

De Ethiek van Kunstmatige Intelligentie: Verantwoord Ontwerp en de Herkansing van Frankenstein

Prof.dr.ir. Peter-Paul Verbeek, Universiteit Twente (september 2019)

Inleiding

Kunstmatige Intelligentie is zonder enige twijfel de meest gevreesde en tegelijkertijd de meest bejubelde technologie van dit moment. Terwijl de VS, China en de EU in een digitale concurrentiestrijd verwickeld zijn geraakt, waarschuwen invloedrijke techneuten als Elon Musk en Bill Gates voor het gevaar dat deze technologie de mens gaat overheersen, en schetsen bestsellerauteurs als Yuval Noah Harari angstaanjagende toekomstbeelden. Enkele jaren geleden richtte de ethiek zich vooral op 'mensverbetering': zijn we niet onderweg naar een 'supermens', die nieuwe ongelijkheid, onrechtvaardigheid en onwaardigheid met zich meebrengt? Inmiddels gaat de discussie eerder over 'mensvervanging': zijn we onszelf niet kwetsbaar aan het maken voor overheersing door onze eigen technologie? Van angst voor 'te veel mens' zijn we opgeschoven naar angst voor 'te weinig mens'.

Om de juiste ethische vragen te kunnen stellen, schets ik in dit *position paper* allereerst wat deze nieuwe technologie zo uitdagend maakt: haar implicaties voor het menselijke denken, en daarmee voor vele domeinen van de samenleving en het dagelijks leven. Vervolgens analyseer ik het huidige ethische debat en geef ik aan waarom het nodig is de aandacht te verbreden van het waarborgen van individuele autonomie naar het floreren van de gehele samenleving. Tot slot trek ik daaruit conclusies voor de mogelijke rol van Europa in deze ontwikkelingen, en voor de specifieke rol die Nederland daarin kan spelen.

Ik zal betogen dat Nederland de potentie heeft om een kernspeler te zijn op het gebied van ethische AI. De Nederlandse techniekfilosofie en -ethiek heeft al decennia een mondiale toppositie. De huidige nadruk binnen de EU op *Responsible Research and Innovation* (RRI) is bijvoorbeeld een directe opschaling van het al veel langer bestaande Nederlandse NWO-programma Maatschappelijk Verantwoord Innoveren (MVI).¹ Recent werd deze toppositie ook nog bevestigd door de toekenning van een zwaartekracht subsidie² van het ministerie van OCW aan een aantal kernspelers in dit onderzoek, die het mogelijk maakt om deze positie de komende 10 jaar verder uit te bouwen. In de Nederlandse aanpak ligt een sterke nadruk op 'ethisch ontwerpen'. Kerngedachte is dat ethische reflectie niet alleen moet uitmonden in mooie beschouwingen maar ook in de technologie zelf: we moeten niet alleen ethisch *praten* over AI maar ook de algoritmes *zelf* moeten ethisch zijn, en recht doen aan centrale waarden als rechtvaardigheid, respect, en solidariteit.

A. Artificial Intelligence: van 'Industry 4.0' naar 'Society 5.0'

Het jaar 2019 zal de geschiedenis ingaan als het jaar van de ethiek van Kunstmatige Intelligentie, meestal aangeduid als "AI": *Artificial Intelligence*. Bedrijven, overheden, maatschappelijke organisaties: er is haast geen speler in onze samenleving te noemen die niet op de één of andere manier heeft gewerkt aan een ethische code voor het ontwikkelen van en omgaan met AI.³ Reden hiervoor is dat de technologische ontwikkelingen momenteel zo snel gaan en maatschappelijk zo goed voelbaar worden, dat er steeds meer zorgen ontstaan over de beheersbaarheid ervan.

AI stelt computers met behulp van algoritmes in staat om dingen te doen die voorheen voorbehouden waren aan mensen: van diagnoses stellen en beslissingen nemen tot vonnissen vellen en robot-lichamen besturen, van het activeren van de stormvloedkering en het analyseren van het studiegedrag van leerlingen tot het

¹ Zie: Jeroen van den Hoven en Peter-Paul Verbeek, 'Nederland is ethisch gidsland in mondiale hightech' (<https://fd.nl/morgen/1251701/nederland-is-ethisch-gidsland-in-mondiale-hightech>).

