Podium voor Bio-ethiek

jaargang 28 · nr. 3 · 2021

De waarde van AI voor de bio-ethiek

Automatiseren van morele oordeelsvorming: antwoorden of vragen? | Marianne Boenink

Helpt artificiële intelligentie bij ingrijpende besluitvorming door dokters? | Jan Hulscher, Elisabeth M.W. Kooi, Caspar Chorus, Annebel ten Broeke en Els Maeckelberghe

Menselijke leiding is cruciaal bij AI-systemen die moreel redeneren ondersteunen | Sophie van Baalen en Linda Kool

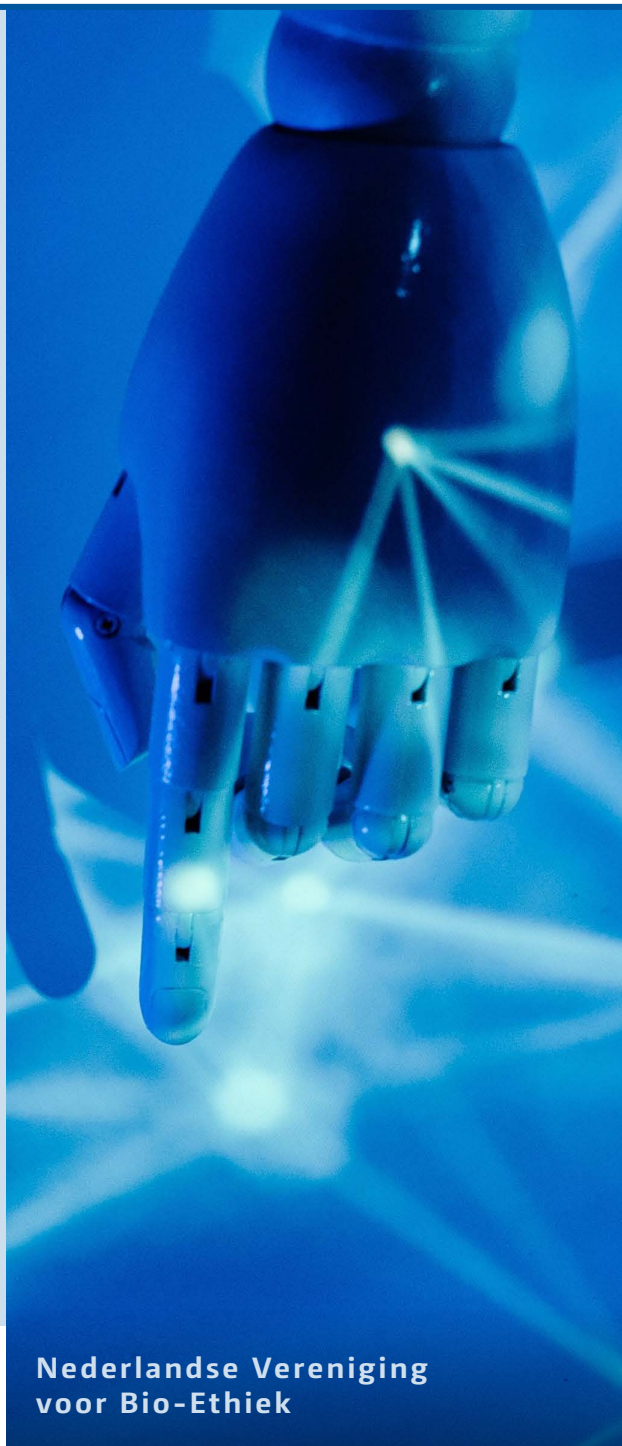
AI en morele oordeelsvorming: van principes naar het vormgeven van ethische AI-praktijken | Rik Wehrens, Sydney Howe, Esra Demir, Kostina Prifti, Klaus Heine en Evert Stamhuis

Kunstmatige intelligentie en/of intelligente levenskunst | Jan Vorstenbosch

Vertrouwen in de geneeskunde en kunstmatige intelligentie | Lily Frank en Michal Klincewicz

Interview met Katleen Gabriels: 'We moeten niet denken dat AI als een orakel antwoord kan geven op morele vragen' | Marieke Bak en Sjaak Swart

Boekbespreking: Rechten en plichten voor AI? | Sjaak Swart



Nederlandse Vereniging voor Bio-Ethiek

Colofon

De NVBe streeft naar:

1. stimulering van de bio-ethiek (humane, dier- en natuurethiek) in relevante sectoren;
2. contacten tussen vertegenwoordigers uit verschillende vakgebieden, instellingen en organisaties die betrokken zijn bij bio-ethische kwesties;
3. erkenning van de waarden van een open discussie over bio-ethische problemen in wetenschap en samenleving;
4. presentatie van discussies in de bio-ethiek in Nederland

Het Podium voor Bio-ethiek (voorheen de Nieuwsbrief) van de vereniging draagt bij aan deze doelen door publicatie van bio-ethisch nieuws (van binnen en buiten de vereniging) en bondige, voor een breed publiek toegankelijke, interdisciplinaire bijdragen over bio-ethische kwesties. Het Podium voor Bio-ethiek verschijnt vier keer per jaar en wordt toegezonden aan leden van de NVBe. Het Podium voor Bio-ethiek, mededelingen uit de Vereniging en bio-ethische informatie verschijnen ook op www.nvbe.nl.

Redactieadres

Secretariaat NVBe, info@nvbe.nl, t.a.v. Podium-redactie

Redactie

drs. Marieke Bak, drs. Nina Breedveld, dr. Rosanne Edelenbosch, drs. Sicco Polders, drs. Amber Spijkers, dr. Dirk Stemerding, dr. Sjaak Swart en dr. Rik Wehrens

Website

dr. Ana Pereira Daoud

Opmaak

drs. Ger Palmboom

Instructie voor bijdragen

Bijdragen in overeenstemming met de doelstelling van Het Podium voor Bio-ethiek zijn van harte welkom. Voor suggesties en vragen kunt u zich wenden tot de redactie via het e-mailadres. Artikelen bij voorkeur rond de 1500 woorden, boekbesprekingen en verslagen van congressen, conferenties, etc. maximaal 500 woorden. Bij voorkeur geen uitgebreide literatuurverwijzingen. Bijdragen kunt u per e-mail sturen naar het redactieadres. De redactie behoudt zich het recht voor bijdragen te weigeren of in te korten.

Lid worden?

Het lidmaatschap van de Nederlandse Vereniging voor Bio-ethiek (NVBe) is er voor iedereen die zich op de een of andere manier betrokken voelt bij de levenswetenschappen in brede zin en de ethische reflectie daarop. Op de website www.nvbe.nl (doorklikken naar 'Lidmaatschap') vindt u een formulier waarmee u zich kunt aanmelden als lid. De ledenadministratie is te bereiken via ledenadministratie@nvbe.nl.

Voor vragen en opmerkingen kunt u contact opnemen met het secretariaat: info@nvbe.nl

NVBe Nederlandse Vereniging voor Bio-Ethiek

Thema: De waarde van AI voor de bio-ethiek Podium voor Bio-ethiek • jaargang 28 • nr. 3 • 2021

Inhoudsopgave

De waarde van AI voor de bio-ethiek

2 **Redactioneel**

7 **Automatiseren van morele oordeelsvorming: antwoorden of vragen?**
Marianne Boenink

13 **Helpt artificiële intelligentie bij ingrijpende besluitvorming door dokters?**
Jan Hulscher, Elisabeth M.W. Kooi, Caspar Chorus, Annebel ten Broeke en Els Maeckelberghe

19 **Menselijke leiding is cruciaal bij AI-systemen die moreel redeneren ondersteunen**
Sophie van Baalen en Linda Kool

25 **AI en morele oordeelsvorming: van principes naar het vormgeven van ethische AI-praktijken**
Rik Wehrens, Sydney Howe, Esra Demir, Kostina Prifti, Klaus Heine & Evert Stamhuis

32 **Kunstmatige intelligentie en/of intelligente levenskunst**
Jan Vorstenbosch

37 **Vertrouwen in de geneeskunde en kunstmatige intelligentie**
Lily Eva Frank en Michal Klincewicz

43 **'We moeten niet denken dat AI als een orakel antwoord kan geven op morele vragen'**
Een interview met Katleen Gabriels
Marieke Bak en Sjaak Swart

49 **Rechten en plichten voor AI?**
Boekbespreking
Sjaak Swart

54 **Nieuws uit de Vereniging**
André Krom

56 **Bericht van het Rathenau Instituut: In Memoriam Melanie Peters**
Bestuur en medewerkers van het Rathenau Instituut

58 **Nieuws van het Centrum voor Ethiek en Gezondheid**
Myrthe Lenselink en Sandra in 't Groen

60 **Berichten van Unesco**
Marike Bontenbal

Redactioneel

Op 11 juni van dit jaar presenteerde Katleen Gabriels (Universiteit Maastricht) op het online NVBe jaarsymposium haar preadvies '*Siri, wat adviseer jij? Over het gebruik van kunstmatige intelligentie voor morele oordeelsvorming*' (Gabriels, 2021, te downloaden via nvbe.nl). Zij vraagt zich daarin af in hoeverre een AI-systeem ingezet kan worden bij morele oordeelsvorming. In dit themanummer zetten we het gesprek voort. Diverse auteurs reageren op het preadvies, en ook Gabriels zelf komt aan het woord in een interview.

De aandacht voor dit thema volgt uit de snelle ontwikkeling van AI-systemen die steeds complexere problemen aankunnen (AI staat voor Artificial Intelligence of kunstmatige intelligentie). Aanvankelijk werkten AI-systemen vaak *top-down*. Daarbij wordt uit bepaalde afleidingsregels en gegevens deductief een uitkomst afgeleid bij een gegeven input of vraag. De ontwikkeling van neurale netwerken in combinatie met de beschikbaarheid van big data heeft in het laatste decennium de ontwikkeling van *bottom-up* machine learning toepassingen mogelijk gemaakt. Deze kunnen patronen en regels in enorme hoeveelheden data opsporen en op grond daarvan tot een resultaat komen, waaronder bijvoorbeeld gezichtsherkenning en tekstvertalingen. In de toekomst wordt verwacht dat AI-systemen zelfs dingen kunnen die we nu nog als uniek menselijk beschouwen, zoals moreel oordelen en reflecteren. Zou AI op dat vlak taken van de mens kunnen ondersteunen of zelfs kunnen overnemen? Inclusief die van (bio-) ethici? Zijn er redenen om specifieke ontwikkelingen op dit vlak te stimuleren of bepaalde toepassingen juist zoveel mogelijk te beperken?

Siri, wat adviseer jij?

Gabriels schetst in haar preadvies in navolging van Moor (2006) vier niveaus van Artificial Moral Agents (AMA's) die zich onderscheiden in de mate van autonomie, complexiteit en morele gevoeligheid. Op niveau 1 functioneert AMA-1, de ethische impact actor. Dat is een systeem dat gegevens aanlevert waarmee de mens tot beter onderbouwde morele oordelen kan komen. In de context van euthanasie bespreekt ze of dit van waarde kan zijn bij een patiënt die uitzichtloos en

ondraaglijk lijdt, niet meer in staat is zijn of haar wil te uiten en geen (duidelijke) wilsverklaring heeft. Zou een AMA-1 behulpzaam kunnen zijn om de wil van de patiënt te duiden of te reconstrueren?

AMA-2 verwijst naar een impliciet ethische actor, een systeem dat functies heeft ingebouwd gekregen om ethische redenen. Een voorbeeld is een browser die automatisch pornografische sites blokkeert als jonge kinderen het internet opgaan. Zo'n systeem kent geen ethiek maar lijkt zich wel als zodanig te gedragen. AMA-1 en AMA-2 worden tot de zwakke AI gerekend omdat deze systemen binnen een vaste en beperkte set van functies en regels werken.

Dat geldt niet voor de sterke AI waartoe AMA-3 en AMA-4 worden gerekend, alhoewel het onderscheid niet scherp is. AMA-3 is de expliciet ethische actor, een systeem dat ethische regels kent en gebruikt om tot een uitkomst of beslissing te komen. Het kan ook van ervaringen leren en suggesties doen voor een 'oplossing'. Een voorbeeld is de volledig zelfrijdende auto. De gedachte is dat de mens bij deze systemen uiteindelijk het laatste woord heeft.

Het meest vergaande AI-systeem is AMA-4, de volledig ethische actor die zelfstandig morele oordelen kan vellen en beslissingen kan nemen, over enorm veel gegevens beschikt, in staat is zich verder te ontwikkelen door opgedane ervaringen, en oordelen en beslissingen ook kan rechtvaardigen. Gabriels noemt als voorbeeld de robotopvoeder die toegang heeft tot alle kennis, de menselijke taal in haar context begrijpt, onpartijdig en gewetensvol handelt en eerdere oordelen kan herzien. Zo'n wilsbekwame robot zou in theorie ingezet kunnen worden bij de opvoeding van kinderen.

Qua techniek bestaan de eerste twee niveaus van AMA's reeds. De twee andere niveaus bestaan nog niet in de praktijk, maar de ontwikkelingen gaan razendsnel. Deze toekomstige systemen roepen tal van vragen op. Over de aard van ethiek, bijvoorbeeld of ethiek wel teruggebracht kan worden tot patronen in data. Vergt ethiek niet ook uitleg, onderbouwing, begrip, empathie, discussie en menselijke interactie? Ook roept het vragen op over het verschijnsel van moreel oordelen zelf. Wat doen mensen eigenlijk precies als zij moreel oordelen en zou AI daartoe in staat zijn? *Last but not least* roept het morele vragen op. Zoals: is de ontwikkeling en het gebruik van dergelijke systemen zelf wel moreel juist? Stuurt het onze ethiek niet in een bepaalde richting? Willen en kunnen wij wel samenleven met moreel oordelende AI-systemen? Hoever moeten we daarin willen gaan? Het zijn vragen waarover Gabriels zich buigt in haar preadvies en waar de auteurs in hun bijdragen op reflecteren.

De bijdragen

Verschillend als ze zijn, hebben de bijdragen aan dit themanummer twee dingen gemeen. Zo wijzen ze ieder op hun eigen manier op tekortkomingen in benaderingen waarbij AI ingezet zou worden voor morele oordeelsvorming. Ook stellen alle bijdragen meer of minder expliciet grenzen aan welke typen AMA ingezet zouden (moeten) kunnen worden. Die grens lijkt ten hoogste bij AMA-3 te liggen, bij de expliciet ethische actor.

Marianne Boenink vraagt zich af wat kwalitatieve oordeelsvorming inhoudt en of dat zich wel tot een algoritme laat reduceren. Zij constateert dat mensen in de praktijk van moreel oordelen de bestaande situatie en regels vaak herinterpreteren en nieuwe vragen stellen. Een proces waarin ethici juist bedreven worden geacht. Volgens de auteur wordt goede oordeelsvorming niet zozeer door de uitkomsten bepaald maar veel meer door de aard van dit proces. AI-systemen moeten daarom geen antwoorden geven maar juist vragen stellen.

Jan Hulscher en collega's stellen dat ook dokters hun overwegingen niet altijd kunnen expliciteren. Zij schetsen een AI-systeem, het al bestaande 'BAIT' model, dat dokters kan helpen bij lastige keuzes. Het systeem genereert verschillende keuzeopties voor een medisch-ethisch dilemma, gebaseerd op eerdere oordelen van menselijke professionals bij een vergelijkbaar probleem. De individuele arts neemt vervolgens de beslissing maar kan zich daarbij spiegelen aan de output van het systeem. Een dergelijke toepassing kan volgens de auteurs uitgebreid worden met perspectieven van patiënten, en zou zo de gezamenlijke besluitvorming kunnen ondersteunen.

Daarop aansluitend stellen *Sophie van Baalen en Linda Kool* dat AI en mensen elkaar kunnen aanvullen in morele oordeelsvorming. Ze beargumenteren dat er geen objectieve criteria zijn voor de juistheid van een moreel oordeel. AI kan weliswaar verschillende menselijke tekortkomingen helpen beperken, maar kent zelf ook een belangrijke beperking: voor oordelen over concrete situaties in de echte wereld zijn altijd menselijke vermogens nodig. Zij bepleiten mede daarom om af te zien van een *human in the loop* benadering ten faveure van een *human in the lead* benadering.

Rik Wehrens en collega's signaleren dat in de huidige benaderingen van AI-ethiek de nadruk veelal ligt op het formuleren van generieke ethische principes en daarop gebaseerde toetsingsinstrumenten. Zo'n 'principiële benadering' van AI-ethiek is volgens de auteurs niet afdoende om ethische AI-praktijken

vorm te geven. Vaak zal alleen in de praktijk blijken wat door betrokkenen als ethisch problematisch wordt beschouwd en wat niet. Bovendien gaat de 'principiële benadering' volgens de auteurs van een te individualistische opvatting van morele oordeelsvorming uit. De auteurs pleiten daarom voor meer empirisch en etnografisch gedreven onderzoek naar moreel handelen in concrete praktijken waaraan AI-systemen een gunstige bijdrage zouden moeten leveren.

Vervolgens vraagt *Jan Vorstenbosch* zich af of toekomstige AI-systemen de patiënt kunnen ondersteunen in de keuzes die hij in een medische situatie moet of kan maken, in het bijzonder bij het geven van 'informed consent' voor een behandeling. Hij laat zien dat de informatiebehoefte en de rol van zowel de patiënt als arts danig verschilt in verschillende contexten zoals curatieve zorg, huisartsbezoek, preventietesten en euthanasieverzoeken. De auteur concludeert dat de AI-mogelijkheden in ieder geval een flinke dosis levenskunst bij de patiënt veronderstellen.

Waar voorgaande bijdragen zich direct richten op morele AI systemen gaan *Lily Frank en Michal Klincewicz* in op de vraag wat de opkomst van medisch AI systemen in het algemeen doet met het institutionele vertrouwen in de geneeskunde. Dit vertrouwen is een belangrijke voorwaarde in de medische praktijk omdat de patiënt grotendeels de kennis ontbeert waarmee artsen oordelen. Toekomstige AI-systemen kunnen dit institutionele vertrouwen in de medische professie ondermijnen door gebrekkige uitleg van morele oordelen en onduidelijkheid over verantwoordelijkheid en aansprakelijkheid bij mogelijke fouten.

In een interview dat de themaredactie met haar hield, komt *Katleen Gabriels* weer aan het woord. We blikken in het interview terug op het advies en bespreken mogelijkheden voor toekomstig onderzoek en beleid. In aanvulling op het preadvies geeft Gabriels aan dat AI-systemen weliswaar consistenter en sneller werken dan mensen maar dat we niet moeten denken dat zij als een orakel een antwoord hebben op onze morele vragen. Haar favoriete AI-systeem is een AMA-3 omdat het de mens op socratische wijze kan ondersteunen bij moreel oordelen door de juiste vragen te stellen.

