontoelaatbaar vinden om een koppel dat binnen bepaalde regels (zoals leeftijdsgrenzen) in Nederland of België in aanmerking komt voor een ivf-traject, die mogelijkheid toch te weigeren.

4.3 Zijn efficiëntie, snelheid en consistentie altijd beter?

Het is niet moeilijk om situaties te bedenken waarin onze menselijke natuur

ons in de weg kan zitten, zoals bij *in-group voorkeuren*, als we leden van de eigen groep onterecht voortrekken en niet-leden discrimineren of minder kansen geven. Onze natuur kan ons ook in de weg zitten als we binnen milliseconden een oordeel over iemand vellen, puur op basis van uiterlijke kenmerken. AI kan in principe minder beïnvloedbaar zijn, bijvoorbeeld voor vleiend gedrag.

Voorstanders van AMA's wijzen onder meer op efficiëntie, snelheid en consistentie, om aan te tonen dat AI het beter kan of zal kunnen dan de mens. In hoeverre zijn deze criteria altijd wenselijk? En zijn ze dat ook in de context van morele oordeelsvorming? Is het onder bepaalde omstandigheden niet beter om niét consistent te zijn? Er zijn contexten die een overtreding van de morele regel 'niet liegen' rechtvaardigen. Denk aan de situatie waarin ouders tegen hun kind zeggen dat hij of zij moet gaan slapen, 'want alle kinderen van de klas liggen al in bed'. En wie doodt uit zelfverdediging en dus de morele regel 'niet doden' overtreedt, krijgt een andere ethische en juridische evaluatie dan wanneer er geen sprake is van zelfverdediging. In veel gevallen zijn afwijkingen van de morele regel, zoals bij het voordringen in de vaccinatievolgorde, wel ethisch problematisch. Als we AMA's willen inzetten bij morele oordelen, dan moeten we hen aanleren wanneer een afwijking van de regel gerechtvaardigd is of kan zijn.

Soms kan ook een overtreding van de wet ethisch gerechtvaardigd zijn. Toen Rosa Parks in 1955 weigerde om in Alabama in de bus haar plaats af te staan aan een blanke passagier deed ze iets illegaals. Het was door de wet opgelegd dat mensen met een donkere huidskleur hun plaats moesten afstaan aan blanken als alle voor hen gereserveerde plaatsen bezet waren. Parks weigerde om haar plaats af te staan, en ook om vervolgens de haar opgelegde boete te betalen, een tweede illegale daad. Maar beide daden zijn wel ethisch verdedigbaar. Ze leidden ook tot belangrijke doorbraken voor de Amerikaanse burgerrechtenbeweging, aangedreven door boosheid en gevoelens van onrechtvaardigheid. In het voorbeeld is het hebben van emoties, zoals het kunnen voelen van woede, juist essentieel om de maatschappij moreel beter te maken. Dat zijn zaken die AMA's (vooralsnog) ontberen.

Ten slotte, ook snelheid en efficiëntie zijn niet noodzakelijk altijd beter. Uiteraard zijn er contexten die dit kunnen rechtvaardigen, bijvoorbeeld wanneer de tijd dringt, zoals in het voorbeeld van het scannen van elektronische patiëntendossiers. Maar ook dan moet steeds gekeken worden naar de *kwaliteit* van het oordeel of advies door de AMA. Er zijn ondertussen talloze voorbeelden bekend

van hoe AI een vooringenomen beslissing nam. Algoritmen oordeelden bijvoorbeeld dat iemand niet in aanmerking kwam voor vervroegde vrijlating omdat het risico op herhaling (recidive) te groot was. Onderzoek liet vervolgens zien dat de algoritmen een patroon puur op basis van iemands naam hadden gevolgd.[33]

4.4 Wat betekent 'beter'?

Een vraag waar we niet aan ontkomen, is: wanneer zou de AMA het onmiskenbaar beter doen dan de mens op het vlak van morele oordeelsvorming? Aan welke voorwaarden moet de AMA daarvoor tenminste voldoen? Het formuleren van betere argumenten? Het overstijgen van *in-group* voorkeuren? Het kunnen omgaan met verschillende (culturele) contexten? Iets anders?

Morele perfectie is niet mogelijk, voor de mens noch voor de AMA. Morele waarden kunnen en zullen conflicteren en dan zal ook de AMA, net zoals de mens, een afweging moeten maken om de best mogelijke (maar niet de perfecte) keuze te kunnen maken. Een AMA moet de gevolgen van handelingen op mensen en andere vormen van leven kunnen inschatten en hen niet beschadigen.