² De schrijver van dit position paper is een van de zes *Principal Investigators* in dit tienjarige project, met de titel 'Ethics of Socially Disruptive Technologies'. Voor dit position paper put hij bovendien – naast zijn rol als universiteitshoogleraar in de filosofie en ethiek van technologie – uit zijn ervaring in zijn rol als voorzitter van de UNESCO World Commission for the Ethics of Science and Technology, en zijn lidmaatschap van het bestuur van het Rathenau Instituut en de programmaraad Maatschappelijk Verantwoord Innoveren van NWO, de Nederlandse Organisatie voor Wetenschappelijk Onderzoek.

³ UNESCO heeft een aantal van deze initiatieven verzameld in dit overzicht: <https://unesdoc.unesco.org/ark:/48223/pf0000369756>

genereren van zoekresultaten op het internet en het live vertalen van gesproken woord. Welke invloed gaat deze technologie hebben op mens en maatschappij? En hoe kunnen we de ontwikkelingen in goede banen leiden?

Het is zeer waarschijnlijk dat AI een grote invloed op onze samenleving zal hebben. En het is niet altijd eenvoudig om met die impact om te gaan, omdat ze door bestaande kaders breekt. Tot nu toe zagen we technologie namelijk vaak als 'passief' en de mens als 'actief', maar AI bezit een technologische vorm van 'autonomie'. Algoritmes kunnen leren, en daardoor ontwikkelen ze zich op manieren die hun programmeurs niet altijd hadden kunnen voorzien, en zijn ze in staat om zelfstandig beslissingen te nemen en mensen te adviseren over zaken die het menselijk denken minder goed kan overzien dan een computer met kunstmatige intelligentie.

AI is onderdeel van wat wel de 'vierde industriële revolutie', wordt genoemd. In deze revolutie, waar we midden in zitten, wordt digitale technologie *fysiek*: het internet wordt een internet-der-dingen, en de computer krijgt een lichaam als *robot*. Dit heeft grote impact op de samenleving, die volgens sommigen aan het veranderen is in 'Society 5.0'.⁴ Na de tijd van de jagers-verzamelaars bracht de ploeg de overgang naar de agrarische samenleving, waarna de stoommachine de industriële samenleving mogelijk maakte en de computer tot de informatiesamenleving leidde. De vierde revolutie leidt nu tot 'de digitale samenleving', in Japan ook wel aangeduid als de 'super smart society'.

De kern van de maatschappelijke impact van AI ligt in haar invloed op het *cognitieve functioneren* van de mens: AI verandert het menselijke denken. Historisch gezien is dat niet voor het eerst. Het schrift heeft bijvoorbeeld grote invloed gehad op hoe ons geheugen werkt: Plato was al ongerust dat we niets meer zouden kunnen onthouden als we alles zouden gaan opschrijven. Ook de boekdrukkunst is zo'n cognitieve technologie: opeens was kennis niet meer opgesloten achter kloostermuren, wat leidde tot een democratisering ervan en uiteindelijk de basis legde voor de wetenschappelijke revolutie. Wat is de nieuwe stap in ons denken die AI teweeg gaat brengen? Wat gaat er gebeuren met ons begrip van de wereld en van onszelf, onze morele en politieke opvattingen en besluiten, onze existentiële vragen en kwetsbaarheden? En hoe zal AI invloed hebben op het functioneren van onze democratie, op de relatie tussen burgers en overheid, en op mogelijkheden tot participatie aan de samenleving?

Zoals ook bij veel andere technologieën het geval is, is het zinvol om in de impact van Kunstmatige Intelligentie drie niveaus te onderscheiden. Op *individueel* niveau beïnvloedt AI de keuzes, gedragingen, en interpretatiekaders van individuele gebruikers: zoekmachines kleuren met welke informatie mensen hun wereldbeeld opbouwen, expertsystemen geven artsen en rechters een kader voor een diagnose of een vonnis. Op *sociaal* niveau werkt deze invloed door in sociale domeinen zoals de zorg en de rechtspraak (via expertsystemen), verkeer en vervoer (via zelfrijdende voertuigen), de journalistiek (in de vorm van 'automated journalism'). Op *maatschappelijk* niveau, tenslotte, beïnvloedt AI de organisatie van de samenleving: 'predictive policing' verandert de aard van de handhavende taak van de politie, zoekmachines beïnvloeden politiek engagement, algoritmes zijn betrokken in militaire besluitvorming.⁵