Sjaak Swart bespreekt tot slot het boek 'Robot Rules. Regulating Artificial Intelligence' van Jacob Turner. Sterke AI onderscheidt zich volgens Turner van alle voorgaande technologieën door 'agency', waardoor het ongelimiteerd toepassingen kent en onbeheersbaar kan worden. Dat wordt maatschappelijk nauwelijks opgemerkt en regulering blijft volgens de auteur vooralsnog beperkt tot zelfregulering door AI-bedrijven.

Enkele overwegingen

In de medische wereld wordt al veel gebruik gemaakt van AI-systemen, bijvoorbeeld voor diagnostiek. Omdat daar ook concrete ethische dilemma's aan verbonden kunnen zijn, ligt het voor de hand dat veel morele AMA-discussies gericht zijn op medisch-ethische problemen, zoals ook blijkt uit de bijdragen in dit themanummer. Maar ook in andere domeinen van de bio-ethiek, zoals dierexperimentencommissies, besluitvorming over natuur en milieu, fokprogramma's voor wilde en gedomesticeerde dieren, kunnen AMA's wellicht een rol gaan spelen.

Wanneer we kijken naar de bijdragen in dit themanummer dan valt zoals gezegd op dat alle auteurs bedenkingen hebben of beperkingen voorstellen en niet verder willen gaan dan AMA-3 waarbij het primaat van de oordeelsvorming nog steeds bij de mens ligt. Daarvoor worden verschillende redenen aangevoerd samenhangend met vertrouwen, verantwoordelijkheid en aansprakelijkheid, de rol van praktijken bij het tot stand komen van morele oordelen en bij concepties van het goede leven. Dat roept de vraag op of we de AI-ontwikkeling – met name waar het gaat om deze typische menselijke functies als (moreel) oordelen – strenger zouden moeten reguleren en wellicht zelfs beperken. De analyse van Jacob Turner in zijn besproken boek doet evenwel vermoeden dat de ontwikkelingen gewoon zullen doorgaan omdat daar grote belangen aan gekoppeld zijn.

Ten slotte, de impliciet ethische actor (AMA-2), een systeem dat zich kenmerkt door haar impliciete morele impact, doet sterk denken aan het concept van normatieve scripts in wetenschaps- en technologiestudies (STS). Een van de mooiste voorbeelden daarin is de kinderlepel die zodanig gebogen is dat die het kind dwingt met de rechterhand te eten. Bij een AMA-2 worden de scripts weliswaar doelbewust ingebouwd en wordt gebruik gemaakt van (grote hoeveelheden) data, maar of dat een relevant verschil is, is de vraag. Misschien is iedere technologie wel vergelijkbaar met een AMA-2.

Sjaak Swart, Marieke Bak en André Krom

Literatuur

Gabriels, K. (2021). *Siri, wat adviseer jij? Over het gebruik van kunstmatige intelligentie voor morele oordeelsvorming*. Preadvies Nederlandse Vereniging voor Bio-ethiek.

Moor, J. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), pp. 18-21.

Automatiseren van morele oordeelsvorming: antwoorden of vragen?

Marianne Boenink

De laatste jaren zie je steeds meer initiatieven om medische oordeelsvorming uit te besteden aan 'slimme' software omdat de betrouwbaarheid van het oordeel dan zal toenemen. Zo zou een computer afwijkingen op scans beter en sneller herkennen, of gezondheidsrisico's beter kunnen inschatten. Artsen reageren hier vaak huiverig op: laat klinische oordeelsvorming zich wel tot een algoritme reduceren? Die vraag geldt ook voor de mogelijke opkomst van AI-systemen voor morele oordeelsvorming.

Het lezen van Katleen Gabriels' preadvies over automatisering van morele oordeelsvorming (Gabriels, 2021) vond ik een confronterende ervaring, omdat het fundamentele vragen oproept over de professionele expertise van ethici. Morele oordeelsvorming is natuurlijk niet voorbehouden aan ethici, dus in principe staat er een algemeen-menselijke activiteit ter discussie. Maar het vakgebied van de ethiek gaat wel bij uitstek over de kwaliteit van morele oordeelsvorming en hoe die te verbeteren. Professionele ethici spelen ook vaak de adviserende rol die in dit preadvies aan sommige vormen van kunstmatige intelligentie (Artificial Intelligence, verder AI) wordt toegedacht.

Een confronterende ervaring, omdat het fundamentele vragen oproept over de professionele expertise van ethici

In deze bijdrage neem ik daarom het fenomeen morele oordeelsvorming onder de loep. Welke opvatting van morele oordeelsvorming ligt ten grondslag aan de hypothese dat ethische AI-systemen (of Artificial Moral Agents, AMA's, zoals Gabriels ze noemt) tot betere morele oordelen kunnen komen dan mensen? Dat vereist meer inzicht in (1) hoe morele oordeelsvorming in software wordt gecodeerd, (2) in hoeverre morele oordeelsvorming bij mensen anders verloopt, en (3) of dat verschil de kwaliteit van oordeelsvorming aantast.

De manier waarop software voor morele oordeelsvorming wordt geprogrammeerd geeft een eerste idee van hoe die oordeelsvorming wordt geconceptualiseerd. Gabriels bespreekt in het preadvies dat er in de praktijk drie manieren zijn om AI-systemen voor morele oordeelsvorming te programmeren. De eerste manier is 'top down', waarbij morele regels (dat kunnen overigens ook deugden zijn) in het systeem worden ingebouwd, die vervolgens op casuïstiek worden toegepast. De tweede manier is 'bottom up', waarbij het systeem veel casussen gevoed krijgt, waaruit het dan zelf (flexibele) regels kan afleiden. En ten slotte is er een hybride manier, die beide combineert.

Elke manier heeft zijn eigen problemen. Voor de top-down benadering is de grote vraag *welke* regels dan eigenlijk opgenomen moeten worden. Daar blijken (niet verrassend) uiteenlopende ideeën over te bestaan. Daardoor zijn er gevolgenethische, deontologische, confucianistische en deugdethische AI-systemen. In principe is een pluralistische combinatie denkbaar, maar om tot een oordeel te komen moet er wel een hiërarchie tussen de verschillende overwegingen worden vastgesteld. Gabriels laat zien hoe ingewikkeld het is ethische regels in software te coderen, zelfs als je maar één theorie gebruikt. Mijn punt is dat top-down programmeren een consensus veronderstelt over de regels (en over de hiërarchie tussen die regels). Die is onder ethici ver te zoeken.

Wellicht is de bottom-up benadering kansrijker, omdat de software hier zelf algemene overwegingen afleidt uit een grote hoeveelheid casuïstiek. Dat woordje 'zelf' is echter misleidend: deze manier van programmeren veronderstelt namelijk dat de ingevoerde casuïstiek (de 'leerset') al van een kwaliteitsoordeel is voorzien. Het roept de vraag op of er op een moreel juiste manier op een casus is gereageerd, en door wie en hoe dat wordt bepaald. Is er wel voldoende consensus over wat telt als een 'moreel verantwoorde (of acceptabele) respons' om zo'n leerset samen te stellen? Juist bij ingewikkelde casuïstiek zal daar vaak onenigheid over zijn, en soms kunnen ook meerdere reacties verdedigbaar zijn.

Hybride programmeren lijkt een middenweg te varen, maar krijgt in feite met de problemen van beide benaderingen te maken. Er zijn zowel breed gedragen regels als niet-controversiële oordelen nodig. En dan moet men het ook nog eens worden over volgorde en hiërarchie: kijk je eerst naar de patronen in de casus, of eerst naar de regels? En wat moet de doorslag geven als die twee in verschillende richtingen wijzen? Het lijkt er op dat de uitdagingen van ethische

oordeelsvorming worden uitvergroot als we die oordeelsvorming in AI-vorm willen vastleggen.

Regels en interpretatie

Los van hoe de algoritmes tot stand komen, alle AI-systemen representeren morele oordeelsvorming als een proces waarin algemene regels worden gekoppeld aan concrete casussen – wat natuurlijk typerend is voor algoritmes. Er wordt eigenlijk een onderliggende taxonomie verondersteld met op de ene as mogelijke kenmerken van een situatie en op de andere as mogelijke relevante normatieve regels. Aan elk hokje van de tabel zit een conclusie vast wat te doen. Oordeelsvorming betekent dan dat je een casus in een van de hokjes plaatst, waaruit de conclusie automatisch volgt. Die conclusie kan overigens ook een spectrum van verdedigbare opties zijn.

Dit idee lijkt op het eerste gezicht wel te sporen met ook onder filosofen gangbare ideeën. Gabriels refereert bijvoorbeeld aan Haidts opvatting dat morele oordeelsvorming een combinatie is van perceptie en redeneren (Haidt, 2013). Mensen vormen zich een indruk van een situatie, hebben daarnaast algemene noties over wat moreel juist is, en die weten ze (meer of minder bewust) te combineren tot een moreel oordeel over wat in deze situatie moreel juist is. Maar zoals Jonsen en Toulmin in hun klassieker *The Abuse of Casuistry* benadrukken, "no rule can be entirely self-interpreting" (1989, p. 8). Regels leggen hun eigen toepassing niet volledig vast. En, zo zou je daar aan kunnen toevoegen, 'no situation is self-describing' (zie ook McLaren, 2006). Anders dan veel AI-systemen impliceren, bestaat morele oordeelsvorming niet alleen uit redeneren, maar ook uit het *interpreteren* van zowel de situatie als abstracte regels of principes. En dat interpretatieve werk zou wel eens het uitdagendste, maar ook het belangrijkste onderdeel van morele oordeelsvorming kunnen zijn.

Interpretatief werk zou wel eens het uitdagendste, maar ook het belangrijkste onderdeel van morele oordeelsvorming kunnen zijn

Dat begint al bij de beschrijving van een casus: wat behoort daar wel en niet toe? Ethiekdocenten en klinisch ethici weten maar al te goed dat je casuïstiek heel plat kunt beschrijven. Een voorbeeld: een patiënt wil X, maar dokter denkt dat Y beter voor hem is. Doorvragen brengt vaak details van de situatie aan het licht die de inbrenger in eerste instantie over het hoofd zag of niet relevant vond,

bijvoorbeeld over de voorgeschiedenis van de arts-patiëntrelatie. Morele perceptie (wat is hier aan de hand en wat is 'moreel saillant' daarin?) is cruciaal. Zulke details helpen om te bepalen welke regels of principes relevant zijn. Gaat het hier ook niet over rechtvaardigheid? En ook helpen ze om die regels te nuanceren of te specificeren: wat betekent het *in deze situatie* om autonomie te respecteren en tegelijkertijd zorgzaam en rechtvaardig te zijn? Al dat interpretatieve werk kan ook helpen nieuwe, vaak meer genuanceerde handelingsopties te bedenken, die men in eerste instantie niet zag.

Deze interpretatieve processen kunnen, maar hoeven niet noodzakelijk bewust plaats te vinden. Een belangrijke rol van ethici in onderwijs en moreel beraad is om ze alsnog expliciet, en daarmee toegankelijk voor reflectie en discussie te maken. Mijn voorlopige conclusie is dan ook dat op AI gebaseerde morele adviseurs vooral goed werk zouden kunnen doen door vragen te stellen, in plaats van antwoorden te geven (zie voor een voorbeeld in deze richting Lara & Deckers, 2020).

Kwaliteit van oordeelsvorming

Mijn eerste punt is dus dat veel AI-systemen voor morele oordeelsvorming het interpretatieve aspect veronachtzamen. Maar misschien is dat helemaal niet erg, *als* dat tot betere oordelen leidt. Het zijn de resultaten die tellen, niet de weg daar naartoe, zou je kunnen redeneren. Deze redenering zie je ook bij AI-toepassingen in de klinische setting. Inmiddels zijn er heel wat studies die laten zien dat AI-systemen beter diagnosticeren of beter voorspellen dan bijvoorbeeld pathologen of radiologen. Zouden we, analoog hieraan, niet moeten onderzoeken wie betere morele oordelen velt, mens of AI-systeem, los van hoe die oordelen geveld worden?

Daarmee komen we op een kwestie waar we al aan raakten bij de bespreking van bottom-up programmeren. Zulk onderzoek vergt een uitkomstmaat waarop je de prestaties van mens en AI kunt vergelijken. Maar het zal in de ethiek bepaald niet eenvoudig zijn om het eens te worden over zo'n uitkomstmaat. Wat zou hier de 'gouden standaard' moeten zijn? Sommige auteurs suggereren dat de voorspelde uitkomst (of liever: evaluatie) vergeleken kan worden met de werkelijke uitkomst (of evaluatie) van de geadviseerde handeling (Wallach et al., 2010), maar dat roept de vraag op of een advies dat feitelijk geaccepteerd wordt, ook in alle gevallen moreel acceptabel is. En, gaat zo'n vergelijking er niet te makkelijk van uit dat elk moreel probleem één beste oplossing heeft?

Toch hoeven we niet in relativisme te vervallen. Ook zonder te veron-

derstellen dat er maar één beste oplossing is, denk ik dat ethici het wel eens zouden kunnen worden over wat betere of slechtere morele oordelen zijn. We beoordelen studenten na ethiekonderwijs vaak meer op hun denkproces dan op de uitkomst van hun casusanalyse. En ook moreel beraad leidt bij deelnemers en ethici vaak tot de overtuiging dat het uiteindelijke besluit beter is. Maar dat kwaliteitsoordeel baseren we eerder op het rijke palet van overwegingen dat vanuit alle denkbare perspectieven in beschouwing is genomen, en op de systematische beoordeling van de relevantie en geldigheid van de overwegingen. Voor de kwaliteit van morele oordeelsvorming lijkt dus niet alleen de uitkomst zelf, maar vooral de organisatie van het interpretatieve proces bepalend. Dat proces hoeft niet altijd vooraf te gaan aan het oordeel, maar het dient op zijn minst tot uitdrukking te komen in de rechtvaardiging van het oordeel.

Met andere woorden: kwalitatief goede ethische oordeelsvorming vergt ook een bepaald proces en/of het vermogen dat proces te expliciteren en onderbouwen. Maar dat is nu juist wat bij veel AI-systemen ontbreekt of onzichtbaar is. Ook op dit punt zou een vorm van AI die vragen stelt behulpzamer zijn dan eentje die antwoorden genereert.

Conclusie

Resumerend denk ik dat de AI-systemen voor morele oordeelsvorming vaak worden gevoed door zeer simpele opvattingen van wat morele oordeelsvorming is, en hoe je de kwaliteit ervan verbetert. Bij de ontwikkeling van AI voor morele oordeelsvorming moet het streven niet zijn de enige goede 'oplossing' te vinden van morele puzzels. Dat morele oordeelsvorming vaak complex interpretatief werk vereist, wordt makkelijk over het hoofd gezien en verhoudt zich ook slecht tot de algoritmische structuur van AI. Daarmee bewijzen zulke AI-systemen de ethiek, maar vooral ook degenen die ze adviseren, geen dienst. Om de kwaliteit van morele oordeelsvorming te bevorderen, moeten vragen worden gesteld in plaats van het stellen van antwoorden. Dat geldt zowel voor AI als voor menselijke ethici.

Dat morele oordeelsvorming vaak complex interpretatief werk vereist, wordt makkelijk over het hoofd gezien

Prof. Marianne Boenink is hoogleraar Ethiek van de gezondheidszorg bij de afdeling IQ Healthcare van het Radboudumc. Haar onderzoek richt zich met name op ethische en filosofische vragen rondom nieuwe biomedische technologie.

Literatuur

Gabriels, K. (2021). *Siri, wat adviseer jij? Over het gebruik van kunstmatige intelligentie voor morele oordeelsvorming*. Preadvies Nederlandse Vereniging voor Bio-ethiek.

Haidt, J. (2013). *The righteous mind . Why good people are divided by politics and religion.* Londen: Penguin Group.

Jonsen, A. R. & Toulmin, S. (1989). *The abuse of casuistry*. Berkeley/Los Angeles: University of California Press.

Lara, F. & Deckers, J. (2020). Artificial intelligence as a socratic assistant for moral enhancement. *Neuroethics*, 13(3), pp. 275-287.

McLaren, B. M. (2006). Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE Intelligent Systems*, 21(4), pp. 29-37.

Topol, E. (2019). *Deep Medicine. How artificial intelligence can make healthcare human again.* New York: Basic Books.

Wallach, W., Franklin, S. & Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science*, 2(3), pp. 454-485.

Thema: De waarde van AI voor de bio-ethiek Podium voor Bio-ethiek · jaargang 28 · nr. 3 · 2021

Helpt artificiële intelligentie bij ingrijpende besluitvorming door dokters?

Jan Hulscher, Elisabeth M.W. Kooi, Caspar Chorus, Annebel ten Broeke en Els Maeckelberghe

Een van de grootste uitdagingen van de invoering van AI in de geneeskunde is het combineren van de *techniek* van de AI met de *praktische wijsheid* van de geneeskunst. In dit essay benoemen we, mede aan de hand van een fictieve casus, enkele van deze uitdagingen met betrekking tot de rol van AI bij het maken van keuzes in de (patiënten)zorg, zonder overigens op alle vragen een antwoord te willen geven. Hierbij sluiten we aan op het preadvies aangaande het gebruik van kunstmatige intelligentie voor morele oordeelsvorming, zoals opgesteld door collega Gabriels (2021). We introduceren een door ons ontwikkelde nieuwe vorm van 'zwakke' AI genaamd Behavioural Artificial Intelligence Technology (BAIT). BAIT kan, zoals we in onderstaand essay betogen, op enkele van deze uitdagingen een alternatief antwoord vormen voor zowel regelgestuurde als datagestuurde AI. BAIT is bij uitstek toegerust om het morele kompas van dokter en patiënt te versterken.

De hypothetische maar al te realistische casus Yente

Het is drie uur 's nachts. U wordt als kinderchirurg geroepen bij Yente, een veel te vroeg geboren baby, nu twee weken oud en 850 gram. Yente heeft een ernstige darmontsteking, 'necrotiserende enterocolitis' (NEC). Zonder operatie komt zij te overlijden. Als u haar wel opereert heeft Yente ongeveer 50% kans om te overleven. Maar: overleving gaat in zo'n 70% van de gevallen gepaard met langetermijn complicaties, zoals ontwikkelingsachterstand of darmproblemen. Ouders twijfelen en vragen uw mening. Adviseert u bij dit dilemma een operatie of palliatief beleid?

