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.