B. Mens en ethiek: van individu naar gemeenschap, en van beoordelen naar begeleiden

De afgelopen tijd is een veelheid aan ethische analyses en kaders verschenen rondom Kunstmatige Intelligentie.⁶ In die analyses komen stevast een aantal kernthema's terug. Allereerst draait de discussie vaak om *bias*: algoritmes kunnen bevooroordeeld zijn, door hun eigen ontwerp of door de data waarmee ze getraind zijn. Een tweede centraal thema is *accountability*: in hoeverre is het mogelijk aansprakelijkheid te organiseren voor het functioneren en de impact van algoritmes, bijvoorbeeld voor besluiten die mede genomen zijn op basis van AI? Daarnaast zijn *transparantie* en *uitlegbaarheid* belangrijke ethische issues: in welke mate kan een systeem uitleggen hoe het tot zijn conclusies is gekomen, zodat mensen kunnen overzien

⁴ Zie: <https://www.gov-online.go.jp/cam/s5/eng/>

⁵ Een korte video met daarin een kernachtig beeld van de ethische uitdagingen van AI is te vinden op: <https://www.youtube.com/watch?v=TiihPx4z0Nw&t=14s>.

⁶ Zie: Roos de Jong, 'Het debat over de betekenis van AI begint pas net': Rathenau Instituut, 2019 (<https://www.rathenau.nl/nl/maakbare-levens/het-debat-over-de-betekenis-van-ai-begint-pas-net>)

op grond waarvan ze het systeem kunnen vertrouwen? En tenslotte wordt ook de *veiligheid* van de systemen wordt vaak benoemd: in hoeverre zijn ze betrouwbaar en bestand tegen vijandig gebruik ('dual use')?

Van individu naar gemeenschap: de noodzaak van een breder debat

Al deze issues gaan impliciet uit van een strijd tussen mens en technologie. Computers kunnen dingen die vroeger alleen mensen konden, maar omdat ze het beter en sneller kunnen, en op een manier die mensen niet altijd helemaal kunnen overzien, worden ze bedreigend. Als computers sommige vormen van kanker al beter kunnen herkennen dan pathologen, wat gaat dat dan betekenen voor de rol van de arts? En wat is de toekomst van de journalistiek als nieuwssites niet alleen gevuld worden door journalisten maar ook door algoritmes? Hoe kunnen we controleren of het systeem de juiste analyse maakt en niet bevooroordeeld is, en kunnen we überhaupt wel verantwoordelijkheid nemen voor beslissingen die tot stand komen op basis van AI?

Ethische issues als deze zijn sterk gericht op het beschermen van de menselijke autonomie tegen de nieuwe vorm van 'autonomie' die de techniek ontwikkelt: is er nog voldoende menselijke controle mogelijk, en sluit het systeem de mens of groepen mensen niet teveel buiten? Niet voor niets richt de *High Level Expert Group for the Ethics of AI* van de EU zich in haar recente rapport⁷ expliciet op 'trustworthy AI': de ethiek moet het vertrouwen tussen mens en techniek waarborgen. Maar er is ook een hele categorie ethische vragen die hier buiten valt. Die richten zich minder op het individu, en meer op de gemeenschap. *Inclusiviteit* is zo'n thema: hoe waarborgen we diversiteit in onze samenleving? *Welzijn* is een ander voorbeeld: hoe ontwikkelen we kunstmatige intelligentie die een positieve bijdrage levert aan het floreren van de gehele samenleving? Deze thema's verdienen meer expliciete ruimte in het debat.