We lijken interessant genoeg soms hogere eisen te stellen aan techniek dan aan de mens. De afgelopen jaren hebben ongevallen die veroorzaakt werden door zelfrijdende auto's van Tesla en Uber wereldwijde aandacht gekregen. Veiligheid moet uiteraard een topprioriteit zijn; mogelijk kunnen zelfrijdende auto's voor meer veiligheid zorgen, bijvoorbeeld door op tijd te remmen (Luetge, 2017). In tegenstelling tot menselijke bestuurders verplaatsen zelfrijdende auto's zich niet onder invloed van alcohol of met een slaaptekort. Zelfrijdende auto's zouden volgens rapporten het aantal verkeersdoden tot 90% kunnen verminderen en het dus veel beter dan menselijke chauffeurs doen (Luetge, 2017). Het is echter onmogelijk om een 100% foutloze machine te ontwerpen en dat zullen we moeten aanvaarden. Een zelfrijdende auto kan namelijk ook een lekke band krijgen of slippen bij vriesweer. We lijken op dit moment fouten gemakkelijker van feilbare mensen te accepteren dan van feilbare machines. Dit komt goed tot uiting in een citaat van AI-wetenschappers Stuart Russell, Daniel Dewey en Max Tegmark: 'Als zelfrijdende auto's de ongeveer 40.000 jaarlijkse verkeersdoden in de VS halveren, krijgen de autofabrikanten misschien geen 20.000 bedankbriefjes, maar 20.000

33 Zie https://www.tijd.be/politiek-economie/belgie/algemeen/luc-steels-belgies-pionier-in-artificiele-intelligentie-ik-voel-een-diepe-bewondering-voor-de-mens/10004353.html Geraadpleegd op 26 mei 2021.

rechtszaken' (Russell, Dewey, & Tegmark, 2015, p. 107 (eigen vertaling)). Misschien verandert die houding in de toekomst, mochten zelfrijdende auto's meer deel uitmaken van ons dagelijks leven en straatbeeld. Waarom zouden we perfectie moeten verwachten van een machine, als die het systematisch beter zou doen dan menselijke chauffeurs (in de zin van minder ongelukken veroorzaken)?

Nu al over toekomstmuziek hebben

Als we kijken naar beschikbare prototypes zoals MoralDM en MedEthEx, dan is de meerwaarde van AMA's er nog niet, ook niet op het vlak van moreel advies. We staan technisch gezien nog ver af van AMA-3 en zeker van AMA-4. Zo moet de mens nog alle data invoeren, wat een behoorlijke investering is qua kosten en tijd. Een belangrijk kantelpunt is zodra de AMA niet alleen modellen of voorspellingen aanlevert, maar ook zelf een normatief advies kan formuleren, zoals in de casus van de (hypothetische) AI-gesprekspartner (AMA-3). Enerzijds mag dit niet zomaar leiden tot blind vertrouwen in de technologie, maar anderzijds zouden er misschien richtlijnen moeten komen dat, als de AMA moreel beter kan oordelen, de mens verplicht is om hier rekening mee te houden. Dat zou nog verschillende vormen aan kunnen nemen. Eén optie is om te vereisen dat de mens zich dan tenminste expliciet tot het oordeel van de AMA verhoudt, en aangeeft op grond waarvan het oordeel al dan niet wordt overgenomen. Een andere optie is dat als de conclusie is dat het inderdaad een beter oordeel is, we dat oordeel dan ook moeten volgen.

Als het in de nabije toekomst mogelijk wordt om drones bij gewapende conflicten zelf te laten beslissen wie ze aanvallen, zonder menselijke supervisie, dan roept dat dwingende vragen op, niet in het minst omdat het dus ook voor een zelflerend systeem moeilijk inschatten is wie burger is en wie militair. Onderscheid maken tussen burgers en strijders is een voorwaarde van het internationaal humanitair recht: enkel militaire doelen mogen worden aangevallen. Gewapende drones die dit verschil niet kunnen maken, overtreden het internationaal oorlogsrecht. Maar zulke systemen roepen uiteraard nog andere prangende vragen op, zoals over onvoorspelbaarheid. Die kan gaan om een oncontroleerbare wapenwedloop, maar ook over onduidelijkheid over hoe een AI-systeem tot een 'besluit' komt. Bij militaire robots is het paradigma om al deze redenen dat, ongeacht of AI het uiteindelijk wel of niet beter doet, de beslissing over leven en dood - en dus de verantwoordelijkheid voor dat besluit - bij de mens moet

liggen (zie Gabriels, 2021). Een soortgelijk paradigma zou ook kunnen gelden als de robotopvoeder ooit ontwikkeld zou worden: zelfs als deze het op bepaalde vlakken beter zou doen dan de mens dan nog kan het ethisch onwenselijk zijn om een kind door een niet-mens te laten grootbrengen (cf. casus 4).

In dit preadvies wordt de AMA apart van de mens besproken. Niet in de zin dat de mens er geen verantwoordelijkheid voor draagt, maar wel dat het om een aparte entiteit gaat, die de mens mogelijk moreel beter kan leren oordelen. In theorie is het ook denkbaar om een AMA ín de mens implanteren, bijvoorbeeld in ons brein, waarbij de kunstmatige neurale netwerken gekoppeld worden aan onze neurale netwerken, en mens en AMA als het ware samensmelten. Hoe moet zo een type AMA eruitzien? En hoe wenselijk is dit?

Dit is allemaal toekomstmuziek. Toch moeten we het er nu al over hebben. In toenemende mate worden beslissingen vandaag de dag al uitbesteed aan algoritmen (*algorithmic decision-making*, ADM). De toeslagenaffaire toont zonder meer aan dat dit niet zonder risico is.