Hoe kan AI helpen bij deze keuze?

Nu u geadviseerd heeft voor operatie of juist voor palliatief beleid: op basis waarvan heeft u dit advies gegeven? Dat is nog niet zo eenvoudig. Ook als je dokters vraagt waar zij hun adviezen op baseren blijken deze overwegingen vaak moeilijk te expliciteren. Om de factoren te onderzoeken die aan onze keuzes bij chirurgische NEC ten grondslag liggen, hebben wij in het Universitair Medisch Centrum Groningen (UMCG) een keuze-experiment uitgevoerd met neonatalogen en kinderchirurgen, gebruikmakend van BAIT. De methodologie van dat onderzoek werd elders uitgebreid beschreven (Ten Broeke, 2021).

Kort samengevat hebben we allereerst aan vier experts gevraagd welke factoren meewegen bij de beslissing om wel of niet te opereren bij een veel te vroeg geboren baby met chirurgische NEC. Dit zijn factoren zoals zwangerschapsduur, bijkomende ziekten, verwachte neurologische uitkomst, en draagkracht van de ouders. Vervolgens zijn 35 fictieve casus gebaseerd op voornoemde factoren, voorgelegd aan 15 UMCG neonatologen en kinderchirurgen met de vraag om in die gevallen wel of niet te opereren. Deze 'papieren patiënten' zijn met statistische technieken zo samengesteld, dat de keuzes van de experts zoveel mogelijk informatie geven over het gewicht dat elke factor heeft bij de betreffende keuzes. Zo is geanalyseerd wat de belangrijkste factoren zijn die onze adviezen bepalen en welk gewicht elke factor heeft. Vervolgens is een model ontwikkeld dat bij elke nieuwe casus, waarbij de eerder genoemde factoren ingevuld worden met 'real life' waarden uit de nieuwe casus, kan voorspellen hoeveel procent van de neonatologen/kinderchirurgen zou adviseren voor of tegen operatie, en welke factoren in welke mate daarbij doorslaggevend zijn (Ten Broeke, 2021). Hierbij is niet onderzocht wat de redenen zijn waarom de deelnemers juist deze factoren het belangrijkst vonden.

De vraag is niet of, maar hoe we ons als dokter en als patiënt moeten leren verhouden tot AI

Dokter AI?

Hoewel BAIT een veelbelovende innovatie lijkt die in staat is om impliciete kennis en intuïtie te expliciteren en onze keuzes te ondersteunen, moeten we goed nadenken over het gebruik van dergelijke hulpmiddelen, gebaseerd op artificiële intelligentie, in de geneeskunde. Zeker als dit raakt aan moreel geladen besluitvorming.

Artificiële intelligentie (AI) 'is here to stay'. De vraag is niet of, maar hoe we ons als dokter en als patiënt moeten leren verhouden tot AI. AI kan immers een

rol spelen in het gehele terrein van de geneeskunde, van preventie via diagnostiek naar behandeling. Maar wat blijft er over van de autonomie van de dokter? En mocht deze afnemen, hoe erg is dat in een tijd waarin de mens vergeleken met de computer als 'een slechte informatieverwerker, een gemankeerde morele beoordelaar en ondermaatse morele actor' wordt gezien (vrij naar Gabriels, 2021)?

Door de vraag zo te stellen komt meteen een van de belangrijkste vooronderstellingen naar voren: dat de computer, niet gehinderd door (on)bewuste vooroordelen, sneller en rationeler beslissingen zou kunnen nemen dan de dokter, hetgeen de patiënt alleen maar ten goede kan komen (Grote, 2020). Wat gebeurt er als dokter en computer het niet eens zijn? Als de dokter de computer niet volgt en er treedt een complicatie op, wie is dan verantwoordelijk, en wellicht juridisch en financieel aansprakelijk? De dokter? Het ziekenhuis? Het IT-bedrijf?

Maar is de computer wel de epistemisch meerdere van de dokters, en hoe kunnen we de epistemische positie van AI onderzoeken? Voor de gemiddelde dokter vormt de data-gedreven AI op basis van machine learning een black box. De - meestal historische - dataset moet groot zijn, en dit, in combinatie met niet-transparante algoritmen, kan leiden tot allerlei niet direct zichtbare bias. Daar tegenover staan de expliciete beslisregels die we in kennisgedreven AI terugvinden. Maar weten we wel hoe we zelf denken als dokter? Is de geneeskundeopleiding er niet juist op gebaseerd om het denken via regels (zoals co-assistenten vaak doen) te doen overgaan in het denken gebaseerd op patronen (zoals medisch specialisten geacht worden te doen)? En zijn deze patronen wel zo transparant als we zelf denken? Om terug te komen op baby Yente: weten we zelf wel waar we onze adviezen op baseren?

Weten we wel hoe we zelf denken als dokter?

BAIT heeft, anders dan datagestuurde AI, geen grote dataset nodig, en is transparant en intuïtief. De in het model gebruikte factoren zijn immers door expertkeuzes tot stand gekomen en gevalideerd. Sterker nog, de techniek kan juist gebruikt worden om via een keuze-experiment, zoals hierboven beschreven, de belangrijkste factoren uit ons keuzegedrag te destilleren en te toetsen aan de dagelijkse praktijk. Er is dus geen directe explicitering van kennis nodig, zoals in klassieke kennisgedreven AI. Daarmee maakt BAIT de dokters bewust van de factoren waar ze hun adviezen op baseren. Door een eventueel (computer) advies te omkleden met redenen waarom de computer dat advies geeft, zal zo'n advies wellicht ook eenvoudiger geaccepteerd worden (Klincewicz, 2016). Het is goed

hier op te merken dat BAIT weliswaar inzicht geeft in wélke factoren van belang zijn, maar niet in waarom die van belang zijn.

Maar volgt de computer in dit model dan niet simpelweg de keuze van de meerderheid van het volk, in dit geval van de dokters die mee hebben gedaan aan het keuze-experiment om het systeem te trainen? Ons systeem probeert dit te ondervangen door alleen een uitspraak te doen over hoeveel procent van alle dokters waarschijnlijk zou kiezen voor optie A of optie B. Het blijft aan de individuele dokter om samen met de patiënt een besluit te nemen, maar nu met de kennis van wat de beroepsgroep onder de huidige omstandigheden zou doen. Daarnaast kan inzichtelijk gemaakt worden op welke specifieke factoren de computer het advies baseert. Desalniettemin is dit ook voor BAIT een uitdaging waar we ons bewust van moeten zijn, want ook dokters zijn uiteraard gevoelig voor de druk van de meerderheid. Inmiddels is in Nederland een landelijk onderzoek gaande waarbij met behulp van BAIT in alle neonatologische centra onderzocht wordt welke factoren het handelen van neonatologen en kinderchirurgen bepalen, waarbij ook gekeken zal worden naar verschillen tussen specialismen en centra. Ook zal ons model met de komst van nieuwe technologische uitdagingen en nieuwe ethische inzichten opnieuw gevalideerd moeten worden.

BAIT en de arts-patiënt relatie

Niet alleen de dokter kan worden beïnvloed, maar ook de patiënt. Als de computer, net zo als vroeger de dokter, als alwetend wordt gezien, is het een kleine stap van ‘doctor knows best’ naar ‘computer knows best’. In het gunstigste geval leidt dit tot herintroductie van paternalisme door dokters, een paternalisme dat we juist kwijt wilden in deze tijd van samen beslissen en waarde-gedreven zorg. BAIT kan hier deels aan tegemoet komen door patiënten c.q. ouders een vergelijkbaar keuze-experiment met verschillende casussen te laten uitvoeren, om zo de voor hen belangrijkste factoren te definiëren. Dit geeft inzicht in wat patiënten de belangrijkste factoren vinden, en versterkt daarmee de autonomie. Door het vergroten van het inzicht in de drijfveren van zowel arts als patiënt versterkt BAIT de mogelijkheid tot samen beslissen.

Door het vergroten van het inzicht in de drijfveren van zowel arts als patiënt versterkt BAIT de mogelijkheid tot samen beslissen

En dat leidt tot wellicht het belangrijkste vraagstuk: wat gebeurt er tijdens

het consult? Hoe verhoudt AI zich tot de geneeskunst? Het gebruik van de computer is fundamenteel anders dan het gebruik van bijvoorbeeld de stethoscoop. Een stethoscoop leidt tot (fysieke) nabijheid, een computer schept afstand. Tot nu toe mist de computer elke vorm van menselijkheid en lichamelijkheid, wijsheid en compassie die de dokter vaak wel kan bieden. Hoe sluit een algoritme aan bij het narratief van de patiënt? Is dat überhaupt mogelijk? Hoe verandert een algoritme het vertrouwen, zo inherent aan de dokter-patiënt relatie? Blijft het algoritme 'gewoon' een hulpmiddel van de dokter of wordt het steeds meer een derde stem in de dokter-patiënt relatie?

(B)AI(T) is here to stay

Wij verwachten dat AI een derde stem in de dokter-patiënt relatie zal worden. Hierbij is het van belang de autonomie van dokter en patiënt binnen deze relatie te versterken en niet te verzwakken. We zullen moeten leren om de techniek van de AI te combineren met de praktische wijsheid van de geneeskunst. BAIT biedt daartoe een nieuwe techniek die fundamenteel afwijkt van zowel regelgestuurde AI als (big) datagestuurde AI. Het is een vorm van 'zwakke' AI, waarbij de mens alleen een beperkte doelstelling oplegt aan de techniek, binnen een vooraf gedefinieerde context. BAIT stelt dokter en patiënt via een slim computerprogramma in staat om binnen deze context de morele intuïtie te expliciteren. BAIT kan daarna gebruikt worden als applicatie in de dagelijkse praktijk, als keuzehulp voor dokter en patiënt om gezamenlijk inzicht te krijgen in de beweegredenen. Het model kan zo ingevoegd worden binnen de narratieve context van het consult. BAIT is niet in staat de morele normen en criteria zelf bij te stellen, daarvoor is en blijft menselijke input noodzakelijk, hetgeen de auteurs ook zeer wenselijk lijkt. Daarmee zouden we BAIT kunnen classificeren als een 'artificial moral agent' klasse 2 (Gabriels, 2021). De komende jaren zal BAIT verder getoetst en gevalideerd worden, zowel als onderzoeksmethode om de factoren die onze keuzes bepalen inzichtelijk te krijgen en als keuzehulp voor dokter en patiënt. Zodat we, dokter en patiënt samen, met een nog beter afgesteld moreel kompas ingrijpende beslissingen kunnen nemen, bijvoorbeeld maar niet uitsluitend, aangaande kinderen als Yente.

Prof. Dr. Jan B.F. Hulscher is kinderchirurg in het UMCG.

Dr. Elisabeth M.W. Kooi is kinderarts-neonatoloog in het UMCG.

Prof. Dr. Ir. Caspar G. Chorus is hoogleraar Choice Behavior Modeling, hoofd van de afdeling Engineering Systems and Services aan de TU Delft en oprichter van Councyl.

Annebel ten Broeke, MSc, is data-analist bij Councyl.

Dr. Els L.M. Maeckelberghe is universitair hoofddocent Medische Ethiek en Onderzoeksethiek aan de RUG/UMCG.

Literatuur

Arnold, M.H. (2021). Teasing out artificial intelligence in medicine: an ethical critique of artificial intelligence and machine learning in medicine. *Bioethical Inquiry,* 18, pp. 121-139.

Gabriels, K. (2021). *Siri, wat adviseer jij? Over het gebruik van kunstmatige intelligentie voor morele oordeelsvorming*. Preadvies Nederlandse Vereniging voor Bio-ethiek.

Grote, T., Berends P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, 46, pp. 205-211.

Klincewicz, M. (2016). Artificial intelligence as a means to moral enhancement. *Studies in Logic, Grammar and Rhetoric,* 48, pp. 171-187.

Ten Broeke, A., Hulscher, J.B.F., Heyning, N., Kooi, E.M.W., Chorus, C.G. (2021). BAIT: a new medical decision support technology based on discrete choice theory. *Medical Decision Making,* 41, pp. 614-609.

Menselijke leiding is cruciaal bij AI-systemen die moreel redeneren ondersteunen

Sophie van Baalen en Linda Kool

In dit artikel betogen we dat AI en de mens elkaar kunnen aanvullen in morele oordeelsvorming. AI kan menselijke intelligentie ondersteunen bij het verzamelen en verwerken van informatie en het inzichtelijk maken van argumenten, waarden en principes. Maar er zijn menselijke vermogens nodig om een oordeel te vellen over concrete situaties in de echte wereld. Daarvoor blijft het cruciaal dat mensen de leiding hebben bij het ontwikkelen en toepassen van *artificial moral agents*.

In haar preadvies onderzoekt Katleen Gabriels de vraag in hoeverre AI ingezet kan worden bij morele oordeelsvorming. Dit doet ze aan de hand van vier typen *Artificial Moral Agents* (AMA-1 tot en met AMA-4) die op een continuüm liggen van oplopende complexiteit en autonomie (Gabriels, 2021). Gabriels concludeert dat morele besluitvorming, in elk geval vooralsnog, bij de mens zal moeten liggen. Wij kunnen ons in deze conclusie vinden en zien net als Gabriels dat AI morele oordeelsvorming zou kunnen ondersteunen daar waar menselijke oordeelsvorming tekortschiet. In dit artikel betogen wij welke eisen dit stelt aan ondersteuning van morele oordeelsvorming door AI. Dit gaat zowel om het ontwerp als de context waarin AI wordt gebruikt.

Morele oordeelsvorming

Onder morele oordeelsvorming verstaan we oordelen waarvan de uitkomst *moreel geladen* is. Dat wil zeggen dat er ethische waarden of principes in het spel zijn. Om tot een moreel oordeel te komen, is vaak een afweging noodzakelijk van waarden, argumenten, ethische principes, geldende regels en wetten, en beschikbare gegevens. Omdat het meewegen van verschillende waarden of

principes kan leiden tot verschillende oordelen, is het 'juiste' oordeel niet evident of objectief vast te stellen. Denk aan een zelfrijdende auto die moet kiezen tussen een aanrijding met een oude man of met een zwangere vrouw. Daarnaast is een moreel oordeel afhankelijk van de oordelaar.

Omdat het meewegen van verschillende waarden of principes kan leiden tot verschillende oordelen, is het 'juiste' oordeel niet evident of objectief vast te stellen

Menselijke oordelaars hebben beperkingen. Ze hebben niet alle informatie tot hun beschikking, of kunnen die niet volledig overzien. Ze zijn wel eens moe of chagrijnig. Ze hebben soms (impliciete) vooroordelen. Volgens Giubilini en Savulescu (2018) zijn mensen *suboptimale informatieverwerkers*, omdat ze niet in staat zijn alle relevante informatie te verwerken. Ze zijn daarnaast *suboptimale morele oordelaars,* omdat ze zich bij het vormen van hun oordeel niet altijd houden aan hun eigen morele principes. Ook zijn ze *suboptimale morele actoren,* omdat ze, zelfs als ze tot een (volgens hun eigen principes) 'goed' moreel oordeel komen, daar niet altijd naar handelen. Kunstmatige intelligentie zou hierbij kunnen helpen, omdat een AI-systeem over een grote hoeveelheid informatie kan beschikken, die snel en efficiënt kan verwerken, en tot een consistent oordeel kan komen, volgens vastgestelde principes.

De ontwikkeling van een kunstmatige morele oordelaar

In haar preadvies laat Gabriels zien dat er haken en ogen zitten aan de ontwikkeling van dergelijke AMA's. Op basis van welke data en met welke normatieve kaders moeten ze bijvoorbeeld worden ontwikkeld? Daarnaast is het onduidelijk of efficiëntere, snellere en meer consistente oordelen ook betere oordelen zijn. Daarbij komt dat ook algoritmen kunnen discrimineren (voor de verschillende manieren waarop vooroordelen in systemen kunnen 'sluipen', zie bijvoorbeeld: Barocas & Selbst, 2016). Die problemen kunnen vrij fundamenteel zijn. Zelfs wanneer een systeemontwikkelaar zich bewust is van het risico op discriminatie, kan het zeer lastig zijn discriminatie te voorkomen. AI-koploper Amazon lukte het bijvoorbeeld niet de bias uit zijn automatische sollicitatiesysteem te halen.

Een AMA-4 is het meest complexe en autonome type AMA. Het is een automatische morele actor die volledig zelfstandig kan opereren, in staat is tot (bewuste) reflectie op het eigen gedrag en daarom ook moreel verantwoordelijk is. Zo'n systeem is volgens ons voorlopig niet haalbaar. De meeste moderne AI-

systemen zijn lerende systemen die leren door patronen te herkennen in grote hoeveelheden data, op basis van statistische technieken, zonder dat ze expliciete regels krijgen om te volgen. Om een AI-systeem te trainen voor morele oordeelsvorming, moet het gevoed worden met grote hoeveelheden data over 'goede' of 'foute' oordelen. Maar, zoals eerder genoemd, is bij morele oordelen de gewenste uitkomst vaak niet (eensgezind) vast te stellen. Daarbij komt dat aangeven of een moreel oordeel goed of fout is, nog niets zegt over de rechtvaardiging van dit oordeel. Dit maakt het moeilijk om een AI-systeem te laten leren wat een moreel oordeel is, en hoe dat wordt gerechtvaardigd. Omdat elke situatie die om een moreel oordeel vraagt anders is, zijn er voor het vellen van oordelen over concrete situaties in de echte wereld menselijke vermogens nodig, zowel cognitieve als empathische. Deze kunnen niet gevat worden in algoritmisch of statistische redeneervormen zoals AI-systemen gebruiken (Van Baalen et al., 2021). Denk aan het benoemen en interpreteren van de morele vraagstelling, het afwegen van principes of waarden die met elkaar in conflict zijn, en het contextualiseren en integreren van de beschikbare informatie. Het is daarom realistischer om na te denken over hoe menselijke en kunstmatige intelligentie elkaar kunnen aanvullen. Zodat de competenties van beide optimaal tot hun recht komen.

