8e Els Borst Lezing

Centrum voor Ethiek en Gezondheid

# Onbegrijpelijke zorg

## Coreferaat 8e Els Borst Lezing

***Dick Willems, hoogleraar Medische ethiek aan de Faculteit der Geneeskunde van de Universiteit van Amsterdam***

Tamar Sharon laat in de Els Borst Lezing van 2020 overtuigend zien wat de toenemende digitalisering van de zorg, en vooral het inzetten van kunstmatige intelligentie (AI) bij het gebruik van big data, doet met belangrijke waarden in onze zorg. Waarden zoals autonomie in de zin van het vermogen om het eigen leven in te richten, of rechtvaardigheid en gelijke kansen, of transparantie en openheid. Het is een belangrijk betoog en ik ben blij dat ik er in dit commentaar op mag reageren. Ik zal mij richten op een beperkt aantal elementen van Tamar's rijke betoog en daarbij meer aandacht besteden aan zorg en behandeling van ziekte dan aan de digitalisering van gezondheid, op de nieuwe technieken van dataverwerking (AI) en ook op de rol die (empirisch) ethici in deze ontwikkeling kunnen spelen.

Maar eerst dit. Termen als ontwrichting en onder druk zetten zijn niet bepaald juichwoorden. Ik ben niet overtuigd dat we dat soort termen moeten gebruiken. Waarden zijn soms aan vernieuwing toe, en of ze er nu aan toe zijn of niet, ze veranderen gewoon als onderdeel van technische en sociale innovaties. Mijn stelling zou zijn dat belangrijke innovaties nooit alleen technisch of sociaal zijn, maar altijd ook ethische innovaties met zich meebrengen. Met andere woorden: technologische innovaties zoals het gebruik van AI in de zorg zijn altijd ook ethische innovaties. Ze vernieuwen niet alleen de praktijk, maar ook de moraal van die praktijk. Ze verdrukken misschien, maar ze vernieuwen ook. Of misschien zetten ze waarden onder druk als eerste stap in de vernieuwing.

De rol van de empirische ethiek is om die ethische innovaties te articuleren en waar nodig te bekritiseren, niet vanuit een 'armchair' buitenstaandersperspectief, maar vanuit waarden van de praktijk zelf.

Het lastige probleem is natuurlijk of je nog iets te zeggen hebt over die nieuwe waarden als ze zozeer deel zijn van de innovatie – kun je nog zeggen dat je sommige morele innovaties niet zo'n goed idee vindt, en van waaruit zeg je dat dan? Het consistente, maar lastige en bewerkelijke antwoord op die vraag is dat je dat niet doet vanuit een van tevoren bepaalde ethisch raamwerk, maar vanuit de waarden van die praktijk zelf. (dus een ethische innovatie die onverenigbaar is met de kernwaarden van de praktijk in kwestie moet worden verworpen), ook al brengt het ons in een cirkel – een onvermijdelijke ben ik bang.

*Terzijde: ik gebruikte zostraks de term 'articuleren'. Empirisch-ethici spreken meestal over hun werk in termen van het articuleren van in de praktijk levende waarden. Die term heeft een interessante link met het woord 'ontwricht' in de titel van Tamar Sharon's lezing. De Engelse term 'articulation' is potjeslatijn voor gewricht. 'Articuleren' heeft niet alleen te maken met helder praten, maar ook met gewrichten, verbindingen, articulaties. Articuleren is daarmee het tegendeel van het ontwrichten dat Tamar in haar titel noemt. Waar de zorg ontwricht, of gedesarticuleerd raakt, moeten we proberen haar weer te articuleren, 'herwrichten', als dat Nederlands zou zijn. Maar dat terzijde.*

***Een voorbeeld: personalisering van ICDs***

Ik wil dit toelichten aan de hand van een waarde die Tamar noemt, misschien wat impliciet, maar toch: de waarde van transparantie, of doorzichtigheid. Dat wil ik doen vanuit een onderzoeksproject waar ik samen met Marieke Bak als 'embedded ethicist' in werk. Ik denk niet dat het een voorbeeld is van wat Tamar gezondheid als superwaarde heeft genoemd, daar kunnen we het nog over hebben.