Een fundamentele vraag is dus: wat willen we op termijn wel en wat niet delegeren aan AI-systemen en onder welke voorwaarden? En binnen welke parameters kan dit oordeel dan een menselijk oordeel vervangen of 'overtreffen'? Zoals eerder vermeld, zijn de kwaliteit en werking van de algoritmen afhankelijk van menselijke input en supervisie. Het is de taak van de mens om algoritmen met een uitgebalanceerde dataset te trainen, om vooringenomenheden te vermijden, en erop toe te zien dat ethische en juridische regels en beginselen worden gerespecteerd. Interdisciplinariteit moet mijns inziens het uitgangspunt worden bij beleidsvorming en in ontwerpteams. Ethici, technici, beleidsmakers, experts uit het betreffende veld en eindgebruikers moeten betrokken worden. Bij elk type AMA is er altijd een cruciale rol voor de mens weggelegd: van het maken van de technologie tot het bepalen van parameters en voorwaarden.

4.5 Wie is verantwoordelijk?

Bij AI-systemen is nog vaak onduidelijk hoe zij tot een besluit komen. Burgers van EU-lidstaten hebben recht op uitleg (cf. *Explainable AI* in de AVG) als een beslissing van een algoritme hun leven beïnvloedt. Bijvoorbeeld wanneer hen een hypotheek geweigerd wordt. Transparantie van het beslissingsproces was al een essentieel onderdeel van AI-expertsystemen in de jaren 1980, maar voor systemen die leren uit big data staat Explainable AI nog in de kinderschoenen. In tegenstel-

ling tot een mens kan zo'n systeem niet reflecteren op de genomen beslissing. De makers moeten in staat zijn om inzage te geven in hoe en op basis waarvan een oordeel of beslissing tot stand is gekomen. Ook al is het niet altijd simpel om de verschillende stappen en de werking van diepe neurale netwerken op een heldere manier uit te leggen aan een leek: de verantwoordelijkheid ligt bij de makers om grip te houden op de technologie, ook als die algoritmen verder leren.[34]

Verder is het mijns inziens aangewezen om ontwikkelaars ook persoonlijk aansprakelijk te stellen voor de algoritmen die ze ontwerpen. De ingenieurs die de algoritmen trainen, hebben de kennis en dus de sleutel in handen om te weten waar het kan mislopen. Bij individuele verantwoordelijkheid is elk lid van de organisatie verantwoordelijk voor zijn of haar bijdrage. Het voordeel van dit model is dat het in principe iedereen motiveert en dat het rechtvaardig is. Bij hiërarchische verantwoordelijkheid is bijvoorbeeld alleen het management (eind) verantwoordelijk. Een individueel verantwoordelijkheidsmodel kan de kans dat een algoritme blind wordt gevolgd helpen beperken. Men weet dan immers: als het misloopt, kan ik aansprakelijk worden gehouden.

Er is een risico dat AMA's tot 'morele luiheid' bij de mens leiden. Zeker zodra AMA's tot morele oordelen kunnen komen (type 3 en 4), maar ook bij de lagere types van AMA's, als deze een rangschikking opstellen, die mensen vervolgens blindelings volgen zonder dat ze er verantwoordelijkheid voor nemen. Dat hangt samen met een belangrijk aspect van onze moraliteit: *moral disengagement*, wat vertaald kan worden als morele ontkoppeling. Dit is een mechanisme dat we in het dagelijkse leven toepassen om ons eigen bedenkelijke gedrag te rechtvaardigen (Bandura, 2002). Een voorbeeld hiervan is door taalgebruik de focus van de handeling weg te leiden. Als in een bedrijf een grote ontslagronde plaatsvindt, dan kun je de focus van het menselijk leed afwenden door te stellen dat je het bedrijf weer 'gezond' maakt of dat je het 'redt' van de ondergang, waardoor het bijna als een moreel aanvaardbare daad klinkt. Een ander voorbeeld is het verleggen van de verantwoordelijkheid. Bij AMA's kan het dus een manier zijn om verantwoordelijkheid af te schuiven op het algoritme of het AI-systeem: 'Het

34 Er zijn al praktische, ethische richtlijnen voor het ontwikkelen van 'ethische' algoritmen. FAT/ML ('Principles for Accountable Algorithms and a Social Impact Statement for Algorithms') staat voor 'fairness', 'accountability en 'transparency' bij machine learning. Een groep van academici en onderzoekers (Google Research en Microsoft Research) ontwikkelde deze lijst, bestaande uit principes en richtlijnen, om bewustwording op het vlak van verantwoordelijkheid (en aansprakelijkheid) bij ontwikkelaars te stimuleren en verhogen. Zie https://www.fatml.org/resources/principles-for-accountable-algorithms Geraadpleegd op 18 april 2021.

algoritme heeft beslist'.

De verantwoordelijkheid van de maker neemt toe naarmate we opklimmen van AMA-1 naar type 4. Bij de volautomatische zelfrijdende auto (AMA-3) kunnen mensenlevens op het spel staan. Bij besluitvorming die een grote impact kan hebben, is toestemming nodig van een onafhankelijke toetsingscommissie en dienen de ontwikkelaars een AI *Impact Assessment* te doen, vergelijkbaar met een *Data Protection Impact Assessment* (DPIA) die deel is van de AVG. In Duitsland is in 2017 een ethische code ontwikkeld voor zelfrijdend transport (zie Tekstbox 2). In de code wordt geadviseerd om relevante scenario's over hoé het zelflerende systeem precies leert en hoé dit de veiligheid verhoogt over te dragen aan een centraal register bij een neutrale instantie, om te komen tot universele standaarden. Dit is een interessante manier om kennis te bundelen, in dit geval over internationale verkeersveiligheid, waarvan wat mij betreft onderzocht moet worden of die ook in andere domeinen op AMA's kan toegepast worden, zoals bij zorgrobots.