Voor het vellen van oordelen over concrete situaties in de echte wereld zijn menselijke vermogen nodig

Van 'human-in-the-loop' naar 'human-in-the-lead'

Het andere geavanceerde type AMA dat Gabriels beschrijft (AMA-3) lijkt daarvoor haalbaarder. Zo'n systeem is technisch nog niet mogelijk, en zou zeer geavanceerd moeten zijn op het gebied van automatische taalherkenning, om zo wetenschappelijke literatuur en andere bronnen te doorzoeken. Daarnaast zou het veel kennis moeten hebben van regelgeving, ethische theorieën en principes en argumentatieleer. Op die gebieden verwachten we wel vooruitgang, maar het blijft afwachten of die vooruitgang voldoende zal zijn voor deze toepassing. Belangrijk is dat zo'n AI-systeem niet wordt ontwikkeld en ingezet om de gebruiker te *adviseren* of *overtuigen* om tot een bepaald oordeel te komen, waarbij mensen vooral supervisie houden ('human-in-the-loop'). In plaats daarvan zou het AI-systeem de gebruiker in verschillende stappen van het redeneerproces moeten voorzien van beschikbare en gewenste ondersteuning, waarbij de gebruiker zelf tot een

oordeel komt ('human-in-the-lead'). Het AI-systeem moet daartoe antwoord geven op verschillende redeneervragen, zoals:

- Welke handelingsmogelijkheden bestaan er in deze situatie?
- Welke argumenten zijn er voor en tegen een bepaalde handelingsmogelijkheid? Op basis van welke regels of principes?

Het systeem kan een deel van deze vragen beantwoorden, zonder daarbij aan te geven welke handelingsopties of argumenten de beste zijn. Dat wil zeggen, zonder een oordeel te vellen, op basis van vooraf geprogrammeerde regels (zoals in regelgestuurde of top-down AI), bijvoorbeeld uit de argumentatieleer en/of uit ethische theorieën. Door het systeem te vragen naar meerdere uitwerkingen van verschillende typen theorieën (bijvoorbeeld deontologie en consequentialisme) in een specifieke situatie, kan een menselijke oordelaar zien tot welke uitkomsten verschillende theorieën kunnen leiden. Vervolgens kan hij of zij zelf afwegen welke het beste past bij een specifieke context en die uitwerkingen gebruiken in het vellen van een oordeel. Andere vragen worden beantwoord op basis van patroonherkenning van eerdere situaties, zoals bij datagestuurde of bottom-up AI. In dat geval wordt een AI-systeem gevoed met een grote hoeveelheid morele casussen. Het gaat daarbij niet om het oordeel goed of fout, maar om de gemaakte afweging. En om de argumenten, waarden en principes die daarbij zijn gevolgd. Dit alles moet concreet worden gemaakt in de set 'trainingsdata'. Wellicht kunnen zelflerende systemen hier in de toekomst deels zelf toe komen, zoals Alpha Go zelf de spelregels kon afleiden uit een groot aantal voorbeelden. Maar dat lijkt voorlopig nog toekomstmuziek omdat morele oordeelsvorming extreem complex is – de 'probleemruimte' is enorm groot (Schermer et al., 2020).

Het blijft cruciaal dat mensen, nu en in de toekomst, de leiding hebben

Daarom blijft het cruciaal dat mensen, nu en in de toekomst, de leiding hebben: niet 'human-in-the-loop' maar 'human-in-the-lead'. Dat geldt niet alleen voor het ontwikkelen van een AMA door het van de juiste regels en trainingsdata te voorzien en het resulterende algoritme te controleren. Het geldt ook voor het toepassen van een AMA, door zelf de vragen te stellen, en de antwoorden van het systeem te wegen en in te passen in de context van de huidige situatie.

Van moreel oordeel naar morele redenering

We pleiten er dus voor om een AMA zo te ontwerpen dat deze het proces van

morele oordeelsvorming door een mens ondersteunt om de eerder genoemde menselijke tekortkomingen aan te vullen. Neem bijvoorbeeld een oordeel van een medisch-ethische toetsingscommissie (METC) over het wel of niet toestaan van een bepaald onderzoek met mensen. Hierbij moeten veel verschillende aspecten worden meegewogen over onder meer risico's, resultaten, regelgeving en ethische principes.

Om risico's in te schatten, is medisch-wetenschappelijke kennis nodig. Hoe werkt bijvoorbeeld het medicijn dat wordt onderzocht en wat is er al bekend over de bijwerkingen op korte of lange termijn in de beoogde onderzoekspopulatie? Een METC weegt dit af tegen de (mogelijke) voordelen van het onderzoek. Een AI-systeem kan dit ondersteunen, bijvoorbeeld door bestaand wetenschappelijk onderzoek te doorzoeken en de uitkomsten daarvan onder elkaar te zetten. Zo wordt de eerste menselijke beperking, *suboptimale informatieverwerking*, aangevuld. Maar het toepassen van informatie in een moreel besluit vereist ook menselijke capaciteiten. Bijvoorbeeld het stellen van de juiste informatievragen, het op relevantie beoordelen van de beschikbare informatie en het interpreteren, integreren en contextualiseren van de beschikbare informatie.

Van belang bij de afweging die een METC moet maken, zijn ook ethische principes, vastgelegd in verklaringen en codes, zoals: respect voor het individu, het recht op zelfbeschikking en het belang van *informed consent* door deelnemers aan onderzoek. In sommige gevallen is het duidelijk of en hoe hieraan voldaan wordt. Maar dat is niet altijd het geval, bijvoorbeeld bij onderzoek met wilsonbekwamen of kinderen, die zelf niet voldoende in staat worden geacht om deze afweging voor zichzelf te maken. Weegt het doel van het onderzoek in dat geval op tegen het gebrek aan *informed consent*? AI kan deze overweging ondersteunen door expliciet te maken welke ethische principes in een casus gelden, en of verschillende ethische benaderingen gerelateerd zijn aan verschillende ethische principes. Een AI-systeem kan beschikken over de afwegingen en uitkomsten van grote aantallen eerdere gevallen of voorbeelden, en vervolgens op basis van overeenkomsten en verschillen met deze casussen laten zien hoe verschillende ethische principes zich tot de huidige casus verhouden. Hiermee wordt een menselijke oordelaar ondersteund om te handelen volgens specifieke principes, door snel en overzichtelijk inzicht te geven in de ethische principes die mogelijk van toepassing zijn, plus hoe deze in vergelijkbare gevallen kunnen worden toegepast. Het AI-systeem maakt deze informatie beschikbaar en inzichtelijk voor de

gebruiker. Op basis daarvan kan hij of zij een oordeel vormen.

Met het inzichtelijk maken van de beschikbare informatie, mogelijke afwegingen, en ethische principes, biedt een AI-systeem ondersteuning aan morele oordeelsvorming door de mens, zonder dat het systeem zelf tot een oordeel of advies komt. De menselijke oordelaar houdt de leiding. Zo ontstaat er geen artificieel moreel adviseur maar een artificieel moreel redeneer-ondersteunend systeem.

Sophie van Baalen is onderzoeker bij het Rathenau Instituut.

Linda Kool is themacoördinator Digitale Samenleving bij het Rathenau Instituut.

Literatuur

Barocas, S.& Selbst, A. (2016). Big data's disparate impact. *Californian Law Review,* 104, pp. 671-732. DOI: http://dx.doi.org/10.15779/Z38BG3

Gabriels, K. (2021). *Siri, wat adviseer jij? Over het gebruik van kunstmatige intelligentie voor morele oordeelsvorming*. Preadvies Nederlandse Vereniging voor Bio-ethiek.

Giubilini, A. & Savulescu, J. (2017). The artificial moral advisor. The "ideal observer" meets artificial intelligence. *Philosophy & Technology,* 31,pp. 169-188. DOI: 10.1007/s13347-017-0285-z

Van Baalen, S., Boon, M. & Verhoef, P. (2021). From clinical decision support to clinical reasoning support systems. *Journal of Evaluation in Clinical Practice,* 27, pp. 520-528. DOI: 10.1111/jep.13541

Schermer, B., van Ham, J. & Falkena, K.W., (2020). *Onvoorziene effecten van zelflerende algoritmen*. Amsterdam: Considerati.

AI en morele oordeelsvorming: van principes naar het vormgeven van ethische AI-praktijken

Rik Wehrens, Sydney Howe, Esra Demir, Kostina Prifti, Klaus Heine & Evert Stamhuis

Het debat over de ethiek van kunstmatige intelligentie (hierna: AI) heeft zich de afgelopen jaren geconcentreerd op het formuleren van generieke ethische principes en daarop gebaseerde toetsingsinstrumenten. In deze bijdrage betogen we dat deze specifieke, principiële benadering van AI-ethiek nog niet afdoende is voor het vormgeven van ethische AI-*praktijken*. We bespreken twee redenen waarom dat het geval is. We eindigen met een pleidooi voor een meer empirisch en etnografisch gedreven onderzoeksprogramma naar moreel handelen in concrete professionele praktijken waarin AI-systemen een gunstige bijdrage zouden moeten leveren.

Inleiding

Met het toenemende gebruik van AI-toepassingen in allerhande aspecten van het leven, waaronder de gezondheidszorg en de wetshandhaving, is het gesprek over criteria voor ethisch-juridisch aanvaardbare AI de afgelopen jaren op verschillende plekken gevoerd. Zo heeft de Europese Commissie de generieke principes voor 'Trustworthy AI' gepubliceerd, die nu ook omgezet zijn in een conceptregeling voor AI, inclusief een mechanisme voor certificering (COM, 2020; COM, 2021). De Europese instellingen vormen niet het enige gremium dat aandacht heeft voor het formuleren van de juiste ethische principes waaraan AI zou moeten voldoen. Talloze bedrijven, professionele associaties en adviesgroepen richtten zich de afgelopen jaren op het formuleren van ethische principes of richtlijnen. Een recente overzichtsstudie identificeerde maar liefst 84 documenten sinds 2016

(Jobin et al., 2019), en de productie heeft sindsdien niet stilgestaan.

Het denkwerk van deze expertgroepen is in meerdere opzichten waardevol; bijvoorbeeld in de erkenning van de cruciale ethische vragen en dilemma's die door AI opgeworpen worden. Ook is het bepalen van de juiste randvoorwaarden voor het ontwikkelen en inzetten van AI een lovenswaardig streven. Kritiek is er echter ook. In een recent artikel betogen Rességuier en Rodrigues (2020) dat de nadruk in AI-ethiek op het formuleren van ethische principes problematisch is omdat ethiek erin gereduceerd wordt tot een substituut voor wetgeving. Dat brengt risico's met zich mee, bijvoorbeeld het risico dat de industrie dergelijke principes kan omarmen zonder juridisch verantwoording te hoeven afleggen als praktijken schadelijk uitpakken.

In deze bijdrage betogen we dat de specifieke invulling die AI-ethiek krijgt – een invulling die we een 'principiële benadering van AI-ethiek' noemen – nog niet afdoende is voor het vormgeven van *ethische AI-praktijken*. We bespreken twee redenen waarom dat het geval is. We eindigen met een pleidooi voor een meer empirisch gedreven onderzoeksprogramma, waarin etnografisch en vergelijkend onderzoek in diverse professionele praktijken in kaart brengt hoe moreel handelen tot stand komt, hoe dergelijke handelingen verantwoord worden en hoe zich dat verhoudt tot concrete normatieve kaders.

Het abstractieniveau van ethische discussies

Een eerste reden waarom de principiële benadering van AI-ethiek op zichzelf genomen niet afdoende is om te komen tot ethische AI-praktijken, heeft te maken met de afstand tussen ethiek en de concrete praktijken waarin ethisch oordelen en handelen vorm moeten krijgen. Een sprekend voorbeeld hiervan is de alomtegenwoordigheid van het fictieve 'trolleyprobleem' bij het bespreken van de ethische vragen rondom zelfrijdende auto's. Daarin wordt vanuit een reeks hypothetische en onwaarschijnlijke situaties besproken welke afwegingen zo'n auto zou moeten maken op het moment dat het maken van slachtoffers onvermijdelijk is. De keuzemogelijkheden zijn geworteld in klassieke ethische theorieën. Vanuit een geprogrammeerde deontologische ethiek komt het AI-systeem tot een ander moreel oordeel dan vanuit een utilitaristisch ethisch raamwerk. Zowel het cre-

De principiële benadering van AI-ethiek is nog niet afdoende voor het vormgeven van ethisch AI-praktijken

eren van dergelijke onrealistische scenario's als het evalueren van de oordeelsvorming aan de hand van ethische theorieën geeft echter weinig inzicht in hoe ethisch handelen concreet vorm zou krijgen in werkelijke praktijksituaties.

Toegegeven, aandacht voor het 'aan de voorkant' inbouwen van ethische criteria en waarden in AI-applicaties heeft de voorkeur boven een achterhaalde 'instrumentele' benadering van technologie, die impliceert dat technologie neutraal is en de ethische en morele vragen pas in de toepassing opdoemen. Tegelijkertijd krijgt morele oordeelsvorming altijd handen en voeten in concrete handelingspraktijken. Het is ook in die handelingspraktijken dat zichtbaar wordt dat morele oordeelsvorming (en de hieruit voortvloeiende acties) meer behelst dan het toepassen van algemene principes of ethische theorieën. Hooguit kunnen computerexperts hieruit enkele algemene kenmerken halen voor de systemen die zij ontwerpen. Maar een ziekenhuis, een bedrijf of een overheidsinstelling komt er niet veel verder mee, zolang deze beginselen niet gecombineerd zijn met componenten uit de eigen handelingspraktijk.

Die praktijk is de ethisch relevante werkvloer. Daar waar de gevolgen voelbaar worden van de inzet van een AI-systeem, zal men de voordelen kunnen incasseren en ook de eventuele schadelijke effecten ondervinden en moeten opvangen. Daar zal men ook de balans moeten bepalen in situaties waarin beginselen niet eenduidig dezelfde kant op wijzen, of er trade-offs aan de orde zijn tussen bijvoorbeeld betrouwbaarheid en snelheid. Vaak zal ook alleen in de praktijk blijken wat door betrokkenen als ethisch problematisch beschouwd wordt en wat niet. De principiële benadering van AI-ethiek kan op hoofdlijnen richting geven aan deze praktische afwegingen, maar ze garandeert op zichzelf genomen geen passende uitkomsten (waarbij 'passend' ook sterk contextafhankelijk is).

Voorbij de individuele afweging

De tweede reden waarom de principiële benadering van AI-ethiek op zichzelf genomen niet toereikend is om te komen tot ethische AI-praktijken, heeft betrekking op de impliciete opvatting van morele oordeelsvorming die erin besloten ligt.

De principiële benadering van AI-ethiek lijkt morele oordeelsvorming vooral voor te stellen als een individuele bezigheid, waarin de beoordelaar langs de weg van een rationele afweging van feiten en argumenten tot een oordeel komt. Ook dat zien we terug in het eerder aangehaalde 'trolleyprobleem'. Ethiek wordt hier in de kern gereduceerd tot een reeks (in eerste instantie individu-

ele) beslissingen, waarin het prevalerende ethische principe uiteindelijk bepaalt wat als een 'ethisch juiste uitkomst' geldt. Morele oordeelsvorming wordt in dit voorbeeld vooral geconstrueerd als een rationeel psychologisch besluitvormingsproces, waarin ethiek de principes en theorieën aanlevert om tot een goede beslissing te komen.

De principiële benadering van AI-ethiek lijkt morele oordeelsvorming vooral voor te stellen als een individuele bezigheid

Deze 'psychologische' modellering van morele oordeelsvorming negeert cruciale inzichten uit disciplines als de sociologie van professies en het wetenschap- en techniekonderzoek (Science & Technology Studies). Ondanks de verschillen tussen deze stromingen benadrukken ze beide hoe (morele) oordeelsvorming ontstaat en verantwoord wordt binnen professionele praktijken en niet een generiek model volgt. Daarin zijn niet alleen richtlijnen (of de toepassing van 'principes') belangrijk. Evenzeer worden praktijken van 'goed handelen' gevormd door socialisatie en de ontwikkeling van onbewuste vormen van kennis en begrip (*tacit knowledge*). Onderdeel daarvan zijn de discretionaire ruimte en inzicht in wanneer de regels juist omzeild of overtreden zouden moeten worden (Collins, 2018; Wallenburg et al., 2019).

Het ethische criterium 'uitlegbaarheid' (*explainability*) kan als goed voorbeeld dienen om de verschillen tussen een psychologisch en een sociologisch georiënteerde invulling van morele oordeelsvorming te verhelderen. Het criterium van 'uitlegbaarheid' wordt vaak gepresenteerd als de oplossing voor het black-box probleem in geavanceerde vormen van *machine learning*, waarbij ook de ontwerpers de uitkomsten van de toepassing niet meer kunnen traceren. Uitlegbare AI in deze betekenis vereist dan transparantie op vele niveaus: van zicht op de principes die de input voor het model zijn geweest tot aan inzicht in hoe het model getraind is en in de voornaamste parameters en waarden, die tot de uiteindelijke uitkomst geleid hebben. De nadruk ligt hier op het zichtbaar maken van alle tussenstappen in het redeneringsproces.

Wanneer we het criterium 'uitlegbaarheid' in de context van morele oordeelsvorming zouden bekijken vanuit een sociologisch georiënteerd perspectief, zien we direct de verschillen met de principiële benadering van AI-ethiek. Een voorbeeld van morele oordeelsvorming in de context van de medische praktijk biedt de vraag wat een goed leven behelst voor een patiënt met meerdere chro-

nische aandoeningen – en welke behandelingen daar het best bij aansluiten. Die discussie krijgt vorm in de conversatie tussen de arts en de patiënt, vanuit de input van collega's, en met behulp van bewijs vanuit de literatuur. 'Uitlegbaarheid' krijgt in deze specifieke praktijk dus veeleer een relationele en interpretatieve betekenis, waarin de patiënt en de arts gezamenlijk hun bedoeling construeren en over de te nemen stappen een (tentatief) besluit nemen, dat ook weer aan verandering onderhevig kan zijn (vgl. Mol, 2008). Het betekent niet dat alle rationele denkstappen van de arts onthuld moeten worden.

Conclusie

Gabriels concludeert aan het eind van haar preadvies dat AI nog niet ver genoeg is ontwikkeld om zelfstandig morele besluiten te kunnen of laten nemen. We delen die conclusie. De complexe ethische afwegingen in het echte leven zijn heel anders dan vereenvoudigde ethische modellen in een gecontroleerde omgeving. Met ethische principes alleen zijn we er niet. Bovendien omvat moreel oordelen volgens ons meer dan een als model te reconstrueren afwegingsproces: het is, in de woorden van Gabriels, een "samenspel van intuïtie en rede" (Gabriels, 2021, p. 13). Een belangrijke aanvulling daarop is dat zowel intuïtie als rede onderdeel zijn van (en vorm krijgen in) professionele ontwikkeling. Denk bijvoorbeeld aan het 'niet pluis'-gevoel bij verpleegkundigen. Daarmee kan moreel oordelen gezien worden als een 'groepsproduct' van professionele praktijken.