De UNESCO *World Commission for the Ethics of Science and Technology* (COMEST)⁸ heeft daarom een heel andere invalshoek gekozen in deze discussie. Kunstmatige Intelligentie, zo is de gedachte, heeft invloed op de menselijke geest. En dat is precies het domein van de UNESCO, met haar focus op onderwijs, wetenschap, cultuur en communicatie. De *Preliminary Study on the Ethics of Artificial Intelligence* die COMEST onlangs publiceerde⁹ richt zich dan ook precies op deze domeinen. Hoe kan het *onderwijs* mensen opleiden voor een wereld waarin technologie sommige banen overbodig maakt, en hoe kan het kritisch denken stimuleren en ingenieurs leren om AI verantwoord te ontwerpen? Wat gaat AI betekenen voor de *wetenschap*: hoe verandert ons begrip van wat een 'wetenschappelijke verklaring' inhoudt, en op welke 'feiten' baseren we straks beslissingen? Wat doet AI met *cultuur* en culturele diversiteit? Hoe kunnen we ook de kleine talen een toekomst geven als automatische vertaling steeds normaler wordt? Hoe kunnen de suggesties die algoritmes van platforms als Netflix en Spotify doen aan hun gebruikers ook ruimte maken voor diversiteit? En, tot slot: hoe kan de kwaliteit van menselijke *communicatie* worden bewaakt, als AI nieuwe vormen van desinformatie mogelijk maakt, en steeds nauwer verweven raakt met de journalistiek?

In deze analyses neemt COMEST expliciet een mondiaal perspectief in, dat niet alleen is gericht op het waarborgen van individuele vrijheid maar ook op het floreren van de gemeenschap. Daarom vraagt COMEST ook aandacht voor een aantal bredere 'global ethics' thema's, die in het werk van UNESCO altijd terugkeren: Afrika, gender, duurzaamheid en vrede. Hoe zorgen we ervoor dat Afrika volop betrokken blijft in de ontwikkeling en implementatie van AI? Hoe kunnen we voorkomen dat AI nieuwe vormen van gender-bias creëert? Hoe realiseren we een duurzame technologische infrastructuur voor AI, in termen van zeldzame aardmetalen en de energieconsumptie van servers? En hoe zorgen we ervoor dat AI vreedzaam en niet gewelddadig wordt gebruikt? Tijdens de Algemene Conferentie in november 2019 zullen de lidstaten van UNESCO gezamenlijk besluiten of er een wereldwijd normatief instrument ten aanzien van AI (een zogeheten 'Recommendation') zal worden ontwikkeld.

Ethiek als begeleiding van technologie: de herkansing van Frankenstein

Om deze ethische thema's goed te kunnen agenderen is het belangrijk om kunstmatige intelligentie niet *tegenover* de mens te plaatsen maar te beschouwen als *verbonden* met de mens. Feitelijk is de discussie over AI een digitale versie van het Frankenstein-verhaal. In de vierde revolutie komt techniek alleen niet tot leven

⁷ Zie: <https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>.

⁸ De auteur van dit *position paper* is voorzitter van deze commissie.

⁹ Zie: <https://unesdoc.unesco.org/ark:/48223/pf0000367823>.

via vlees en bloed, maar via algoritmes en robots: of het nu gaat om technologische werkeloosheid, algoritmische besluitvorming of geautomatiseerde journalistiek, steeds is er de vrees dat technologie op de stoel van de mens gaat zitten, en schade aanricht. En hoewel de vrees voor negatieve implicaties van AI zeker niet ongegrond is, is het tegenover elkaar plaatsen van mens en technologie dat wel. Om de implicaties van AI helder te kunnen begrijpen is het immers juist nodig om te zien hoe AI *onderdeel* wordt van menselijke praktijken en vorm geeft aan interpretatiekaders en beslissingen – net zoals het schrift en de boekdrukkunst dat eerder deden.

Dit is precies wat er mis ging bij Frankenstein. Zijn monster werd alleen maar een monster omdat hij het in de steek liet. Victor Frankenstein raakte na de euforie over het tot leven komen van zijn schepsel namelijk zo in paniek dat hij wegvlochtte en het ‘monster’ aan zijn lot overliet. Dat monster moest daardoor helemaal alleen zijn weg zien te vinden in de wereld. Vrij vertaald: als we willen voorkomen dat algoritmes en robots ontspreiden, dan moeten we ze niet vrezen, maar opvoeden. Als we ons uitsluitend laten regeren door de angst voor wat er allemaal kan gebeuren, laten we ze in de steek in plaats van ze een goede plek te geven in de samenleving. Wie blijft hangen in de keuze tussen ‘toestaan’ of ‘verbieden’ slaagt er niet in om door te dringen tot de werkelijke impact van AI.