Duitsland heeft reeds een ethische code opgesteld voor zowel gedeeltelijk als volautomatisch zelfrijdend transport (zie Luetge, 2017). Als een ongeluk onvermijdelijk is, dan stelt de ethische code dat het strikt verboden is om bijvoorbeeld een verschil te programmeren op basis van persoonlijke kenmerken, zoals leeftijd, geslacht, fysieke of mentale eigenschappen. Dat sluit uit dat in het eerdergenoemde scenario, met een zwangere vrouw en een oudere man, een keuze wordt gemaakt op grond van *deze* kenmerken. Het algoritme moet iedereen gelijk behandelen, vanuit een universele morele regel. Zo vermijd je ook dat de auto beslissingen neemt die moreel moeilijk verdedigbaar zijn.

Tekstbox 2: Een ethische code voor zelfrijdend transport

4.6 Belang van consensus en duidelijke begripsbepaling

Ik begon dit hoofdstuk met het benoemen van een gebrek aan consensus. Dat speelt zowel bij de vier typen AMA's als op metaniveau, bijvoorbeeld bij de definities van cruciale begrippen zoals moraliteit en morele oordelen. De beslissing van welk normatief ethisch kader je volgt bij het bouwen van AMA's, stuurt de output. Vooral bij AMA-3 en 4 kan dit verregaande gevolgen hebben. Een robotopvoeder die consequentialisme aanhoudt, zal andere klemtonen leggen op het belang van genot naargelang die met de inzichten van Jeremy Bentham wordt

getraind, of met die van John Stuart Mill. Dit lijkt triviaal, zeker in een gedachte-experiment, maar bij opvoeding kan het ethisch wenselijker zijn om meer in te zetten op intellectuele vaardigheden en dus Mill te verkiezen boven Bentham. En het is sowieso niet simpel om complexe ideeën over 'goede opvoeding' in welomlijnde criteria vast te leggen.

In dit preadvies staat de vraag naar AI en morele oordelen centraal. Weten we eigenlijk wel wat de belangrijke elementen zijn voor een goed moreel oordeel bij *mensen*? Die basis is nodig om morele oordelen in AI te ontwerpen. Dus de voorwaarde om AMA's te mogen gebruiken bij morele oordelen, is om eerst helderheid te scheppen over hoe de mens tot morele oordelen komt.

Duidelijke definities en begripsbepaling zijn geen overbodige luxe. Spreken we nog wel over hetzelfde als we termen zoals begrip van context, inzicht, morele intuïtie, bedoeld voor menselijke actoren, op AMA's toepassen? Of verschuift de betekenis fundamenteel? De discussie over AMA's (hoofdstuk 1) maakt duidelijk dat het begrip kunstmatige morele actor *an sich* tot veel verwarring leidt. Computersystemen omschrijven als 'morele actoren' kan de verantwoordelijkheid wegleiden van de mensen die erachter zitten en controle op de systemen moeten behouden (Johnson & Miller, 2008). Dit alles maakt de discussie over AMA's, en bij uitbreiding AI, zo ingewikkeld: je moet eigenlijk vooraf duidelijk maken over welk type je spreekt en welke definitie je aanhoudt. Critici stellen dat, zodra je AI definieert als 'moral agent', je mensen op het verkeerde been zet. Filosofen zoals Deborah Johnson en Keith Miller (2008) pleiten ervoor om termen zoals intelligentie, autonomie en 'moral agency' nooit te gebruiken binnen de context van AI. Johnson (2006) stelt voor om AI niet te definiëren als morele actor maar wel als morele entiteit, en wel om drie redenen: AI valt onder onze morele evaluatie, AI-systemen kunnen ons handelen beïnvloeden en elk AI-systeem is het resultaat van menselijke keuzes. Ook begrippen zoals vrije wil maken de discussie erg complex: we moeten eerst de conceptuele, filosofische uitdagingen aangaan voordat we een model van 'vrije wil' in een technologie zouden kunnen bouwen, om tot AMA-4 te kunnen komen (Himma, 2009).

Het zal u als lezer niet verbazen, dat ik die grote vragen niet kan beantwoorden in het bestek van dit preadvies.

5 Besluit

In dit preadvies zocht ik uit in hoeverre kunstmatige intelligentie ingezet kan worden bij morele oordeelsvorming. Op dit moment is AI nog vooral een *tool* en de leersystemen hangen af van de data waar wij, mensen, hen mee trainen. AI is maar zo goed als de data waar die mee gevoed wordt en kan zich bijvoorbeeld niet voorstellen hoe de wereld *zou kunnen zijn*. De AMA's die nu al technisch mogelijk zijn (types 1 en 2), hebben zelf geen moraal, handelen niet vrij en zijn dus niet moreel verantwoordelijk of aansprakelijk. AI-systemen en robots zijn door mensen gemaakte artefacten. Ze vallen onder onze morele evaluatie en controle, en niet omgekeerd. Een algoritme kan zelf geen ethische vragen stellen. Het zal een probleem proberen op te lossen, maar als de beschrijving van het probleem geen ethische kanttekeningen bevat, dan zal het algoritme er geen rekening mee houden. Bovendien kent AI alleen het verleden, niet de toekomst, door in bestaande data (bijvoorbeeld menselijke keuzes uit het verleden) te graven. De modellen kunnen verschillende scenario's berekenen, maar niet wat betreft menselijk gedrag dat nooit eerder vertoond is. Bovendien valt een aantal zaken niet zomaar te voorspellen, bijvoorbeeld vanwege de rol van toeval. Je kan die factoren wel proberen mee te nemen in de analyse, maar het blijft bij kansberekening.