Het op een verantwoorde manier inbedden van AI-toepassingen, zowel in zorg, welzijn, bestuur en rechtspraak, vraagt dus om meer dan louter technische oplossingen, formele certificeringswetgeving en het formuleren van ethische principes. We dienen juist te bestuderen hoe morele beslissingen tot stand komen en onderhandeld worden in de dagelijkse praktijken van professionals (zorgverleners, functionarissen, datawetenschappers en andere stakeholders (zoals patiëntgroepen of getroffen burgers). Dit vraagt om een etnografisch en vergelijkend onderzoeksprogramma dat in de praktijk onderzoekt hoe moreel handelen tot stand komt, hoe dergelijke handelingen verantwoord worden en hoe zich dat verhoudt tot concrete normatieve kaders voor die praktijken.

Het inbedden van AI-toepassingen vraagt om meer dan louter technische oplossingen, formele certificeringswetgeving en ethische principes

Wie naar concrete professionele praktijken gaat kijken, kan immers van alles

verwachten als het gaat om het gedrag van en tussen de actoren. We moeten rekening houden met de attitudes en percepties van professionele groepen binnen hun praktijk, hun specifieke taal en manieren van probleemkadering, alsmede hun persoonlijke beroepsethos. Evenzeer moet in beeld komen hoe de verschillende betrekkingen juridisch vormgegeven zijn, met welke rechtsregels en beginselen de belangen van de diverse betrokken personen/organisaties geharmoniseerd zijn en welke flexibiliteit daarin ingebouwd is. Hoe zit het met asymmetrie in macht, kennis en invloed en in welke mate biedt het recht concrete correcties daarop, bijvoorbeeld door middel van zorgplichten en de aansprakelijkheidsverdeling? Even belangrijk is het te kijken naar de percepties en verwachtingen van de personen over wier belangen met behulp van het AI-systeem een oordeel of beslissing gegeven wordt en naar wat er voor die personen op het spel staat.

In de samenleving interacteert een AI-systeem met de reeds aanwezige kenmerken van de professionele praktijk. De expliciete en impliciete normen, doelstellingen en de waardenafwegingen die daaruit volgen zullen het uiteindelijke resultaat meebepalen en beslissend zijn voor de vormgeving, het gebruik en de mogelijk positieve en negatieve effecten van AI-toepassingen. Beginselethiek kan hierin een belangrijke rol spelen, maar zou niet moeten opgaan in een 'scheidsrechter of certificeringsrol'. In plaats van te toetsen of specifieke systemen voldoen aan algemene ethische principes, ligt de meerwaarde van ethiek in diens mogelijkheid tot het *openhouden* van dialoog en reflectie door het voortdurend bevragen van ontwikkelingen en het bieden van nieuwe invalshoeken (vgl. Rességuier & Rodrigues, 2020). Een meer empirische en etnografisch georiënteerde benadering van AI-ethiek kan juist daar goed aan bijdragen. Door gedetailleerd in kaart te brengen hoe morele beslissingen tot stand komen in specifieke praktijken, en door deze inzichten terug te koppelen, ontstaat ruimte om gezamenlijke reflectie in te bouwen en toe te werken naar ethisch verantwoorde AI-praktijken.

Rik Wehrens is universitair docent aan de Erasmus School of Health Policy & Management, Erasmus Universiteit. Hij doet onderzoek naar de sociale en ethische dimensies van data-gedreven technologie in de gezondheidszorg in een nationale en internationale context.

Sydney Howe is promovenda aan de Erasmus School of Health Policy & Management, Erasmus Universiteit. Ze doet onderzoek naar de legitimiteit en kosteneffectiviteit van AI-applicaties in de diagnostisering van huidkanker.

Esra Demir is promovenda aan de Erasmus School of Law, Erasmus Universiteit Rotterdam. Ze doet onderzoek naar eigenaarschap van menselijke biodata.

Kostina Prifti is promovendus aan de Erasmus School of Law, Erasmus Universiteit. Hij doet onderzoek naar 'regulation by design' voor autonome robots die in de gezondheidszorg worden gebruikt.

Klaus Heine is hoogleraar Law and Economics, Erasmus School of Law, Erasmus Universiteit Rotterdam. Hij bekleedt een Jean Monnet leerstoel en is directeur van het Jean Monnet Centre of Excellence op het gebied van Digital Governance.

Evert Stamhuis is hoogleraar Law & Innovation, Erasmus School of Law en senior fellow Jean Monnet Centre of Excellence in Digital Governance.

Literatuur

Collins, H. (2018). *Artifictional intelligence: against humanity's surrender to computers*. Cambridge: Polity Press.

COM (2020). *White Paper on Artificial Intelligence - A European approach to excellence and trust.* Brussels, pp. 27.

COM (2021). *Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts.* Brussels, pp. 206.

Gabriels, K. (2021). *Siri, wat adviseer jij? Over het gebruik van kunstmatige intelligentie voor morele oordeelsvorming*. Preadvies Nederlandse Vereniging voor Bio-ethiek.

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), pp. 389-399.

Mol, A. (2008). *The logic of care: Health and the problem of patient choice*. London/New York: Routledge.

Resseguier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, 7(2), DOI: 2053951720942541.

Wallenburg, I., Weggelaar, A.M. & Bal, R. (2019). Walking the tightrope: how rebels "do" quality of care in healthcare organizations. *Journal of Health Organization and Management*, 33(7/8), pp. 869-883.

Kunstmatige intelligentie en/of intelligente levenskunst

Jan Vorstenbosch

In deze bijdrage stel ik voor om vanuit patiëntenperspectief naar Artificial Intelligence (verder: AI) te kijken. Welke potentieel ethisch-relevante rol kan AI spelen in medische contexten waarin patiënten beslissingen moeten nemen in het licht van hun persoonlijke opvattingen? Het gaat dan om een vraag over hoe ze willen leven en wat voor rol gezondheid daarin speelt. Dit wordt soms ook wel als levenskunst aangeduid. Het principe van Informed Consent (verder IC) neem ik als leidraad om vragen te genereren rond AI in verschillende medische contexten.

In haar preadvies voor het jaarsymposium van de NVBe over kunstmatige intelligentie en morele oordeelsvorming baseert Katleen Gabriels (Gabriels, 2021) zich op het begrip 'Artificial Moral Agent' (AMA). Het preadvies over de rol van AMA's is sterk georiënteerd op het begrip 'morele oordeelsvorming' over medisch-ethische kwesties en minder op de interactie tussen arts en patiënt. Voor de patiënt speelt in die interactie ook de toepassing van een bredere, soms ethisch genoemde, opvatting over hun leven die geen 'morele' strekking in strikte zin heeft (Williams, 1985). Moraal gaat dan over rechten en plichten *tussen* mensen en is daarom gericht op maatschappelijke consensus, of zelfs universele waarden zoals mensenrechten. Ethiek draait om de vraag naar het goede leven die in een liberale samenleving een zaak van individuen is.

Ik wil vooral vanuit het patiëntenperspectief naar AI-toepassingen binnen min of meer alledaagse medische contexten kijken

Dit artikel gaat over de betekenis van AI voor de patiënt en deze vraagstelling. Dit uitgangspunt sluit dat van Gabriëls natuurlijk niet uit, nog los van de vraag of het onderscheid tussen morele en ethische vragen zo strikt is. Het gaat me in dit artikel dan ook meer om een accentverschuiving of aanvulling dan een kritiek op het preadvies. Ik wil vooral vanuit het patiëntenperspectief naar

AI-toepassingen binnen min of meer alledaagse medische contexten kijken. De toegevoegde waarde is dat dit perspectief licht kan werpen op de rol die AI in die contexten kan spelen bij het informeren van en het helpen beslissen door patiënten over opties en behandelingen. De snelheid waarmee AI op medisch terrein zijn intrede zal doen, maakt deze toespitsing op het patiëntenperspectief ook actueel. Beslissingsondersteunende systemen en beslissingssystemen zijn al volop in gebruik. We zijn bekend met zoeksystemen als Google of het privacybeschermende DuckDuckGo en navigatiesystemen zoals TomTom. Er zijn inmiddels allerlei kennissystemen die met algoritmes werken en nu al een grote rol spelen in tal van contexten, zoals bij de overheid en in de juridische context. Hun bredere inzet in medische contexten lijkt een kwestie van tijd.

Informed Consent en autonomie

Voor de beslissing wat de arts mag of moet doen, gegeven de persoonlijke opvatting en het oordeel van de patiënt, zijn er twee leidende ideeën: het Informed Consent (verder IC) principe en het idee van een 'gedeeld beslissingsproces' (shared decision-making) voor een behandeling. Ik richt me hier vooral op het begrip en het proces van IC. Ook toestemming is in zekere zin een gedeeld beslissingsproces maar de rechten en plichten zijn bij IC duidelijker verdeeld dan bij 'shared decision-making'. De arts heeft bij IC de plicht om de patiënt adequaat te informeren over de diagnose en het behandelingsplan *zodat* die patiënt het recht om een eigen beslissing te nemen (of hij/zij de behandeling wil of niet) op goede gronden kan uitoefenen.

Toestemming (consent) is iets heel anders dan een keuze maken op eigen initiatief, bijvoorbeeld de keuze om bij hoofdpijn paracetamol te slikken. In beide gevallen gaat het om een beslissing, maar bij toestemming is er sprake van een behandelingsoptie die alleen mag worden uitgevoerd door een arts en waar de patiënt ja of nee tegen moet zeggen. Wil die toestemming zinvol en geldig zijn, dan moet ze ook geïnformeerd zijn. Het is een notoir probleem – ook moreel maar vooral juridisch – welke informatie daarvoor moet worden gegeven (bijvoorbeeld over risico's) en hoe deze informatie wordt overgebracht. Op dit punt wordt het interessant om naar AI te kijken. Want met behulp van AI en bronnen zoals big data, evidence-based medicine, elektronische patiëntendossiers en 'personalized medicine' lijkt hier een wereld te winnen voor de IC-procedure. De pretenties die met informatiebronnen zoals big data en personalized medicine

zijn verbonden, raken immers het hart van het idee dat de persoon *zelf* kan en moet beslissen. Heeft de patiënt vaak niet meer aan AI dan aan de arts, niet alleen voor de beste behandeling maar ook om beter te weten wat hij of zij zelf wil? Door een combinatie van 'diepe', systematische informatie over de persoonlijke leefpatronen en voorkeuren van het individu enerzijds en de geobjectiveerde en op de patiënt-toegespitste informatie over de meest ideale behandelingsoptie anderzijds, zou tijdens de IC-procedure in elk geval de informatie beter kunnen worden geordend en de toestemming beter worden onderbouwd.

Heeft de patiënt vaak niet meer aan AI dan aan de arts, niet alleen voor de beste behandeling maar ook om beter te weten wat hij of zij zelf wil?

Zo lijkt het althans. Google, Amazon en Facebook, en allerlei 'spelers' op de commerciële markt voor 'deelgebieden' van ons persoonlijke leven claimen dat zij dankzij hun data ons beter kennen dan wij onszelf, in ieder geval in bepaalde opzichten. Maar uiteindelijk beslissen we toch zelf over onze aankopen, denken we dan. Maar is dat zo? Wat de medische sfeer speciaal maakt, is dat de inzet vaak hoger is, onze kennis als patiënt (en arts?) beperkter en onze eigen inbreng als patiënt in een beslissing lastiger. Het voert te ver om deze principiële zaken over autonomie en 'zelf' hier uit te werken. Misschien is dat ook niet nodig. Misschien varieert die autonomie, en misschien zelfs ons 'zelf', per medische context en al naar gelang de medische vraag. De contexten waarin medische beslissingen moeten worden genomen verschillen immers nogal. De verschillen betreffen o.a. de urgentie van de medische situatie, de complexiteit van de informatie en de rol van vertrouwen in de arts-patiëntrelatie. Zelfs zaken als de tijd en de ruimte die de arts en de patiënt wederzijds hebben, nemen of geven, kunnen een grote rol spelen. Ik schets globaal vier contexten en benoem telkens een paar aspecten.

Vier contexten

IC speelt vooral een centrale rol in de context van specialistische curatieve zorg in ziekenhuizen. Die specialistische zorg is vaak sterk handelingsgericht. Voor een goed onderbouwde optie kan de arts een beroep doen op databases en interpretatiemodellen met als leidraad de juiste Diagnose-Behandel-Combinatie. De kenniskloof tussen arts en patiënt is groot en voor veel specialismen is de staalkaart van (over-)wegingen zoals de risico's en de mate van invalidatie bij ingrepen (of de risico's op invalidatie) omtrent wat te doen, relatief helder te communiceren.

De patiënt zal in zijn of haar beslissing vaak volgend zijn, zoals ook de arts zich zal laten leiden door evidence-based praktijken, zo niet door protocollen. AI lijkt in deze context meestal een logische stap. Maar een belangrijke en vaak gestelde vraag is of de uitlegbaarheid van de algoritmen waarop het AI-advies draait, voor de arts geen probleem zal worden in de communicatie met de patiënt. Als de arts zijn of haar vertrouwen in het systeem niet kan funderen in een eigen oordeel, wat doet dat dan met het vertrouwen van de patiënt in de arts (en het systeem)?

Heel anders is de context van de huisartspraktijk. Daar is het klachtenpatroon van de patiënt het uitgangspunt en dat is vaak diffuus. Diagnostiek staat centraal en de vraag hoe en of er iets (bijvoorbeeld een medicament) zal worden ingezet is geen voorwerp van een *expliciete* IC-procedure. Er is hier vaker sprake van een proces dat afhankelijk is van houding en karakter van arts en patiënt en van hun onderlinge vertrouwensrelatie. In deze context kunnen beslissingssystemen misschien een objectiverende rol spelen. Maar ook hier geldt: wat als de huisarts zelf nauwelijks inzicht heeft in de achterliggende algoritmen? Is een goed gesprek dan niet beter?

Een derde context, waarin huisartsenpraktijken overigens ook een centrale rol spelen, is het aanbieden van vaccins en preventieve testen op bijvoorbeeld kanker. Hier ligt de bal volop bij de 'potentiële patiënt' die dit aanbod op de deurmat vindt en moet beoordelen en beslissen wel of niet mee te doen. Hier zou een beslissingsondersteunend systeem op internet waarin een aantal factoren worden gepresenteerd misschien goede diensten kunnen bewijzen, in combinatie met zelf in te voeren persoonlijke gegevens. Met enige aanpassingen zou ook het aanbod van een coronavaccin onder deze context kunnen worden geschaard, maar daarmee zitten we tegelijk op het macro-niveau van de overheid en in een veel complexer en dynamisch maatschappelijk proces.

Wat als de huisarts zelf nauwelijks inzicht heeft in de achterliggende algoritmen?

Een vierde context is de existentiële context van beslissingen over leven en dood. Euthanasie komt ook bij Gabriels aan de orde. Daar is duidelijk sprake van een *moreel* geladen situatie gezien de cruciale rol van de arts. Maar hier past Informed Consent weer niet omdat zo'n 'behandelings'-aanbod op initiatief van de arts juist *niet* verwacht wordt en zelfs dubieus is. Het verzoek moet van de patiënt komen. Het is in zekere zin juist omgekeerd: de arts moet geïnformeerd worden over de patiënt en daardoor overtuigd raken dat hij/zij mag handelen.

Dat wordt ook duidelijk uit het voorbeeld van AI van Gabriels waarbij informatie wordt vergaard over de opvattingen van de persoon over lijden via 'life-logging'. De vraag is of dit alleen data-vergaring is of ook AI-bemiddelde data-ordening en interpretatie. In de huidige wetgeving gaat het immers om het criterium of de patiënt *hier en nu* 'ondraaglijk lijdt' naar de overtuiging van de arts. Dat is een lastige (inter-)subjectieve vraag waarover al veel is geschreven door ethici. Het gaat niet om informatie over de vraag wat de persoon in het verleden van ondraaglijk lijden 'vond' of als ondraaglijk ervoer.

Besluit

In de eerste fase van de coronapandemie is er veel geschamperd over Rutte's 'intelligente' lockdown. Helaas is bij die gelegenheid de kans gemist om een brede maatschappelijke discussie te beginnen over wat het betekent om intelligent, niet alleen oplossingsgericht, maar vooral wijs en verstandig om te gaan met onze gezondheid en onze gezondheidszorg. Voorlopig komt het me voor dat we, in ieder geval als patiënten, daarbij meer zullen hebben aan intelligente levenskunst dan aan kunstmatige intelligentie. Al zou het mooi zijn als we de verwerking van de enorm toegenomen informatie en inzichten over gezondheid met behulp van AI konden integreren in die intelligente levenskunst.

Jan Vorstenbosch was tot zijn pensioen in 2018 als UD Toegepaste Ethiek verbonden aan het Departement Wijsbegeerte van de Universiteit Utrecht.

Literatuur

Gabriels, K. (2021). *Siri, wat adviseer jij? Over het gebruik van kunstmatige intelligentie voor morele oordeelsvorming.* Preadvies voor het jaarsymposium van de NVBe.

Giubilini A. & Savulescu, J. (2018). The artificial moral advisor. The 'ideal observer' meets artificial intelligence. *Philosophy and Technology*, 31, pp. 169-188.

Williams, B. (1985). *Ethics and the limits of philosophy.* Oxford: Oxford UP, Vooral: Ch. 10. Morality, The Peculiar Institution.

Vertrouwen in de geneeskunde en kunstmatige intelligentie

Lily Eva Frank en Michal Klincewicz

Kunstmatige intelligentie (AI) en systemen die met *machine learning* (ML) werken, kunnen veel onderdelen van het medische besluitvormingsproces ondersteunen of vervangen. Ook zouden ze artsen kunnen helpen bij het omgaan met klinische, morele dilemma's. AI/ML-beslissingen kunnen zo in de plaats komen van professionele beslissingen. We betogen dat dit belangrijke gevolgen heeft voor de relatie tussen een patiënt en de medische professie als instelling, en dat dit onvermijdelijk zal leiden tot uitholling van het institutionele vertrouwen in de geneeskunde.

Machine learning en kunstmatige intelligentie in de geneeskunde

AI/ML-gestuurde systemen kunnen veel impact hebben op het medische besluitvormingsproces zoals diagnoses, prognoses, of de keuze van een behandeling. Daarnaast kunnen deze systemen, zoals aangegeven door Gabriels (2021), artsen helpen met het omgaan met morele dilemma's in de klinische praktijk, bijvoorbeeld als de plicht om de persoonlijke informatie van de patiënt te respecteren in strijd is met de plicht om te handelen in het belang van de patiënt. In dergelijke situaties kunnen AI/ML worden opgevat als ethische ondersteuning in de vorm van een kunstmatige morele adviseur (Giubilini & Savulescu, 2018; Klincewicz, 2016).