Dat ‘opvoeden’ heeft in het geval van AI bovendien een heel letterlijke betekenis. De crux van kunstmatig intelligente systemen is immers dat ze *leren*. Ze ontwikkelen zich zelfstandig in interactie met hun omgeving en gaan dingen doen die hun ontwerpers er nooit zelf in hebben gestopt. Uiteindelijk is daardoor niet meer de ontwerper of de omgeving verantwoordelijk, maar het systeem zelf. En dus moet ethiek zich ook op die systemen zelf richten en op de manier waarop ze leren. Welke datasets gebruiken we om ze te trainen? En hoe bewerken de algoritmes deze datasets?

Ethiek van kunstmatige intelligentie moet dan ook veel meer zijn dan het stellen van de vraag welke AI we wel en niet willen. In plaats van AI alleen van buitenaf te beoordelen moet de ethiek AI van binnenuit begeleiden. Door centrale ethische waarden een plek te geven in het verantwoord ontwerpen ervan, en daarnaast in haar inbedding en gebruik. Algoritmes belichamen waarden, ontwikkelen zich in de richting van bepaalde waarden door de manier waarop ze getraind worden, en worden vanuit waarden geïmplementeerd en gebruikt. Het selectie-algoritme waarmee Amazon nieuw personeel zocht is hiervan een iconisch voorbeeld geworden: door het profiel van de zittende succesvolle medewerkers als model te nemen, bleek het algoritme uiteindelijk alleen nog blanke mannen tussen de 40 en de 50 te selecteren en werd het personeelsbestand wel erg eenzijdig en discriminatoir. Doordat alleen de waarde van ‘succesvol zijn’ een plaats had, en ‘inclusiviteit’ en ‘diversiteit’ niet expliciet werden meegenomen in de ontwikkeling en training van het algoritme, had deze ontwikkeling negatieve gevolgen die voorkomen hadden kunnen worden.

‘Responsible Design’ en ‘Responsible Guidance’ van technologie zijn daarom minstens zo belangrijk voor de ethiek van AI als het opstellen van codes en kaders. Op het gebied van Responsible Design is veel expertise in Nederland¹⁰; Responsible Guidance, dat zich sterker richt op implementatie en gebruik, is een thema dat momenteel op landelijk niveau wordt uitgewerkt vanuit de werkgroep Ethiek en Digitalisering van ECP, waarbij niet alleen kernspelers uit de Nederlandse overheid en techniekethiek aangesloten zijn, maar ook grote bedrijven als IBM, Microsoft en Facebook.¹¹

C. Europa: van ‘derde weg’ tot verbinder

De afgelopen jaren is de rol van Europa in de ontwikkeling van AI vaak gepositioneerd als alternatief voor die van de Verenigde Staten en China. Als een soort *revival* van de 20^e-eeuwse zoektocht naar een ‘derde weg’ tussen het kapitalistische Westen en het communistische Oosten, zoekt de EU nu naar een kader voor AI waarin niet alleen commerciële belangen van bedrijven de hoofdrol spelen, en evenmin de behoefte van de

¹⁰ Deze expertise bevindt zich onder andere in het [Delftse ‘Design for Values Institute’](#) en het [DesignLab](#) van de Universiteit Twente. Vanuit de Universiteit Twente wordt ook een [MOOC](#) aangeboden op het gebied van de filosofie en ethiek van techniek en design. Zie ook: Verbeek, P.P. (2011), [Moralizing Technology: Understanding and Designing the Morality of Things](#). Chicago and London: University of Chicago Press.

¹¹ Een breed toegankelijke publicatie over begeleidingsethiek is voorzien begin 2020.

staat om de samenleving te kunnen sturen. Deze zoektocht heeft onder andere geleid tot het al genoemde ethische kader van de EU *High Level Expert Group on the Ethics of AI*.

In plaats van deze 'derde weg' te positioneren als een *alternatief* voor China en de VS zou het beter zijn deze uit te werken als *verbinding*. Vanuit het mondiale kader dat de UNESCO *World Commission for the Ethics of Science and Technology* heeft neergelegd, is juist een goede verbinding van individuele vrijheid en verantwoordelijkheid voor het collectief essentieel. Wanneer de EU zich presenteert als 'het ethische alternatief' of zoekt naar ruimte om de derde 'grote speler op de markt' te zijn, wordt die kans op verbinding onnodig beperkt. Gezien de te verwachten mondiale impact van AI, is het juist zaak om niet verder te polariseren, maar de verbinding op te zoeken, op mondiale fora.