Het EU-gesponsorde onderzoeksproject waar ik het over wil hebben heet PROFID (1) en het doel ervan is, heel kort gezegd, om het gebruik van Implantable Cardioverter Defibrillators (ICDs) te personaliseren. Een ICD is in feite een in het lichaam geïmplanteerde Automatische Externe Defibrillator, de bekende groene kastjes in winkelcentra en publieke gebouwen, dus een apparaat dat het hart met een elektrische schok weer op het rechte pad brengt wanneer dat begint te ontsporen en dreigt te stoppen. Deze apparaten worden geplaatst bij mensen die een hartaanval hebben overleefd en die een verhoogd risico lopen om er weer een te krijgen. Nederland heeft ongeveer 60.000 ICD-dragers, er komen er per jaar 3-4000 bij. (2)

Probleem is dat de criteria voor plaatsing niet goed werken, waarschijnlijk omdat ze te algemeen zijn; die houden alleen rekening met de resterende hartfunctie na het infarct (voor de technici: ≤35% ejectiefractie), maar dat blijkt niet zo'n goede voorspeller van de kans om een nieuw infarct te krijgen. Gevolg is dat relatief veel mensen het ding krijgen terwijl ze het niet nodig hebben (het 'klapt' of 'vuurt' nooit), en aan de andere kant ook veel mensen het ding niet krijgen terwijl ze er wel door gered hadden kunnen worden.

Wat is het antwoord? Personalised medicine. Plaatsing van de ICD, zo is het idee van het PROFID project, moet gebeuren op basis van kenmerken van de persoon, niet op algemene richtlijnen, dus geen 'one size fits all'.

*Terzijde: er zijn mensen die het begrip 'personalized medicine' willen vervangen door 'precision medicine'. Precisie is ongetwijfeld 'cooler' dan zoiets vaags als een persoon - precisie is een technologische term met alle associaties van objectiviteit, terwijl 'persoon' ons bij de sociale wetenschappen brengt, of - helemaal erg - bij de humaniora - met alle associaties van subjectiviteit. Ik denk dat we de term 'personalized' moeten behouden, omdat hij veel beter dan 'precision' zegt wat het is: het verfijnen van diagnostiek en behandeling op basis van allerhande, voornamelijk sociale en persoonlijke gegevens van mensen. Maar dat terzijde.*

# 8e Els Borst Lezing

Centrum voor Ethiek en Gezondheid

Het PROFID-project gaat een model ontwikkelen, dat *persoonlijker* kan worden dan de huidige algemene richtlijnen, dat persoonlijker kan voorspellen of iemand een ICD nodig heeft of niet. Op zichzelf een, ook ethisch gezien, loffelijk idee: het voorkomt, als het werkt, onnodige *schade* bij wie het apparaat toch niet nodig heeft, en het '*doet wel*' voor mensen die het apparaat nodig hebben maar het nu niet krijgen. Voor het ontwikkelen van dat model heeft PROFID een immense hoeveelheid gegevens nodig over een immense hoeveelheid mensen die ooit een hartinfarct hebben doorgemaakt en overleefd. PROFID is dan ook een duur project: het kost de EU ruim 20 miljoen. Ik denk dat je daar ook ethische vragen over kunt stellen, maar dat ga ik nu niet doen.

Allerlei gegevens worden verzameld: biologische gegevens over het eerdere hartinfarct, MRI-gegevens, misschien genetische gegevens, maar ook gegevens over leefstijl, sociaal-economische omstandigheden, etc etc. Dit zijn teveel gegevens om in een normale statistische analyse te verwerken tot een voorspellend model. PROFID zal dus Machine Learning (ML) gaan gebruiken.

***(On)doorzichtigheid***

En nu kom ik bij de ondoorzichtigheid waar Tamar het kort over heeft, en die volgens mij één van de belangrijke terreinen van normatieve innovatie zal worden bij deze technische innovatie.