Op al deze gebieden (en mogelijk andere) kunnen zij medische professionals vervangen die hun opleiding, ervaring en professionele intuïtie gebruiken om beslissingen te nemen maar ook de gezamenlijke beslissingen die genomen worden met de patiënten en/of andere leden van het medische personeel. Hieronder vallen ook juridische adviseurs en mensen die zitting hebben in speciale besluitvormingsorganen, zoals bijvoorbeeld ethische commissies. Met andere woorden, AI/ML-beslissingen kunnen mogelijk in de plaats komen van professionele beslissingen.

Dit heeft twee belangrijke consequenties voor de relatie tussen een patiënt en de medische wereld als instelling die samenhangen met de uitlegbaarheid en mogelijke fouten van AI/ML-gestuurde besluitvorming. Dit zal onvermijdelijk leiden tot een afbraak van het institutionele vertrouwen in de geneeskunde.

Vertrouwen van patiënten in individuele medische professionals

Patiënten hebben vaak niet de opleiding, ervaring, of intuïtie om medisch adviezen en medische beslissingen te begrijpen. Zelfs wanneer klinische informatie goed wordt gecommuniceerd en begrepen, worden deze beslissingen beïnvloed door de juridische, ethische en culturele context van de arts of van het ziekenhuis. Deze asymmetrie van kennis betekent voor de zorgverlening dat patiënten erop zullen moeten vertrouwen dat de medische professional handelt in hun belang en dat de professional de normen en waarden van de patiënt respecteert. Dit schept een speciale set van plichten voor medische professionals, waarvan de belangrijkste zijn dat ze als professionals betrouwbaar zijn en het ook inderdaad verdienen om dat vertrouwen te krijgen (Rhodes, 2020). Dat is het geval als patiënten hen dat vertrouwen omtrent medische beslissingen geven en ook laten zien. Natuurlijk zijn er ook omstandigheden waarin de arts te vertrouwen is maar de patiënt hem toch niet vertrouwt.

Asymmetrie van kennis betekent voor de zorgverlening dat patiënten erop zullen moeten vertrouwen dat de medische professional handelt in hun belang

Het vertrouwen van de patiënt in het tijdperk van kunstmatige intelligentie

Het vertrouwen van de individuele patiënt in een medische professional of beslissing wordt beïnvloed door het maatschappelijke vertrouwen in de medische wereld als geheel. Dit meer algemene vertrouwen is het resultaat van de mate waarin medische professionals, als groep, in het verleden aan hun plichten hebben voldaan (Rhodes, 2020). In dat geval zullen toekomstige patiënten eerder vertrouwen hebben, zo niet, dan zal het vertrouwen lager zijn (Jacobs et al., 2006).

Dit algemene niveau van institutioneel vertrouwen in de geneeskunde kan echter worden ondermijnd door het *onkritische* gebruik van AI/ML in de geneeskunde. Ten eerste zal in een klinische setting een medische professional de prestaties van een AI/ML-gestuurde medische aanbeveling of beslissing moe-

ten kunnen uitleggen en deze ook kunnen rechtvaardigen. Als bijvoorbeeld een diagnose van kanker gedeeltelijk door een algoritme is vastgesteld, dan zal de patiënt vermoedelijk willen weten waar die diagnose op gebaseerd is en kan hij of zij ook vragen hoe de AI/ML werkt en waarom de arts de diagnose vertrouwt. De arts heeft daarbij de plicht om eerlijk te zijn over de rol van de AI/ML en zijn of haar eigen beperkte begrip van de werking ervan.

De arts heeft de plicht om eerlijk te zijn over de rol van de AI/ML en zijn of haar eigen beperkte begrip van de werking

Een deel van dit probleem vloeit voort uit de ondoorzichtige of *black-box* aard van de algoritmes in *machine learning* systemen die adviseren over diagnose, prognose en behandelingen. Inzichten die een AI/ML verschaft zijn niet gemakkelijk in begrijpelijke taal uit te leggen aan de doorsnee patiënt. Dit in tegenstelling tot andere diagnose-instrumenten, zoals CT-scans en bloedtesten, die een interpretatie door een medische professional zelf vergen, wat ook onderdeel is van het klinische besluitvormingsproces. In het geval van een AI/ML-besluit wordt de patiënt gevraagd erop te vertrouwen dat het vertrouwen van de medische professional in het AI/ML besluit gerechtvaardigd is. Alhoewel vertrouwen niet zomaar inwisselbaar is, moeten we ons afvragen of deze overdracht van het vertrouwen van de patiënt in de medische professional naar het AI/ML-systeem niet ten koste kan gaan van het institutionele vertrouwen in de medische professie zelf. Een verkoper die ons merkproducten van hoge kwaliteit verkoopt, zal niet noodzakelijk ons vertrouwen in verkopers in het algemeen verhogen, maar wel in het merk en het product. Op vergelijkbare wijze zal het overdragen van de besluitvorming naar AI/ML het algemene vertrouwen in deze systemen en de ingenieurs die ze ontwerpen, vergroten, ten koste van het vertrouwen van de maatschappij in het medische vakgebied.

Vertrouwen en fouten door kunstmatige intelligentie

Een ander gevolg van de introductie van kunstmatige intelligentie met betrekking tot het vertrouwen in de geneeskunde, is dat AI/ML-gestuurde besluitvorming, ondanks de belofte dat zij wellicht nauwkeuriger is dan uitsluitend menselijke besluitvorming, kan leiden tot fouten met schadelijke gevolgen voor de patiënt, zoals een misdiagnose, een fout recept, of gewoon een gebrek aan gevoeligheid voor de waarden van de patiënt. Fouten kunnen tot juridische consequen-

ties leiden zoals rechtszaken en financiële sancties. Maar hoe zal de juridische (en morele) verantwoordelijkheid worden vastgesteld in gevallen waar AI/ML-gestuurde besluitvorming bij betrokken was? Zal de medische professional verantwoordelijk worden gehouden? Of het bedrijf dat de AI/ML-systemen leverde? Dit is een specifieke versie van een algemeen probleem in ethiek van robotica en kunstmatige intelligentie: de verantwoordelijkheidskloof (Matthias, 2004; de Sio & Mecacci, 2021). Dit is het fenomeen waarbij een onkritische invoering van automatisering in complexe besluitvormingsprocessen het onmogelijk maakt om verantwoordelijkheid te leggen bij een individu. Een belangrijk gevolg hiervan is de afbraak van vertrouwen omdat van AI/ML-ingenieurs, programmeurs en bedrijven over het algemeen niet wordt verwacht dat zij hetzelfde niveau of dezelfde soort morele verplichtingen hebben als medische beroepsbeoefenaren.

Behoud van het vertrouwen in de geneeskunde als beroep

Algemeen wordt ervan uitgegaan dat het belang van de patiënt voorop staat bij de medische professional en dat dit de medische institutie kenmerkt. Maar als AI/ML erbij betrokken wordt, dan kan de medische professional geen verantwoordelijkheid nemen over fouten en beslissingen, omdat de medische professional geen ingenieur is en uiteindelijk niet verantwoordelijk is voor het juist functioneren van het AI/ML-systeem. Door deze complexe situatie is het lastig om met voldoende zekerheid te generaliseren over toekomstige juridische procedures of ethische oordelen. Er moet daarom worden nagedacht over de mate waarin de arts misschien gedeeltelijk verantwoordelijk kan worden gesteld hiervoor.

Er moet worden nagedacht over de mate waarin de arts misschien gedeeltelijk verantwoordelijk kan worden gesteld

Er is echter een groot verschil tussen de toewijzing van verantwoordelijkheid voor misdiagnoses bij gebruik van AI/ML en situaties waar alleen professionals bij betrokken zijn. In dat laatste geval wordt gebruik gemaakt van het institutionele vertrouwen van mensen in de beroepsgroep om medische professionals niet aansprakelijk te stellen (moreel of juridisch) voor ingrepen waar dat in andere contexten wel het geval zou zijn geweest, zoals snijden, “vergiftigen” (medicijnen toedienen), of verkeerd informeren. In het geval van AI/ML beslissingen ligt de aansprakelijkheid in eerste instantie bij de professional vanwege zijn vertrouwen in het AI/ML-systeem, maar dat roept het beeld op dat de professio-

nal zelf niet betrouwbaar is. Als er meer van dit soort situaties zich voordoen, kan het vertrouwen in de medische beroepsbeoefenaars als groep worden aangetast.

Om de onvermijdelijke erosie van institutioneel vertrouwen in de geneeskunde tegen te gaan moeten we sterk benadrukken dat de beoefening van geneeskunde meer is dan het sleutelen aan het lichaam om de optimale werking te behouden, alsof het een auto is. Het beoefenen van geneeskunde gaat gepaard met rechten en plichten voor professionals, die anderen niet hebben. Zoals Rosamond Rhodes opmerkt kunnen artsen in een lichaam van iemand snijden, medicatie verstrekken, en onze naakte lichamen onderzoeken, maar tegelijkertijd moeten zij zich ook aan de geheimhoudingsplicht houden, handelen in het belang van anderen, eerlijk zijn, en streven naar vertrouwen en dat vertrouwen ook verdienen (2020). Dit alles op een manier die veel verder gaat dan wat we verwachten van mensen in andere beroepen. De geneeskunde is een fundamenteel moreel toegewijd beroep. De introductie van door AI/ML ondersteunde besluitvorming in de klinische situatie moet zorgvuldig beoordeeld worden in het licht van deze bijzondere plichten en rechten.

We moeten een manier vinden waarop we het institutionele vertrouwen, dat in het hart van de medische praktijken ligt, kunnen beschermen, aangezien dat vertrouwen het mogelijk maakt voor patiënten om zichzelf kwetsbaar op te stellen tegenover interventies die zij misschien niet volledig kunnen waarderen of begrijpen. Als we dit niet doen, kunnen medische professionals en de geneeskunde als instelling hun speciale status, en het hoge vertrouwen dat ze krijgen, verliezen. Om AI/ML op een ethische manier in de medische praktijk te integreren, is op zijn minst een aanvullende medische (en voortgezette) opleiding nodig. Dit leerplan moet worden ontwikkeld in samenwerking met artsen, patiëntvertegenwoordigers, ethici en AI/ML-wetenschappers.

Dr. Lily Eva Frank is assistant professor Industrial Engineering and Innovation Science bij de Philosophy and Ethics Group van Eindhoven University of Technology.

Dr. Michal Klincewicz is assistant professor Cognitive Science and Artificial Intelligence bij Tilburg University.

Literatuur

De Sio, F.S. & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, https://doi.org/10.1007/s13347-021-00450-x

Giubilini, A. & Savulescu, J. (2018). The artificial moral advisor. The "ideal observer" meets

artificial intelligence. *Philosophy & Technology*, 31(2), pp. 169-188.

Jacobs, E.A., Rolle, I., Ferrans, C.E., Whitaker, E.E. & Warnecke, R.B. (2006). Understanding African Americans' views of the trustworthiness of physicians. *Journal of General Internal Medicine*, 21(6), pp. 642-647.

Klincewicz, M. (2016). Artificial intelligence as a means to moral enhancement. *Studies in Logic, Grammar and Rhetoric*, 48(1), pp. 171–187. https://doi.org/10.1515/slgr-2016-0061

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), pp. 175–183. https://doi.org/10.1007/s10676-004-3422-1.

Rhodes, R. (2020). *The trusted doctor: Medical ethics and professionalism*. Oxford: Oxford University Press.

‘We moeten niet denken dat AI als een orakel antwoord kan geven op morele vragen’

Een interview met Katleen Gabriels

Marieke Bak en Sjaak Swart

Katleen Gabriels schreef voor het online NVBe jaarsymposium ‘*Het gebruik van kunstmatige intelligentie voor morele oordeelsvorming*’ op 11 juni 2021 het preadvies ‘*Siri, wat adviseer jij? Over het gebruik van kunstmatige intelligentie voor morele oordeelsvorming*’ waarin ze ingaat op AI-systemen die moreel kunnen oordelen. Een verbazingwekkende technologie die onder ethici veel vragen oproept. Podium ging in gesprek met haar naar aanleiding van het preadvies.[1]

Waar komt jouw belangstelling voor kunstmatige intelligentie (AI) en ethiek vandaan?

Ik heb me altijd geïnteresseerd voor de impact van technologie op moraliteit. Tijdens mijn PhD-onderzoek bij de Vrije Universiteit Brussel over virtuele moraliteit maakte ik kennis met *Second Life*, een digitale driedimensionale wereld waarin mensen zich konden voordoen als degene die ze graag wilden zijn. Dat leverde vragen op over wie dat dan is en wat de onderlinge regels zijn. Later kwam het concept van *Internet of Things* op waarnaar ik onderzoek deed als postdoc. Nu ben ik als universitair docent bij de Universiteit van Maastricht bezig met AI en blijkt het concept van virtuele moraliteit weer ontzettend relevant te zijn bij de vraag in hoeverre moraliteit aan een machine geleerd kan worden.

Waarom is het belangrijk de opkomst van AMA’s (Artificial Moral Agents) te onderzoeken?

De inzet van AI voor medische doeleinden, bijvoorbeeld in de diagnostiek, wordt

steeds belangrijker. Vanwege de verwevenheid van technologie en moraliteit is het dan nog maar een kleine stap naar morele oordeelsvorming zelf, die bij AMA's expliciet in de technologie besloten ligt. Echter, als we AI inzetten voor bijvoorbeeld het opsporen van verdacht borstweefsel, dan is het doel duidelijk: we willen zo snel mogelijk een behandeling starten om iemands leven te redden. Bij moraliteit is het vaak niet zo duidelijk.

En waarom heb je euthanasie als casus gekozen bij AMA-1 (een AI-systeem dat alleen moreel relevante data aanlevert)? Daar zijn immers zonder AI al genoeg morele problemen?

Dat is waar. Maar euthanasie prikkelt meer als casus. Het laat namelijk zien dat de vraag of we dit wel moeten willen, ook bij de minst vergaande AMA's al relevant is. Ik had ook de casus kunnen kiezen van de verdeling van IC-bedden op basis van medische dataprofielen. Gelukkig is Code Zwart tijdens de coronacrisis niet opgetreden, maar ik vraag me af of het gebruik van een AMA in dat geval geaccepteerd zou zijn.

Wat zijn de belangrijkste uitdagingen voor de toepassing van AMA's in de medische sector?

Ten eerste de vaagheid en dubbelzinnigheid van taal waarmee AMA's getraind moeten worden. AI-systemen hebben grote moeite om de betekenis van woorden in context te begrijpen. Overigens niet alleen in het medische domein. Een ander obstakel is de mogelijk beperkte bereidheid om morele oordeelsvorming aan een machine over te laten. In hoeverre willen we vertrouwen op de technologie? Als een systeem alleen ondersteunend is, is er geen probleem, want er kan van afgeweken worden. Ik werk nu samen met een Amerikaanse hoogleraar aan een artikel over AMA-3 waarin een smartphone app je informeert over verschillende morele opties en bijvoorbeeld zegt: 'Bekijk het eens zo' of 'Dat is een drogreden'. Van zo'n systeem moet je geen moreel superieur antwoord verwachten maar het kan wel veel meer data verzamelen dan de mens en goede argumenten bij een ethische afweging vinden. Daarvoor heb je wel een systeem nodig dat goed in taal is en in staat is om bijvoorbeeld drogredenen en metaforen te herkennen. Voordat die systemen er echt zijn, ben ik met pensioen, denk ik.

Als een systeem alleen ondersteunend is, is er geen probleem, want er kan van afgeweken worden

Zouden AMA's de mens ook kunnen overrulen, of misleiden?

Dat gebeurt nu al. Zelfrijdende auto's kunnen de bestuurder overrulen, bijvoorbeeld. Maar dat kunnen we wellicht accepteren omdat ze efficiënter, sneller, en op meer data gebaseerd zijn. We schrikken van een ongeluk met zo'n auto, maar tegelijkertijd kun je je afvragen hoeveel levens er worden gered door die automatische piloot. We wéten dat machines in sommige opzichten beter zijn dan mensen. De vraag is of dat ook zo is bij moreel oordelen. Mensen kunnen beter omgaan met de context waarin een moreel dilemma zich voordoet, maar zijn sterk beïnvloedbaar. AMA's zullen niet voor een egoïstische verleiding vallen zoals de Vlaamse burgemeester die bij de COVID-vaccinatie ten onrechte voordrong. Daar was overrulen door een AMA nog niet zo gek geweest. Misleiden daarentegen vind ik een lastig begrip omdat daar een intentie achter zit. Natuurlijk kan er wel sprake zijn van misleiding door de programmeur.

Moeten we bang zijn voor het gebruik, of zelfs misbruik, van AMA's in de commerciële sector?

Bedrijven zouden AMA's kunnen gaan gebruiken zodra zij daarmee tijd en geld kunnen besparen, bijvoorbeeld voor een soort ethiek-audit. Het is de vraag of dat wenselijk is. Algoritmes spelen nu al een grote rol in het beleid van bedrijven en dat gaat niet altijd goed, ook niet bij overheden overigens. Het maakt de vraag naar verantwoordelijkheid veel lastiger. Denk bijvoorbeeld aan de toeslagenaffaire, een wel heel grimmig voorbeeld. Het is eigenlijk idioot dat we politici niet dwingen hun verantwoordelijkheid te nemen in dat soort situaties. Als je ziet hoe Rutte wegkomt met 'dat had ik niet in mijn kortetermijngeheugen opgeslagen', dan denk ik, geef die man een AI-implantaat!

Als je ziet hoe Rutte wegkomt met 'dat had ik niet in mijn kortetermijngeheugen opgeslagen', dan denk ik, geef die man een AI-implantaat!

In het preadvies schrijf je dat AI-ontwikkelaars persoonlijk verantwoordelijk moeten zijn voor de algoritmes die ze ontwikkelen. Maar die ontwerpers zijn toch onderdeel van een machtssysteem waarin de managers de opdrachten geven?

Dat is zo, maar dan heb je het meer over de negatieve opvatting van verantwoordelijkheid en over aansprakelijkheid. Diverse voorbeelden laten zien dat het misgaat omdat individuen geen verantwoordelijkheid kregen en er niet naar hen

werd geluisterd. Dat speelde bij de Challenger-ramp in 1986 (het ruimteveer dat ontplofte vlak nadat het opsteeg, red.). Ook bij de casus van Cambridge Analytica, het bedrijf dat onrechtmatig verkregen Facebookgegevens voor politieke doeleinden gebruikte, speelde een soortgelijk gebrek aan positieve verantwoordelijkheid. De bedrijfscultuur moet dus worden aangepakt opdat er meer horizontale inspraak komt in het ontwerpproces.