AI is nog niet ver genoeg ontwikkeld om zelfstandig morele besluiten te kunnen of laten nemen. De besluitvorming moet in ieder geval voorlopig bij de mens liggen en het blijft een open vraag in hoeverre het ooit wenselijk is om dit te wijzigen (hoewel ik ernaar neig om te zeggen dat de verantwoordelijkheid bij de mens moet blijven liggen). Daarvoor zouden tenminste drempelwaarden nodig zijn die bepalen vanaf wanneer AI beter moreel kan oordelen dan de mens, en wat dan in specifieke contexten geldt als 'beter'.

Hoewel de kans zeer klein is, is het niet uitgesloten dat moraliteit op een

zeker moment spontaan ontstaat in een multi-agent systeem. En misschien laat AI in de toekomst wel toe om op een aantal tot nu toe filosofische vragen – denk aan vragen over de aard van intelligentie en/of bewustzijn – een meer empirisch antwoord te geven. We hebben veel meer data voorhanden om de systemen mee te trainen, dus er is meer potentieel dat AI ook fundamenteel nieuwe inzichten kan geven in wat de essentie is van intelligentie of bewustzijn. Misschien komt er ooit een dag waarop AI zelfstandig moreel kan handelen, zoals de robots in *Machines Like Me*. Mocht sterke AI er komen, die kan functioneren als een volledig ethische actor (type 4), dan zal deze net zoals de mens verantwoordelijk kunnen worden gehouden. In dat geval moet er worden nagedacht over AI-verantwoordelijkheid en passende straffen voor AI-systemen. Tegelijkertijd moet dit type AMA ook beschermd worden: omdat het dan een hogere vorm van bewustzijn heeft, kan het ethisch gerechtvaardigd zijn om dit niet zomaar te mogen uitschakelen.

Tot die dag is de mens de morele mentor van AI. We moeten dus vooral hopen dat de mens intelligent met AI blijft omgaan. De sleutel tot betere AI ligt in *interdisciplinariteit*: in ontwerpteams, onderzoek, onderwijs en beleid. Ethici en computer- en datawetenschappers kunnen elkaar versterken. Wie ontwerpt, maakt keuzes: niet alleen functioneel maar ook keuzes met ethische impact. Ingenieurs, computerwetenschappers en andere ontwikkelaars in spe moeten de kans krijgen om zich tijdens hun opleiding de ethische vragen eigen te maken en methodes aan te leren om hierop antwoorden te bieden. Hoe beter de makers zélf de vraagstukken begrijpen, hoe beter hun beslissingen zullen zijn (zie ook Gabriels, 2021). Er is ook meer interdisciplinariteit bij beleidsvorming nodig, aangezien maar weinig beleidsmakers een technische achtergrond hebben. Dit kan leiden tot situaties waarin het belang van regulering wordt onderschat. Of dat niet kritisch genoeg wordt gekeken naar het gebruik van het woord AI: verschillende technologieën waarvan start-ups of organisaties beweren dat ze AI zijn, bevatten lang niet altijd AI-technieken.

Een kritische blik op AI is dus te allen tijde aangewezen. Vandaag de dag kunnen we (ook ongemerkt) al meer uit de technologie halen dan dat we erin stoppen. In de toekomst zal die mogelijkheid alleen maar toenemen. Bij elke vorm van vooruitgang van AI is en blijft daarom de meest cruciale rol weggelegd voor de mens.

Referenties

Allen, A. L. (2008). Dredging up the Past: Lifelogging, Memory, and Surveillance. *The University of Chicago Law Review 75*(1), pp. 47-74.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence 12*, pp. 251–261,

Allen, C., Smit, I., & Wallach, W. (2005). Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches. *Ethics and Information Technology 7*(3), pp. 149–155.

Allen, C., & Wallach, W. (2014). Moral Machines: Contradiction in Terms or Abdication of Human Responsibility? In P. Lin, K. Abney, & G.A. Bekey (eds.) (2014). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press, pp. 55-68.

Allott, R. (1991). Objective Morality. *Journal of Social and Biological Structures 14* (4), pp. 455-471.

Anderson, M., Anderson, S. L., & Armen, C. (2005). MedEthEx: Toward a Medical Ethics Advisor. In *Proceedings of the AAAI 2005 Fall Symposium on Caring Machines: AI in Elder Care*, pp. 9–16.

Anderson, M., Anderson, S. L., & Armen, C. (2006). MedEthEx: A Prototype Medical Ethics Advisor. In *Proceedings of the National Conference on Artificial Intelligence 21*(2), Cambridge, MA: AAAI Press/MIT Press, pp. 1759–1765.