Onder het kopje 'De ondoorzichtigheid van voorspellende analyses' bespreekt zij hoe de ondoorzichtigheid van het 'redeneren' van AI allerlei vormen van bias kan reproduceren. 'Met alle nieuwe categorieën van gegevens die in aanmerking worden genomen bij het doen van gezondheid gerelateerde voorspellingen, zoals eerder besproken, wordt het voor zowel artsen als patiënten erg moeilijk om te doorzien welke overwegingen bij een beslissing een rol spelen.'
Transparantie en uitlegbaarheid staan centraal in de ethische literatuur, en de termen worden vaak door elkaar gebruikt met termen als traceerbaarheid en begrijpelijkheid. *Transparantie* slaat vooral op de ingevoerde gegevens en traceerbaarheid voor de redenering of het mechanisme van de AI. Uitlegbaarheid is dus de overkoepelende term, waarvoor de definitie die wordt voorgesteld door Arrieta et al. handig is: 'Een uitlegbare kunstmatige intelligentie produceert zelf details of redenen om haar eigen werking duidelijk of gemakkelijk te begrijpen te maken.' (3) In deze definitie legt de AI zichzelf uit, net als een dokter haar eigen handelen uitlegt.

# 8e Els Borst Lezing

Centrum voor Ethiek en Gezondheid

De vraag is: in hoeverre moeten professionals en patiënten in staat zijn om de redenering te begrijpen waarmee een ML-gebaseerd model tot zijn conclusies komt? Een gerelateerde vraag is: moeten eisen van transparantie en uitlegbaarheid vergelijkbaar zijn met standaard menselijke redenering en besluitvorming, of is er meer of andere transparantie nodig, juist omdat er sprake is van ML? Immers, in de reguliere zorgsituaties hoeft een patiënt de gegevens en de redenering van een arts niet volledig te begrijpen om vertrouwen in diens advies te kunnen hebben. Maar aan de andere kant kan artsen of onderzoekers worden gevraagd om hun beslissing of advies te rechtvaardigen. Het is misschien moeilijk, maar *niet onmogelijk* om inzicht te krijgen in hun redeneringen. In die zin is de ondoorzichtigheid van de meer complexe vormen van AI, zoals deep learning en neurale netwerken, radicaler dan de ondoorzichtigheid van de menselijke dokter. Want in het geval van complexe AI snapt *niemand* hoe de AI precies tot haar diagnose, prognose of behandelingsadvies komt. Een echte black box, die ook niet te openen valt.
De ethische vraag wordt dan: in welke mate kan en moet een rechtvaardiging worden geëist van ML-gebaseerde modellen? Maar vooral: wat is zo'n rechtvaardiging en wanneer is deze goed genoeg? Hier ontstaan nieuwe concepten (in dit geval van rechtvaardiging) en nieuwe normen – de normatieve innovatie waar ik eerder op doelde.

Bij dit alles is er een belangrijk onderscheid tussen verklaarbaarheid van de manier waarop het model wordt gevormd, met andere woorden de "redenering" van de machine (in sommige literatuur traceerbaarheid genoemd), en verklaarbaarheid van het resultaat. Een deel van de redenering is misschien moeilijk of zelfs onmogelijk uit te leggen, terwijl de uitkomst (intuïtief of anderszins) perfect begrijpelijk kan zijn. AI is meer dan ondoorzichtig, het is vaak onbegrijpelijk en dat maakt kritiek moeilijk of onmogelijk. Daarmee staat een andere waarde onder druk dan de waarden die Tamar noemt, namelijk de bereidheid om je beslissingen uit te leggen. De waarde die we hechten aan verklaarbaarheid en begrijpelijkheid heeft geleid tot een heel nieuw onderzoeksveld van de verklaarbare AI (explainable AI, xAI), een recente ontwikkeling die volgens een van de oprichters tot doel heeft 'machine learning-technieken te creëren die menselijke gebruikers in staat stellen de informatie van komende generaties van kunstmatig intelligente partners te begrijpen, te vertrouwen als dat terecht is en effectief te managen'. (5) De waarde van verklaarbaarheid en bereidheid om uit te leggen wat je doet krijgt in xAI een nieuwe vorm en inhoud.

Het is belangrijk op te merken dat de mate van verklaarbaarheid van het 'redeneren' afhangt van het type AI / ML: neurale netwerken en 'deep' of 'unsupervised' machine learning zijn meer black-box-achtig en minder transparant dan eenvoudigere vormen van ML. In die mate dat de redeneringen van neurale netwerken inherent onbegrijpelijk kunnen blijven - het kunnen zwarte dozen zijn die zelfs niet kunnen worden geopend. Veel AI-onderzoekers en –enthousiastelingen verwachten dat AI steeds moeilijker te begrijpen wordt voor mensen - vooral degenen die geloven dat AI op een gegeven moment over het algemeen intelligenter zal worden dan mensen (de veelbesproken maar ook twijfelachtige "singulariteit"). Bovendien kan het begrijpen van ML-gebaseerde modellen complexer worden als ML wordt gebruikt als een vorm van continu leren, waarbij het model elke keer dat het wordt gebruikt wordt verbeterd. Als dat klopt dan wordt steeds onduidelijker wat de AI precies doet met al die heterogene gegevens over ons.