Stel dat ook de AMA's zich verder ontwikkelen dan niveau 4, hoe verhoudt zich dat tot de mens en wat betekent dat bijvoorbeeld voor verantwoordelijkheid en aansprakelijkheid?

Morele AI kan het natuurlijk ook mis hebben, net als mensen. Dat probleem kennen we nu ook al. Een Tesla die een witte vrachtwagen ziet als witte lucht, kan een crash veroorzaken. Wie is dan verantwoordelijk? Tesla of de bestuurder? Als de bestuurder tijdens de crash de krant leest terwijl hij of zij geacht wordt het stuur vast te houden, dan is het wel duidelijk. Of denk aan het voorbeeld met de IC-bedden: ik denk dat de AMA daar een hulpmiddel zou moeten zijn om de situatie door te rekenen en dat de beslissingen door mensen genomen moeten worden. Maar als AMA's ooit ingebouwd worden in ons lichaam, zoals met pacemakers of 'deep brain stimulation' nu al het geval is, wordt het wel complexer en ontstaan er bijzonder veel interessante vragen die in het preadvies nog niet werden besproken.

Terug naar de ethiek zelf. Je suggereert in het preadvies dat je een AMA ook kunt gebruiken om verschillende ethische theorieën te exploreren. Maar leidt dat niet tot cherry picking en relativisme?

Ik denk dat dat meevalt. We gebruiken immers nu ook al verschillende bronnen om tot een oordeel te komen. Vergelijk het met boeken: je kunt altijd meer boeken lezen om je kennis bij te spijkeren en dan nog steeds zelf de beslissing nemen. Zo kun je met een AMA verschillende ethische theorieën exploreren om tot een meer afgewogen oordeel te komen. Bij het schrijven vond ik AMA-4 het leukst om over na te denken, omdat daar het oordeel in het systeem zelf zit en er dus een keuze plaatsvindt tussen verschillende theorieën. Daarbij zijn er overigens ook variaties binnen die theorieën. Zou een AMA die getraind werd met het utilisme, kiezen voor een aanpak gebaseerd op Jeremy Bentham of op John Stuart Mill? Dat is nog best een verschil.

Lopen we niet het gevaar dat met name de westers georiënteerde ethiek dominant wordt bij de ontwikkeling van AMA's? En zijn bepaalde theorieën zoals consequentialisme niet makkelijker te implementeren in een AMA, dan bijvoorbeeld de zorgethiek die een affectieve relatie met het systeem veronderstelt?

Absoluut. Er zijn enorme verschillen in uitkomsten als je de AMA-4 robotopvoeder uitrust met een confuciaanse of een utilistische ethiek. Toen ik nog aan de TU Eindhoven werkte, zagen veel studenten vooral heil in consequentialisme, omdat die ethische theorie het best toelaat om op alles een getal te plakken. Maar in een brainstorm met mensen die bij DeepMind werkten (een door Google overgenomen AI-bedrijf, red.) was men huiverig om overal een getal aan te verbinden, want hoe kun je aan deugden als eerlijkheid of samenwerking een getal verbinden? We moeten onze beperkingen hierin echt toegeven en duidelijk maken. Zelfs consequentialisme is problematisch. Want uiteindelijk moet je daar ook grenzen stellen over wat je wel of niet wilt meenemen in de afweging. De techniek kan die grenzen niet stellen. In mijn colleges voor ingenieurs zijn er studenten die hopen dat alles met stappenplannen en pijlen weergegeven kan worden. Maar ethiek werkt niet zo.

In mijn colleges voor ingenieurs zijn er studenten die hopen dat alles met stappenplannen en pijlen weergegeven kan worden

Ben jij als ethicus zelf bang om je baan te verliezen door de opkomst van AMA's?

Wat een grappige vraag. Tien jaar geleden zei men dat je iets ingenieursachtig moest gaan studeren en zeker geen filosofie – want daar kon je geen brood mee verdienen. Nu zegt men dat we filosofen nodig hebben vanwege de maatschappelijke en ethische uitdagingen. We hebben in ieder geval breed opgeleide mensen nodig die kunnen argumenteren en analyseren. En of dat nou empirisch of theoretisch georienteerd onderzoek is, daar ben ik zelf vrij liberaal in. Ze hebben allemaal een plek en ik ben vooral voor samenwerking. En nee, ik ben niet bang voor banenverlies. Misschien worden we straks ondersteund door AI, bijvoorbeeld door een AMA-3.

Je legt de eigen grens van wenselijkheid AI-systemen tussen AMA-3 en -4?

Ja, ik hou meer van het socratische idee van een AMA-3 systeem dat mensen begeleidt in het stellen van vragen en het zoeken naar antwoorden. Een systeem dat 24/7 beschikbaar is en je nooit uitlacht als je een domme vraag stelt. Het is eigenlijk wat dat betreft een 'veilige' technologie. Maar ook een AMA-3 is niet

zonder risico's: zo'n systeem kan gehackt worden en tot morele luiheid leiden. Aan AMA-4 systemen zijn echter nog veel meer onduidelijkheden verbonden en wellicht ook negatieve consequenties voor de samenleving. Maar voorlopig zie ik een AMA-4 nog niet ontwikkeld worden.

Goed ethiekonderwijs en de inpassing van AI-systemen in de samenleving en instituties zijn wel belangrijk. Daar besteden we in het onderwijs aan ingenieurs nog weinig aandacht aan. Bij artsen is dat beter, daar zit de ethiek er altijd al meer in omdat het duidelijk is dat abortus, euthanasie en het leven geen zwart-wit zaken zijn. We moeten niet denken dat AI als een orakel antwoord kan geven op morele vragen.

Het was opvallend dat er bij het NVBe jaarsymposium geen mensen vanuit bedrijven als Google of Facebook zijn uitgenodigd

Wat is je advies aan ethici?

Mijn advies, sterk beïnvloed door negatieve ervaringen tijdens de Belgische opleiding, is dat je ethiek niet separaat moet zien van niet-ethische praktijken, domeinen en disciplines. Kom eens uit je silo! Praat met andere mensen met andere denkkaders en zoek de dialoog op. In de medische ethiek wordt voortdurend samengewerkt met artsen maar ik weet niet of er veel ethici zijn die op bezoek gaan bij bedrijven. Het was opvallend dat er bij het NVBe jaarsymposium geen mensen vanuit bedrijven als Google of Facebook zijn uitgenodigd. Die kruisbestuivingen kunnen juist veel opleveren.

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 tevens bestuurslid van de International Society for Ethics and Information Technology (INSEIT), voorzitter van ETHICOMP en geaffilieerd lid van 4TU Centre for Ethics and Technology.

Marieke Bak, MSc MA, is postdoc onderzoeker Medische Ethiek bij het Amsterdam UMC (locatie AMC). Haar onderzoek richt zich op ethische vragen rondom gebruik van big data en AI in medisch onderzoek. Zij is tevens redactielid van het Podium voor Bio-ethiek en van het bestuur van de NVBe.

Dr. J.A.A. (Sjaak) Swart was tot zijn pensionering universitair hoofddocent bij de onderzoeksgroep Integrated Research on Energy, Environment and Society (ESRIG) van de Rijksuniversiteit Groningen en is daar nog steeds aan verbonden als onderzoeker. Hij is tevens redactielid van het Podium voor Bio-ethiek

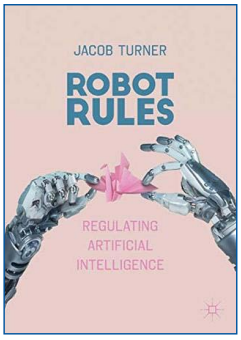
Noten

1. Het interview vond online plaats op 30 augustus 2021.

Rechten en plichten voor AI?

Sjaak Swart

Boekbespreking:



Robot Rules. Regulating Artificial Intelligence
Jacob Turner
Springer Nature Cham, Zwitzerland, 2019
ISBN 978-3-319-96234-4

De bijdragen in dit nummer gaan vooral over AI-systemen in de medische ethiek. Het is echter duidelijk dat AI-systemen steeds meer deel gaan uitmaken van ons leven. Zij worden daarbij steeds geavanceerder en slimmer en overtreffen reeds op sommige gebieden menselijke intelligentie. Zouden we aan AI-systemen misschien rechten en plichten moeten toekennen omdat ze over een (moreel) oordeelsvermogen beschikken en autonoom kunnen handelen?

Sterke AI en agency

'Robot Rules' (Turner, 2019) gaat vooral over sterke AI, al is het onderscheid met wat zwakke AI wordt genoemd vaag. Sterke AI-systemen hebben, aldus Turner, de 'ability to achieve an unlimited range of goals, and even to set new goals independently, including in situations of uncertainty or vagueness' (p. 6). Het zijn karakteristieken die we juist aan menselijke intelligentie toekennen.

Toekomstige AI-systemen zullen ook zelflerend zijn waarbij door mensen gemedieerde en gestuurde data-input nagenoeg of geheel ontbreekt. Dat kan tot onvoorziene en onverwachte resultaten leiden. Een voorbeeld is het Alpha Go Zero systeem dat het eerder ontwikkelde Alpha Go systeem, dat op big data was gebaseerd, binnen enkele dagen wist te verslaan in het go-spel zonder enige

data-input. Zulke systemen zijn ook in staat zijn de eigen algoritmes aan te passen om een gesteld doel te bereiken. Daarmee creëren ze dus hun eigen AI.

Dat is overigens niet uniek voor een technologie. Genetisch gemodificeerde organismen kunnen hun genetische code ook aanpassen tijdens hun reproductie. Maar zij hebben geen kennis van en houden geen rekening met de menselijke regelsystemen waaronder ze vallen. Turner concludeert dat "the combination of the ability to take decisions and to take those decisions based on their predicted effect within a system of rules and norms is what renders an entity as an agent" (p. 79). Door agency onderscheidt sterke AI zich volgens hem van andere opkomende technologieën. Een AI-systeem kan niet alleen zelfstandig oordelen en handelen maar dat ook interpreteren in bestaande maatschappelijke kaders en systemen. Dat roept vragen op over hun morele status en aansprakelijkheid en verantwoordelijkheid in zowel juridische als morele zin.

Uitdagingen

De auteur exploreert verschillende verantwoordelijkheid- en aansprakelijkheidssystemen en concludeert dat hun toepasbaarheid beperkt is vanwege onduidelijkheid of een AI-systeem als een object, subject of persoon moet worden aangemerkt. De overweging dat een AI-systeem agency heeft en trekken heeft van een subject, roept volgens hem de vraag op of aan AI-systemen rechten moeten worden toegekend. Velen vinden dat ridicuul maar de auteur beargumenteert dat we ook rechten toekennen aan dieren en soms ook aan ecosystemen en culturele goederen. Afhankelijk van de mate waarin een AI-systeem zou kunnen lijden (samenhangend met een mate van AI-bewustzijn), de mogelijke inherente waarde van het systeem en de mate waarin dat voor de mens van belang is, zouden we aan AI-systemen ook rechten en plichten kunnen toekennen. Daaraan voegt hij toe dat toekomstige AI-systemen wellicht worden gecombineerd met mensen tot hybride systemen of cyborgs. Turner concludeert dat we ooit voor de keuze komen te staan welke verantwoordelijkheid en rechten we AI-systemen gaan toekennen. Daarbij maakt hij geen scherp onderscheid tussen morele en juridische rechten omdat juridische rechten volgens hem vaak voortkomen uit morele rechten.

Turner concludeert dat we ooit voor de keuze komen te staan welke verantwoordelijkheid en rechten we AI-systemen gaan toekennen

AI-persoonlijkheid

De erkenning van AI-rechten impliceert dat we AI-systemen ook als rechtspersoon moeten erkennen, aldus Turner. Ook nu al kennen we niet-menselijke rechtspersonen zoals organisaties en bedrijven. Alhoewel mensen daarbinnen uiteindelijk de beslissingen nemen, hebben deze instituties als zodanig ook rechten en plichten die niet identiek zijn aan die van hun bestuurders. Zo kan een CEO of politicus die verantwoordelijk is voor een beslissing die tot grote maatschappelijke schade heeft geleid, de laan uit worden gestuurd, maar wordt hij of zij doorgaans niet opgezadeld met die schade zelf. Die blijft bij de organisatie.

Ingewikkelder wordt het als organisaties in de toekomst AI-systemen als medebestuurder opnemen of zelfs geheel zullen geleid worden door of gaan bestaan uit AI-systemen. In hoeverre kunnen deze systemen verantwoordelijk en aansprakelijk worden gesteld voor schade die ze mogelijk veroorzaken? Volgens Turner is het nodig rechtspersoonlijkheid met bijbehorende rechten en plichten aan deze AI-systemen toe te kennen, zoals bijvoorbeeld het recht op eigendom, toegang tot en onderwerping aan rechtssystemen en contractrecht, en registratieplicht. De kritiek dat we hiermee AI al te menselijk maken is volgens hem een “Android Fallacy” (p. 189), namelijk het gelijkstellen van persoonlijkheid aan menselijkheid. Rechtspersoonlijkheid is volgens hem een juridische constructie.

Regulering

Turner benadrukt de noodzaak van regulering van de opkomende AI-systemen. Tot nu toe vindt vooral zelfregulering plaats door de bedrijven als IMB en Microsoft die deze systemen ontwikkelen. Nuttig, maar volstrekt onvoldoende vanwege de grote commerciële belangen die aan AI-systemen verbonden zijn, aldus de auteur. De rol van overheden of andere publieke instituties is daarom onmisbaar. Daarbij is een unificerend ethisch raamwerk nodig in plaats van een veelheid van nationale en/of domeinspecifieke reguleringen omdat met name sterke AI zich ontwikkelt tot breed toepasbare, dat wil zeggen in meerdere domeinen bruikbare systemen. De auteur schetst een aantal voorbeelden van overheids- en professionele initiatieven waar ethische en juridische standaarden worden voorgesteld voor AI-systemen. Voorbeelden zijn de General Data Protection Regulation (GDPR) van de EU en de set van principes en waarden geformuleerd tijdens de Asilomar conferentie over AI in 2017.

Het gaat daarbij niet alleen om de rol van instituties en technologen, ook

aan AI-systemen zelf moeten eisen worden gesteld. Zo zouden ze als zodanig identificeerbaar en kenbaar moeten zijn in hun interactie met mensen (iets dat op dit moment vaak ontbreekt in bijvoorbeeld sociale media systemen). Een ander vereiste is dat een AI-systeem de redeneringen die ten grondslag liggen aan een oordeel kan uitleggen, wat met name voor machine learning toepassingen niet gemakkelijk is omdat artificieel denken niet gelijk is aan menselijk denken. Ook dient een systeem gevrijwaard te zijn van vooroordelen als gevolg van selectieve datasets en trainingsprocedures.

Zouden we de autonomie van een systeem moeten beperken door de eis dat bij een beslissing altijd een mens betrokken is?

Zouden we de autonomie van een systeem moeten beperken door de eis dat bij een beslissing altijd een mens betrokken is of geraadpleegd wordt? Dat zou volgens Turner ten koste gaan van doelmatigheid en de voordelen, zeker als de prestaties van een autonoom AI-systeem beter zijn dan die van een mens. Maar dat betekent niet dat de mens niet zou kunnen ingrijpen. Dat kan bijvoorbeeld door een zogenaamde *killswitch* die het systeem uitschakelt of neutraliseert bij foute of onwenselijke oordelen. Of dat voldoende is, is de vraag. Denkbaar is het dat de echt slimme AI-systemen van de toekomst in staat zijn hierop te anticiperen!

Pan met kokend water

Turner verwijst naar mensen als Elon Musk en wijlen Stephan Hawking die waarschuwen voor de existentiële dreiging en onbeheersbaarheid van AI-systemen als die in de toekomst de menselijke intelligentie zullen overtreffen (en dat soms al doen) en onvoorspelbaar en onbeheersbaar blijken te zijn. Dat is wellicht iets voor de toekomst, maar Turner stelt vast dat we ons nu reeds zorgen zouden moeten maken over de huidige (nog niet zo intelligente) AI-systemen die de wereld al lijken over te nemen. In dit verband waarschuwt hij voor twee onjuiste aannames, namelijk dat we de mate waarin AI-technologie binnendringt in onze huidige wereld voldoende kennen en dat het menselijk vernuft in staat zou zijn problemen tijdig op te lossen. De ontwikkelingen die Turner beschrijft, lijken op science fiction en ver weg. Maar de huidige AI-systemen ontwikkelen zich razendsnel. Door het cumulatieve karakter van verbeteringen

We zijn als de spreekwoordelijke kikker in een pan met water die we langzaam aan de kook brengen

merken we dat nauwelijks op. We zijn volgens Turner als de spreekwoordelijke kikker in een pan met water die we langzaam aan de kook brengen terwijl het dier dat niet opmerkt.

Relationele netwerken?

Het boek van Turner geeft een interessant overzicht van recente ontwikkelingen in de wereld van AI en de noodzakelijke regulering daarvan. Zijn pleidooi voor het toekennen van rechten en persoonlijkheid blijft echter hangen in een wat formeel, juridisch en pragmatisch kader. De vraag of er ook inhoudelijke, morele argumenten zijn om AI-systemen rechten en persoonlijkheid toe te kennen, bijvoorbeeld omdat ze een moreel subject zouden kunnen zijn, wordt helaas niet uitgewerkt. Daarnaast gaat het boek nauwelijks in op de rol van moraliteit in het dagelijks leven waarbij mensen elkaar immers morele rechten, plichten en persoonlijkheid toekennen mede op grond van intermenselijke relaties. In dat verband is het zinnig om te vragen in hoeverre toekomstige AI-systemen door mensen zullen worden ingepast in hun relationele netwerken waardoor er wederkerige relaties zouden kunnen ontstaan tussen AI en mensen, met daarbij ook wederkerige verwachtingen omtrent elkanders rechten en plichten. In hoeverre zullen bijvoorbeeld de AMA's type 3 en 4 van Katleen Gabriels (zie preadvies 2021) als een sociale partner worden gezien in plaats van alleen een tool om moeilijke morele beslissingen te kunnen nemen?

Dr. J.A.A. Swart was tot zijn pensionering universitair hoofddocent bij de onderzoeksgroep Integrated Research on Energy, Environment and Society (ESRIG) van de Rijksuniversiteit Groningen en is daar nog steeds aan verbonden als onderzoeker. Hij is tevens redactielid van het Podium voor Bio-ethiek.