Anthony, T. et al. (2020). Learning to Play No-Press Diplomacy with Best Response Policy Iteration. *34th Conference on Neural Information Processing Systems (NeurIPS 2020).*

Arkoudas, K., Bringsjord, S., & Bello, P. (2005). Toward Ethical Robots via Mechanized Deontic Logic. In: *AAAI Fall Symposium on Machine Ethics*, pp. 17–23.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine Experiment. *Nature 563*, pp. 59-64.

Bandura, A. (2002). Selective Moral Disengagement in the Exercise of Moral Agency. *Journal of Moral Education 31*(2), pp. 101-119.

Bauer, W. A. (2018). Virtuous vs. Utilitarian Artificial Moral Agents. *AI & Society 35*, pp. 263–271.

Biller-Andorno, N., & Biller, A. (2019). Algorithm-Aided Prediction of Patient Preferences — An Ethics Sneak Peek. *The New England Journal of Medicine 381*(15), pp. 1480-1485.

Boden, M. A. (2016). *AI. Its Nature and Future*. Oxford: Oxford University Press.

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

Brody, G. H., & Shaffer, D. R. (1982). Contributions of Parents and Peers to Children's Moral Socialization. *Developmental Review 2*, pp. 31-75.

Brooks, R. (2003). *Flesh and Machines: How Robots Will Change Us*. New York: Vintage Books.

CEG, Centrum voor ethiek en gezondheid (2020). *Sensoren in de thuissituatie. De ethiek van e-health. Deel III* (2020).

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics 26*, pp. 501-532.

Cheok, A. D., Edirisinghe, C., & Karunanayaka, K. (2017). Confucius Computer: A Philosophical Digital Agent for Intergenerational Philosophical Play. *Personal and Ubiquitous Computing 21*, pp. 327-343.

Coeckelbergh, M. (2016). Care Robots and the Future of ICT-Mediated Elderly Care: A Response to Doom Scenarios. *AI & Society 31*, pp. 455-462.

Danielson, P. (1992). *Artificial Morality: Virtuous Robots for Virtual Games*. New York: Routledge.

Dietrich, E. (2001). Homo Sapiens 2.0: Why We Should Build the Better Robots of Our Nature. *Journal of Experimental & Theoretical Artificial Intelligence 13*(4), pp. 323–328.

Dodge, M., & Kitchin, R. (2007). 'Outlines of a World Coming into Existence': Pervasive Computing and the Ethics of Forgetting. *Environment and Planning B: Planning and Design 34*(3), pp. 431-445.

Dreyfus, H., & Dreyfus, S. (1986). Why Computers May Never Think Like People. *MIT Technology Review 89*(1), pp. 42-61.

Du Sautoy, M. (2020). *De code van creativiteit. Hoe AI leert schrijven, schilderen en denke*n. Amsterdam: Uitgeverij Nieuwezijds. Oorspronkelijk: *The Creativity Code. How AI Is Learning to Write, Paint and Think* (2019).

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature 542*, pp. 115-118.

Fitzpatrick, K. K., Darcy, A. M., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *Journal of Medical Internet Research (JMIR) Mental Health 4*(2): e19.

Fjelland, R. (2020). Why General Artificial Intelligence Will Not Be Realized. *Humanities and Social Sciences Communications 7*(10). https://www.nature.com/articles/s41599-020-0494-4

Fossa, F. (2018). Artificial Moral Agents: Moral Mentors or Sensible Tools? *Ethics and Information Technology 20*, pp. 115-126.

Gabriels, K. (2019). *Regels voor robots. Ethiek in tijden van AI*. Brussel: VUBPRESS.

Gabriels, K. (2020). *Conscientious AI: Machines Learning Morals*. Brussel: VUBPRESS.

Gabriels, K., & Lanzing, M. (2020). Ethical Implications of Onlife Vitriol. In S. A. Polak, M. Williams, D. Trottier, & T. Olteanu (Eds.) (2020). *Violence and Trolling on Social Media. History, Affect, and Effects of Online Vitriol.* Amsterdam: Amsterdam University Press, pp. 197-214.

Gabriels, K. (2021). Robots met moreel kompas en compassie. *Filosofie-Tijdschrift 31*(2), pp. 20-24.

Gabriels, K., & Volkman, R. (2021). On the Limits of Safe Moral Enhancement through AI Mentors. *CEPE/IACAP 2021: The Philosophy and Ethics of Artificial Intelligence*, Universität Hamburg, Hamburg, (5-9 juli 2021). Online omwille van COVID-19.

Giubilini, A., & Savulescu, J. (2018). The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence. *Philosophy & Technology 31*, pp. 169-188.

Haidt, J. (2003). The Moral Emotions. R. Davidson, K. Sherer, & H. Goldsmith (ed.) (2003), *Handbook of Affective Sciences*. Oxford: Oxford University Press, pp. 852-870.

Haidt, J. (2012/2013). *The Righteous Mind. Why Good People Are Divided by Politics and Religion*. Londen: Penguin Group.

Himma, K. E. (2009). Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent? *Ethics and Information Technology 11*(1), pp. 19–29.

Jacobs, A., Tytgat, L., Maus, M., Meeusen, M. & Vanderborght, B. (red.) (2019). *Homo Roboticus. 30 vragen en antwoorden over mens, robot & artificiële intelligentie*. Brussel: VUBPRESS.