# 8e Els Borst Lezing

Centrum voor Ethiek en Gezondheid

***Vertrouwen***

Voor zover transparantie te moeilijk, te duur of simpelweg onmogelijk is, zal *vertrouwen* cruciaal worden, zoals dat al het geval is bij complexe besluitvorming zonder big data of AI. Het algemene beeld dat uit empirisch onderzoek naar voren komt, is dat mensen de neiging hebben om AI veel minder te vertrouwen dan menselijke intelligentie, vooral voor belangrijke en vérstrekkende beslissingen met persoonlijke implicaties. Daarom zal een cruciale vraag voor de toekomst van het gebruik van AI in de geneeskunde gaan over vertrouwen, en meer specifiek het onderscheid tussen goed of gepast (on)vertrouwen en slecht of ongepast wantrouwen. (4)
Er is nogal wat ethische en filosofische theorie over wat gerechtvaardigd vertrouwen is en wat niet, bijvoorbeeld in het werk van Annette Baier en anderen, en over het onderscheid tussen overmatig of juist te zwak vertrouwen. Maar die literatuur is tot nu toe grotendeels gericht op vertrouwen tussen mensen en minder op vertrouwen tussen mens en machine of tussen mens en algoritme.
Radicale ondoorzichtigheid en onbegrijpelijkheid maakt vertrouwen tegelijk onmisbaar en moeilijk. Onmisbaar: je hebt vertrouwen nodig voor dingen die je zelf niet snapt of kunt, die je niet in de hand hebt. Als ik de band van mijn fiets plak, hoef ik er niet op te vertrouwen dat het goed is, dat weet ik gewoon. Ik weet precies wat ik gedaan heb, en ik weet ook wanneer het niet goed is. Ook als de fietsenmaker het doet, kan ik nog steeds controleren of het goed gaat (wat in de literatuur vaak ‘confidence’ genoemd wordt); echt vertrouwen (‘trust’) is nodig als je geen idee hebt van waarom en hoe iemand iets doet, en waar je aan kunt zien of deze het goed gedaan heeft. Ik vertrouw erop (in de zin van ‘trust’) dat mijn garagehouder de remmen van mijn auto goed afstelt, ik heb geen enkele manier om dat te controleren.

De kenmerken van gepast vertrouwen (en wantrouwen) in machines, vooral als het echt zo wordt dat die in toenemende mate hun eigen gang gaan, is één van de vraagstukken rond de rol van digitalisering en AI in de zorg, die op het bord van de empirische ethiek liggen.

# 8e Els Borst Lezing

Centrum voor Ethiek en Gezondheid

***Conclusie***

Tamar Sharon heeft in haar Els Borst Lezing laten zien hoe disruptief en vernieuwend de digitalisering en het gebruik van big data in de zorg kan worden. Zeker als de verwerking en het gebruik ervan steeds vaker worden overgelaten aan geautomatiseerde systemen die zelfstandig diagnoses stellen, prognoses maken en behandelingen voorstellen, zeker dan wordt het voor filosofen en ethici een mer à boire voor *empirisch* onderzoek. Want om de normatieve innovatie die deel is van de technische innovatie goed te articuleren (in de dubbele zin van helder verwoorden en verbinden), kunnen ethiek en filosofie niet zonder empirie. Tamar laat zien dat waarden onder druk staan – dat is erg belangrijk, maar het is even belangrijk om te onderzoeken waar en hoe die druk onze waarden ook verbetert.

_____


1. https://profid-project.eu/
2. https://www.stin.nl/
3. Arrieta AB, Diaz-Rodriguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inform Fusion. 2020;58:82-115.
4. Adjekum A, Blasimme A, Vayena E. Elements of Trust in Digital Health Systems: Scoping Review. J Med Internet Res. 2018;20(12):e11254.
5. Gunning D. Explainable Artificial Intelligence (xAI). Defence Advanced Research Projects Agency (DARPA); 2017.