De EU is voor deze verbinding van cruciaal belang, niet alleen als mede-ontwikkelaar van technologie, maar ook als proeftuin voor succesvol beleid ten aanzien van technologie. Hoeveel kritiek er soms ook mag worden geuit op de effectiviteit en zinvolheid van Europees beleid en Europese wetgeving: als er één domein is waarin de EU uitermate effectief en belangrijk werk heeft gedaan, dan is het wel de informatietechnologie. De recent EU-breed ingevoerde General Data Protection Regulation is daar een levend voorbeeld van. De EU slaagt er daadwerkelijk in om individuele en collectieve belangen met elkaar in balans te brengen, en staatsinterventie samen te laten gaan met vrijheid en democratie. Naar dergelijke vormen van balans moeten we blijven zoeken: de mogelijkheid om verantwoord om te gaan met de impact van AI op de mensheid is ervan afhankelijk. De voorhoedeopdracht die de EU heeft op het gebied van *Responsible Research and Innovation* vormt hiervoor een goede basis: er is ruime ervaring met het integreren van ethische en maatschappelijke reflectie in technologische ontwikkeling, en die ervaring moet ten volle worden benut en uitgebouwd.

D. Organisatie en beleid: naar *Responsible AI* en een Minister en een Raad voor Digitalisering

De verreikende impact van AI op de samenleving, de noodzaak om tot een mondiaal kader voor ethische reflectie te komen, en de ruime Nederlandse expertise in Maatschappelijk Verantwoord Innoveren maken gezamenlijk een aantal richtingen voor organisatie en beleid zichtbaar:

- Er zou een nationaal programma voor *Responsible AI* moeten komen, waarin bedrijven, overheden en onderzoekers samenwerken volgens het beproefde kader van Maatschappelijk Verantwoord Innoveren (www.nwo-mvi.nl). Parallel daaraan moet Nederland zich inspanssen voor een EU-breed programma met dezelfde missie.
- Nederland zou binnen UNESCO het voorstel voor een *Recommendation on the Ethics of AI* moeten steunen, en een actieve rol moeten spelen in de verdere uitwerking ervan, met bijzondere aandacht voor *Responsible Research and Innovation*.
- Vanaf de komende kabinetsformatie zou Nederland een minister voor digitalisering moeten hebben. Deze minister heeft interdepartementale verantwoordelijkheid, stelt samen met mede-bewindspersonen een digitaliseringsprogramma op in alle beleidsgebieden, en is verantwoordelijk voor de uitvoering daarvan; daarnaast werkt zij of hij actief aan Europese en mondiale initiatieven voor het verantwoord ontwikkelen en implementeren van nieuwe digitale technologie.
- Er zou een Raad voor Digitalisering moeten komen, naar het model van de Gezondheidsraad, met onafhankelijke experts die de overheid gevraagd en ongevraagd kunnen adviseren en grotere studies kunnen uitvoeren om het digitaliseringsbeleid een steviger basis te kunnen geven.

Peter-Paul Verbeek (1970) is universiteitshoogleraar Filosofie van Mens en Techniek en wetenschappelijk co-directeur van het DesignLab van de Universiteit Twente. Hij is tevens *honorary professor Techno-Anthropology* aan Aalborg University, Denemarken. Hij is voorzitter van de *World Commission on the Ethics of Science and Technology* (COMEST) van UNESCO en lid van de *Programmaraad Maatschappelijk Verantwoord Innoveren* van NWO, het bestuur van het *Rathenau Instituut*, de *Raad voor Geesteswetenschappen* (KNAW), de *Commissie voor de Vrijheid van Wetenschapsbeoefening* (KNAW) en de *Raad van Toezicht van TNO*. Verbeek is lid van de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), de Academie voor Technologie en Innovatie (AcTI) en de Koninklijke Hollandse Maatschappij van Wetenschappen (KHMW). In 2012 ontving hij de Prof. Borghgraef prijs voor biomedische ethiek (KU Leuven), in 2016 de World Technology Award in Ethics (World Technology Network). Meer informatie: www.ppverbeek.nl.