Johnson, D. G. (2006). Computer Systems: Moral Entities but not Moral Agents. *Ethics and Information Technology 8*(4), pp. 195–204.

Johnson, D. G., & Miller, K. W. (2008). Un-Making Artificial Moral Agents. *Ethics and Information Technology 10*(2–3), pp. 123–133.

Johnson, D. G., & Verdicchio, M. (2017). Reframing AI Discourse. *Minds & Machines 27*, pp. 575-590.

Klincewicz, M. (2016). Artificial Intelligence as a Means to Moral Enhancement. *Studies in Logic, Grammar, and Rhetoric 48*(61).

Kurzweil, R. (2005). *The Singularity is Near. When Humans Transcend Biology*. New York: Penguin Books.

Lara, F., & Deckers, J. (2020). Artificial Intelligence as a Socratic Assistant for Moral Enhancement. *Neuroethics 13*, pp. 275-287.

Leben, D. (2019). *Ethics for Robots. How to Design a Moral Algorithm.* Londen / New York: Routledge.

Lin, P., Abney, K., & Bekey, G. A. (eds.) (2014). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press.

Luetge, C. (2017). The German Ethics Code for Automated and Connected Driving. *Philosophy & Technology 30* (4), pp. 547-558.

Malle, B. F. (2016). Integrating Robot ethics and Machine Morality: The Study and Design of Moral Competence in Robots. *Ethics and Information Technology 18*(4), pp. 243–256.

Mayer-Schönberger, V. (2009). *Delete: The Virtue of Forgetting in the Digital Age*. Princeton: Princeton University Press.

McEwan, I. (2019). *Machines Like Me.* Londen: Jonathan Cape.

Miller, K. W., Wolf, M. J., & Grodzinsky, F. (2017). This "Ethical Trap" Is for Roboticists, Not Robots: On the Issue of Artificial Agent Ethical Decision-Making. *Science and Engineering Ethics 23*(2), pp. 389–401.

Moor, J. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, pp. 18-21.

Moor, J. (2007). Taking the Intentional Stance Toward Robot Ethics. *APA (American Philosophical Association) Newsletters 6 (2)*.

Mortier, P., Kiekens, G., Auerbach, R. P., Cuijpers, P., Demyttenaere, K., Green, J. G., Kessler, R. C., Nock, M. K., Zaslavsky, A. M., & Bruffaerts, R. (2017). A Risk Algorithm for the Persistence of Suicidal Thoughts and Behaviors during College. *Journal of Clinical Psychiatery 78*(7), e828-e836.

Nagenborg, M. (2007). Artificial Moral Agents: An Intercultural Perspective. *International Review of Information Ethics 7*, pp. 129–133.

Nietzsche, F. (1983/2006). *Over nut en nadeel van geschiedenis voor het leven*. Epe: Historische Uitgeverij.

Nyholm, S. (2021). De ethiek van zelfrijdende auto's. *Filosofie-Tijdschrift 31*(2), pp. 25-29.

Ochs, E., & Kremer-Sadlik, T. (2007). Introduction: Morality as Family Practice. *Discourse & Society 18*(1), pp. 5-10.

Podschwadek, F. (2017). Do Androids Dream of Normative Endorsement? On the Fallibility of Artificial Moral Agents. *Artificial Intelligence and Law 25*(3), pp. 325–339.

Ponzetti, J. J. Jr. (2010). Virtue Ethics as the Basis for Family Interaction. In S. Scales, A. Potthast, & L. Oravecz (eds.) (2010). *The Ethics of the Family*. Cambridge: Cambridge Scholars Publishing, pp. 308-316.

Rachels, J. (2007). *The Elements of Moral Philosophy* (Fifth edition, by Stuart Rachels). New York: McGraw-Hill.

Royakkers, L. (2021). Robotisering als rationalisering. Het probleem van dehumanisering. *Filosofie-Tijdschrift 31*(2), pp. 4-8.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine 36*(4), pp 105-114.

Russell, S. J., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach* (Third Edition). Essex: Pearson Education Limited.

Savulescu, J., Gyngell, C., & Kahane, G. (2021). Collective Reflective Equilibrium in Practice (CREP) and Controversial Novel Technologies. *Bioethics*, pp. 1-12.

Sharkey, N., & Sharkey, A. (2010). The Crying Shame of Robot Nannies: An Ethical Appraisal. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems 11*(2), pp. 161–190.

Sharkey, N., & Sharkey, A. (2014). The Rights and Wrongs of Robot Care. In P. Lin, K. Abney, & G.A. Bekey (eds.) (2014). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press, pp. 267-282.

Sharkey, A. J. C. (2016). Should We Welcome Robot Teachers? *Ethics and Information Technology 18*, pp. 283-297.

Sharkey, A. (2017). Can Robots Be Responsible Moral Agents? And Why Should We Care? *Connection Science 29*(3), pp. 210–216.

Smetana, J. G. (1999). The Role of Parents in Moral Development: A Social Domain Analysis. *Journal of Moral Education 28*(3), pp. 311-321.

Sullins, J. P. (2006). When Is a Robot a Moral Agent? *International Review of Information Ethics 6*(12), pp. 23–30.