Nieuws uit de Vereniging

André Krom

Overlijden Melanie Peters (Rathenau Instituut)

Het bestuur heeft geschokt gereageerd op het overlijden van Melanie Peters van het Rathenau Instituut, eerder deze zomer. Reeds sinds lange tijd zijn er goede en warme banden tussen het Rathenau Instituut en de NVBe. We wensen familie, vrienden en bekenden van Melanie ook graag via deze weg alle sterkte, evenals alle medewerkers van het Rathenau Instituut.

Gezamenlijk overleg redactieleden Podium voor Bio-ethiek en NVBe-bestuur

Op 20 oktober komen de redactieleden van Podium voor Bio-ethiek en het NVBe-bestuur voor het eerst sinds het begin van de coronapandemie fysiek bij elkaar voor een gezamenlijk overleg. Voornemen is om dit jaarlijks te doen. De redactie is inhoudelijk onafhankelijk van het bestuur, en dat moet vooral zo blijven. Ook dan is er alle gelegenheid om met elkaar van gedachten te wisselen over gemeenschappelijke zaken, zoals het verkennen van mogelijkheden om met onze publicaties en bijeenkomsten nog meer en beter, geïnteresseerden en belanghebbenden te bereiken.

Onderwijsmiddag 2021 op 5 november a.s.

Op vrijdag 5 november wordt van 14:00 – 17:30 uur de NVBe Onderwijsmiddag 2021 georganiseerd. Ook deze bijeenkomst zal sinds lange tijd weer fysiek plaatsvinden, te Utrecht. Het thema dit jaar is het gebruik van kunst in ethiekonderwijs. Het belooft een zeer interessant en interactief programma te worden. Sprekers met een eigen visie op het gebruik van kunst in ethiekonderwijs zullen deze middag aan het woord komen. Het zal gaan over literatuur (auto-biografieën, graphic novels, romans, poëzie), en over films en beeldende kunst, met Megan Milota (UMCU), Rob Houtepen (UM), Josette Jacobs (WUR), Anne-Fleur van der Meer (Radboudumc) en Maaike Haan (Radboudumc). Zij vertellen over de waarde van kunst en laten zien hoe zij hun onderwijs vormgeven. Heeft u zich al aangemeld?

2022 en verder

Het Preadvies 2021 is nog maar net verschenen, of het is alweer tijd om na te gaan denken over mogelijke onderwerpen voor het Preadvies voor volgend jaar, en daarmee tevens over het thema voor het jaarsymposium 2022. Dat geldt ook voor het vormen van de eerste ideeën over het volgende lustrum van de NVBe. In 2023 bestaat de NVBe alweer 30 jaar. Alle reden voor een feestje. Op korte termijn zal het bestuur hier een eerste brainstorm over organiseren. U hoort daar vervolgens via de gebruikelijk kanalen zo spoedig mogelijk meer over.

Bericht van het Rathenau Instituut: In Memoriam Melanie Peters

Bestuur en medewerkers van het Rathenau Instituut

Melanie Peters, onze energieke en bevlogen directeur, is op 11 augustus overleden. Het bericht van haar overlijden komt als een schok voor ons allemaal. Melanie was sinds 2015 directeur van het Rathenau Instituut. Onder haar vleugels maakte het instituut een geweldige sprong voorwaarts. Melanie kon als geen ander de uitwerking van wetenschap, technologie en innovatie op de samenleving tastbaar en invoelbaar maken. We zullen haar gedreven en warme persoonlijkheid missen.

Melanie stond voor het Rathenau Instituut als geheel. 'We doen het samen', hoorden we Melanie vaak zeggen. 'Goed teamwork!', schreef ze graag. En ze vond het belangrijk dat collega's van verschillende teams en afdelingen van elkaar wisten waar zij aan werkten. Dat maakte onze organisatie sterker.

Samenwerking werd ook een rode draad in de onderzoeken en de werkwijze van het Rathenau Instituut. Onder haar bezielende leiding knoopte het Rathenau Instituut samenwerkingen aan met vele organisaties in binnen- en buitenland. We onderzochten samenwerkingsverbanden in wetenschap, technologie en innovatie en benadrukten het belang van het gesprek tussen partijen met conflicterende belangen. Of het nu ging om de ammoniakkwestie of het aanpassen van DNA in menselijke embryo's. Voor Melanie betekende samenwerken: oog hebben voor ieders taak en verantwoordelijkheid. Om oplossingen te vinden voor grote maatschappelijke vraagstukken, hebben we de diversiteit aan disciplines en perspectieven nodig.

In een goede samenwerking mag het schuren, of zelfs knetteren. Dat deed het soms ook, want Melanie was vol vuur als het ging over de kwaliteit van de discussie over wetenschap, technologie en innovatie in de samenleving. Voor Melanie was het maatschappelijk perspectief van waaruit het Rathenau Instituut werkt, allerminst een vaag begrip. Het staat voor het beschermen van

publieke waarden in onze samenleving, zoals inclusiviteit, rechtvaardigheid en burgerrechten. 'We moéten met elkaar in gesprek' klinkt misschien vriendelijk, maar Melanie bedoelde dit niet vrijblijvend. Het was een ferme aansporing om te zeggen wat gezegd moet worden. Zowel tegen collega's onderling, als in het publieke en politieke debat, waaraan zij als directeur intensief meedeed.

Melanie had een uitzonderlijk groot netwerk – een teken van haar brede interesse en betrokkenheid – waarop het Rathenau Instituut kon bouwen. Ze had een geweldig geheugen, een rijke associatieve geest, was razendsnel in haar denken en doen, en legde verbanden die anderen nog niet zagen. Niet alleen figuurlijk, maar ook letterlijk was ze soms onnavolgbaar. Achter haar bureau was ze eigenlijk nooit te vinden; als een spring-in-het-veld rende ze van afspraak naar afspraak. Wonderlijk genoeg wist zij altijd precies waar elk overleg om draaide, merkten we.

Bovenal zullen we ons de warme en betrokken persoon die Melanie was herinneren. Voor Melanie stond de mens centraal. Wie in het leven klem kwam te zitten, ontmoette haar zachte en meelevende kant. En als het over ons onderzoek ging, stelde zij vragen als: 'Kan iedereen meedoen in de digitale samenleving?' 'Plukt iedereen de vruchten van de kenniseconomie?' Met Melanie als directeur waren dit leidende vragen voor het Rathenau Instituut. Toen het coronavirus ons dwong om vanuit huis te werken, onderstreepte ze het belang van menselijk contact. 'Houd afstand, maar wees nabij!', drukte ze ons op het hart.

In onze gedachten zijn we bij haar man Albert, haar dochters Eva en Swati en de naaste familie en vrienden, naar wie ons innig medeleven uitgaat. Wij wensen hen sterkte in deze moeilijke tijd.

Melanie, we gaan je enorm missen.

Nieuws van het Centrum voor Ethiek en Gezondheid

Myrthe Lenselink en Sandra in 't Groen

Het Centrum voor Ethiek en Gezondheid (CEG) signaleert over actuele en beleidsrelevante ethische vraagstukken over gezondheidszorg en biomedisch onderzoek. Het CEG brengt signalementen uit en organiseert bijeenkomsten, waarbij u uiteraard van harte welkom bent. Op de website van het CEG (www.ceg.nl) vindt u alle publicaties en actualiteiten. In deze bijdrage geven wij een toelichting op lopende projecten en ander nieuws.

Nieuw presidiumlid

Per 1 september 2021 is Hafez Ismaili M'hamdi vicevoorzitter van de CEG-commissie en lid van het CEG-presidium. Ook Jet Bussemaker, Bart-Jan Kullberg en Maartje Schermer maken deel uit van het presidium. Hafez Ismaili M'hamdi vervangt Jeannette Pols.

Hafez Ismaili M'hamdi is ethicus en universitair docent aan de afdeling Medische Ethiek, Filosofie en Geschiedenis van de Geneeskunde van het Erasmus MC. Hij is opgeleid in de ethiek en wijsbegeerte aan de Universiteit Utrecht en Universiteit Leiden en als musicus aan het Koninklijke Conservatorium van Den Haag. Ismaili M'hamdi is ook gastdocent esthetiek en filosofie van de muziek aan de *Academy of Creative and Performing Arts* van de Universiteit Leiden.

Voor zijn promotie deed Ismaili M'hamdi onderzoek naar de verdeling van verantwoordelijkheden voor de gezondheid en het welzijn van kinderen. Momenteel doet hij onderzoek naar gezondheidsverschillen en ethische kwesties rondom embryo's. Hafez Ismaili M'hamdi is ook vicevoorzitter van de CEG-commissie.

Els Borst Lezing: 22 november 2021

De Els Borst Lezing gaat dit jaar over de verbondenheid tussen de gezondheid van mens, dier en milieu volgens de One Health benadering. Dr. Bernice Bovenkerk, universitair hoofddocent dier- en milieu-ethiek bij de Filosofie Groep van

Wageningen Universiteit, presenteert haar visie.

Wereldwijd leven mensen en dieren vaak dicht op elkaar. De One Health benadering vertrekt vanuit de wederkerige relaties tussen de gezondheid van mens, dier en milieu. Zo liet de coronapandemie ons zien dat dieren ziektes kunnen overbrengen op mensen, maar ook dat dieren ziek kunnen worden als gevolg van menselijk ingrijpen. Hoe te handelen bij conflicten tussen de gezondheid van mens, dier en milieu? Mogen we dieren preventief slachten om de volksgezondheid te beschermen? Hoeveel gezondheidsrisico zijn we bereid te lopen om voldoende en betaalbaar voedsel te kunnen produceren?

In haar lezing beargumenteert Bernice Bovenkerk dat door een veranderende leefomgeving een transformatie van het systeem waarin mens, dier en milieu samenleven, hoognodig is. Noteert u 22 november 2021 alvast in uw agenda? De bijeenkomst is van 15.00 – 17.00 uur, met aansluitend een borrel. Meer informatie volgt binnenkort via de CEG-website.

Over de Els Borst Lezing: Sinds 2013 organiseert het CEG iedere november een lezing waarin een gerenommeerde deskundige spreekt over een ethisch thema of vraagstuk binnen de (gezondheids)zorg. Deze lezing is vernoemd naar Els Borst-Eilers, minister van Volksgezondheid, Welzijn en Sport van 1994 tot 2002. Als minister stond zij aan de wieg van het CEG. Els Borst-Eilers heeft tijdens haar carrière veel aandacht besteed aan verschillende ethische thema's, zoals euthanasie, het stelsel van donorverklaringen, en wetenschappelijk onderzoek met embryo's, geslachtscellen en foetaal weefsel.

Berichten van Unesco

Marike Bontenbal

Unesco is de organisatie van de Verenigde Naties belast met bio-ethiek. De 193 lidstaten voeren hierover een mondiale dialoog en maken internationale afspraken. In deze rubriek vertelt de Nederlandse Unesco Commissie meer over het (bio-) ethiekwerk van Unesco. Deze keer over een verklaring die de ethiekcommissies van Unesco op 1 juli uitbrachten over het belang van ethische afwegingen bij het gebruik van zogenaamde coronapaspoorten.

Coronapaspoorten zouden nooit ingezet moeten kunnen worden om bepaalde privileges voor te behouden aan gevaccineerden. Daarnaast mogen coronapaspoorten de keuzevrijheid van mensen niet beperken, aldus de Unesco ethiekcommissies IBC en COMEST. Zij wijzen op het gevaar dat de coronacertificaten gebruikt kunnen worden voor oneigenlijke doeleinden, waarmee bewust of onbewust discriminatie of sociale scheidslijnen in de hand worden gewerkt. Daarnaast moet er met het gebruik van de paspoorten meer oog zijn voor internationale solidariteit met landen waar het vaccinatieprogramma nog in de kinderschoenen staat.

Ongelijkheid

COMEST en IBC zien het gebruik van het coronapaspoort als een belangrijk middel om de vrijheden van burgers te herstellen, zoals het zich vrij kunnen bewegen in de openbare ruimte, het bezoeken van locaties en evenementen en het deelnemen aan het internationale verkeer. Tegelijkertijd dreigt het gevaar dat de paspoorten tot nieuwe vormen van uitsluiting en discriminatie leiden. Bijvoorbeeld wanneer mensen vanwege gebrek aan financiële middelen, vervoer of kennis- en informatievoorziening minder goede toegang hebben tot de mogelijkheid zich te laten testen of te vaccineren. Ook heeft niet iedereen goede toegang tot het gebruik van een digitale app die de vaccinatie- en testbewijzen registreert, zoals ouderen en digitaal laaggeletterden.

Het is bekend dat de pandemie juist de mensen treft die behoren tot kwetsbare groepen, minderheidsgroepen (zoals vrouwen), of mensen die leven

in armoede. Het is ook bekend dat de pandemie sociale en economische ongelijkheid wereldwijd, maar ook op lokaal niveau, heeft verdiept. Juist daarom moet de inzet van coronacertificaten zorgvuldig worden georganiseerd.

Mensenrechten

De invoering van het coronapaspoort kan een aantal grondrechten ondermijnen, zoals bewegingsvrijheid. En wanneer mensen zich gedwongen voelen om zich te laten vaccineren omdat zij anders bepaalde privileges (toegang, de mogelijkheid om te reizen, etc.) mislopen, komt het recht op lichamelijke integriteit in het geding. Dit wordt benadrukt in een opiniestuk in NRC van hoogleraar Peter-Paul Verbeek, lid van de Nederlandse Unesco Commissie en internationaal voorzitter van COMEST: "De belangrijkste en gevoeligste kwestie rond de coronapaspoorten is de vraag of zo'n paspoort meer vrijheid biedt aan wie gevaccineerd is en zo vaccinatie tot een indirecte verplichting maakt. Niet iedereen kan of wil zich laten vaccineren. [...] Lichamelijke integriteit is een mensenrecht dat in het huidige beleid zorgvuldig wordt gerespecteerd; het coronapaspoort mag daar geen verandering in brengen."

Meer lezen

Verbeek, P.P. (2021) Coronapaspoort kan ongelijkheid vergroten. *NRC,* 8 juli, pp. 18.

De volledige Unesco-oproep voor de zorgvuldig afgewogen inzet van het coronapaspoort kunt u hier lezen: https://en.unesco.org/news/unescos-ethics-commissions-call-address-ethical-issues-covid-19-certificates.

Meer over Unesco in Nederland: www.unesco.nl.

In memoriam

Melanie Peters is lid van Unesco's internationale bio-ethiekcommissie IBC, en tevens lid van de Nederlandse Unesco Commissie. We hebben met verslagenheid kennisgenomen van het overlijden van Melanie in augustus 2021. Ze was zeer betrokken bij het werk van Unesco in het algemeen, en (bio-) ethiek in het bijzonder.

Podium
voor Bio-ethiek

De NVBe streeft naar:

1. Het stimuleren en expliciteren van de bio-ethiek (medische ethiek, dier- en natuurethiek) rondom actuele maatschappelijke thema's;
2. Het verbeteren van contacten tussen vertegenwoordigers uit verschillende vakgebieden, instellingen en organisaties die betrokken zijn bij bio-ethische kwesties;
3. Open en gelijkwaardige discussies met en tussen stakeholders en andere betrokkenen over bio-ethische kwesties in wetenschap, technologie en samenleving;
4. Aansprekende publicaties over actuele bio-ethische kwesties in Nederland.

Het Podium voor Bio-ethiek draagt bij aan deze doelen met de publicatie van bondige, voor een breed publiek toegankelijke, interdisciplinaire bijdragen over bio-ethische kwesties in de Nederlandse taal en van bio-ethisch nieuws, zowel van binnen als van buiten de vereniging.

Het Podium verschijnt vier keer per jaar en wordt toegezonden aan leden van de NVBe in een gedrukte en/of digitale versie. Het Podium en de mededelingen uit de vereniging zijn ook te vinden op www.nvbe.nl. Nieuwe podiumnummers komen op de website beschikbaar drie maanden na de officiële publicatiedatum.

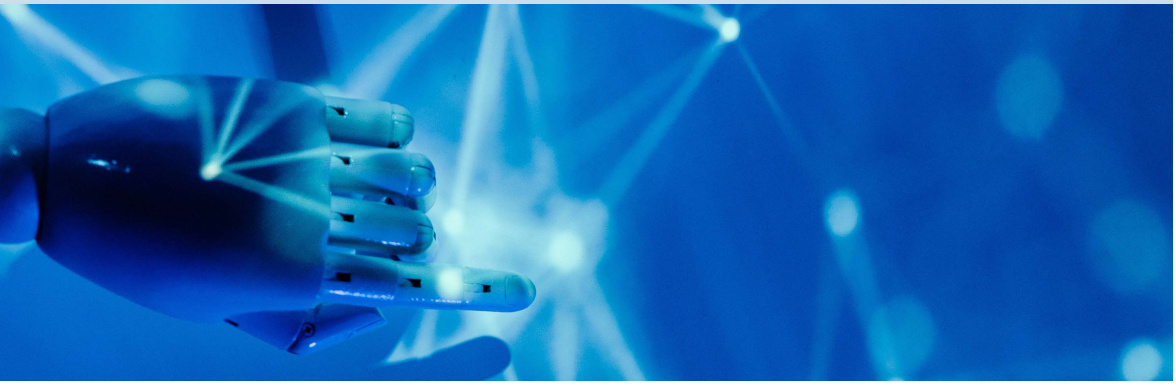
Lid worden?

Het lidmaatschap van de Nederlandse Vereniging voor Bio-ethiek (NVBe) is er voor iedereen die zich op de een of andere manier betrokken voelt bij de levenswetenschappen in brede zin en de ethische reflectie daarop.
Op de website www.nvbe.nl (doorklikken naar 'Lidmaatschap') vindt u een formulier waarmee u zich kunt aanmelden als lid. De ledenadministratie is te bereiken via ledenadministratie@nvbe.nl

Wilt u reageren? Schrijf een brief!

Wilt u reageren op een van de bijdragen in dit nummer, of heeft u iets toe te voegen aan het thema van dit nummer of aan andere onderwerpen die in recente podiumnummers zijn besproken? Dat kan door uw reactie van maximaal 300 woorden te mailen naar podium@nvbe.nl. Gelieve duidelijk in het onderwerp te vermelden 'Brief Podium'. Als uw boodschap een inhoudelijke bijdrage levert aan de discussie en tijdig bij ons binnen is, plaatsen we deze in het eerstvolgende nummer.

NVBe
Nederlandse Vereniging voor Bio-Ethiek



www.nvbe.nl