Tegmark, M. (2017). *Life 3.0. Mens zijn in het tijdperk van kunstmatige intelligentie*. Amsterdam: Maven Publishing.

Van de Poel, I., & Royakkers, L. (2011). *Ethics, Technology, and Engineering: An Introduction*. Chichester, West Sussex: Wiley-Blackwell.

Van Mens, K., Elzinga, E., Nielen, M., Lokkerbol, J., Poortvliet, R., Donker, G., Heins, M., Korevaar, J., Dückers, M., Aussems, C., Helbich, M., Tiemens, B., Gilissen, R., Beekman, A., & de Beurs, D. (2020). Applying Machine Learning on Health Record Data from General Practitioners to Predict Suicidality. *Internet Interventions 21*.

Van Wynsberghe, A., & Robbins, S. (2019). Critiquing the Reasons for Making Artificial Moral Agents. *Science and Engineering Ethics 25*, pp. 719-735.

Walker, L. J., & Hennig, K. H. (1999). Parenting Style and the Development of Moral Reasoning. *Journal of Moral Education 28*(3), pp. 359-374.

Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Walton, D. (2016). Some Artificial Intelligence Tools for Argument Evaluation: An Introduction. *Argumentation*, pp. 1-23.

Yilmaz, L., Franco-Watkins, A., & Kroecker, T. S. (2017). Computational Models of Ethical Decision-Making: A Coherence-Driven Reflective Equilibrium Model. *Cognitive Systems Research 46*, pp. 61-74.

Yudkowsky, E. (2012). Friendly Artificial Intelligence. In A. H. Eden, J. H. Moor, J. H. Søraker, & E. Steinhart (eds.) (2012). *Singularity Hypotheses. A Scientific and Philosophical Assessment*. Berlijn / Heidelberg: Springer-Verlag, pp. 181-193.

Dankwoord

De auteur wenst haar oprechte dank uit te spreken aan het bestuur van de NVBe voor de uitnodiging om het preadvies te schrijven. NVBe-bestuursleden Marieke Bak en Verna Jans en voorzitter André Krom gaven gedetailleerde feedback op eerdere versies van het manuscript. Hun waardevolle, diepgaande suggesties, bijvoorbeeld ten aanzien van de term AMA of bij de inzet van AI bij euthanasie (casus 1), verbeterden de analyses niet alleen aanzienlijk, maar vergrootten ook de leesbaarheid ervan. De auteur dankt ook Stephan Van der Waart van Gulik en Bart De Vylder voor de brainstormsessies.

Over de auteur

Dr. Katleen Gabriels is moraalfilosofe, gespecialiseerd in computer- en machine-ethiek. Ze werkt als universitair docent in techniekfilosofie en -ethiek aan de universiteit van Maastricht, waar ze ook opleidingsdirecteur is van de bachelor-opleiding Digital Society. Ze is de auteur van *Conscientious AI. Machines Learning Morals* (VUBPRESS, 2020). Dat is de herwerkte, Engelstalige versie van *Regels voor robots. Ethiek in tijden van AI* (VUBPRESS, 2019). Ze is ook de auteur van *Onlife. Hoe de digitale wereld je leven bepaalt* (Lannoo, 2016) en, samen met wiskundige Ann Dooms, van *Van Melkweg tot moraal* (Academia Press, 2020). Katleen is bestuurslid van de International Society for Ethics and Information Technology (INSEIT), bestuurslid en plaatsvervangend voorzitter van ETHICOMP en geaffilieerd lid van 4TU Centre for Ethics and Technology. De Leerstoel Willy Calewaert 2018-2019 (demens.nu) werd aan haar toegekend door de faculteit Ingenieurswetenschappen van de VUB.

Preadvies

uitgebracht t.b.v. het jaarsymposium van de
Nederlandse Vereniging voor Bio-ethiek 11 juni 2021

Over de NVBe

De Nederlandse Vereniging voor Bio-ethiek (NVBe) is opgericht in 1993 en wil bijdragen aan een open discussie over bio-ethische kwesties in samenleving en wetenschap. Ze doet dat door het contact te bevorderen tussen vertegenwoordigers uit verschillende vakgebieden, instellingen en organisaties die bij bio-ethische kwesties betrokken zijn, de bio-ethische reflectie in de relevante sectoren te stimuleren en bio-ethische discussiebijdragen te presenteren. Bio-ethiek houdt zich bezig met morele vragen die worden opgeroepen door allerlei ontwikkelingen betreffende het leven en de natuur.
De NVBe publiceert elk jaar een Preadvies over een onderwerp dat maatschappelijk actueel en wetenschappelijk uitdagend is. Vier keer per jaar verschijnt het Podium voor Bio-ethiek. Daarnaast organiseert de NVBe jaarlijks een Jaarsymposium en een Onderwijsmiddag.
Het Preadvies wordt op uitnodiging van de NVBe geschreven door een wetenschapper. Het dient als aanleiding en katalysator voor een gedachtewisseling tijdens het Jaarsymposium en ook daarna in wetenschap en samenleving.
Voor meer informatie over de NVBe, eerdere Preadviezen, alle jaargangen van Podium en de voordelen en de kosten van het lidmaatschap (ook voor instellingen): zie de website www.nvbe.nl.

Nederlandse Vereniging
voor Bio-Ethiek

www.nvbe.nl