Position Paper AI
Thema: Mens en ethiek

Parlement en Wetenschap

Dr. Katleen Gabriels
Maastricht University
k.gabriels@maastrichtuniversity.nl

September 2019

Inhoudstafel

A. WAT IS AI? 2
B. VERDIEPING 3
C. EUROPESE DIMENSIE 6
D. ORGANISATIE EN BELEID 7

Bijlagen
Leessuggesties (boeken) 10
Ethische code (Duitsland) zelfrijdend transport 11

Eindnoten 14

A. WAT IS AI?
Beschrijving AI en het wenkend perspectief

Artificiële intelligentie (AI) begon in de jaren 1950 als onderzoeksveld om menselijke intelligentie in een machine na te bootsen.[1] De opzet was om machines op een vergelijkbaar niveau als de mens taken te laten uitvoeren, bijvoorbeeld redeneren.[2] Een hedendaagse, meer specifieke definitie omschrijft AI als "een verzamelnaam voor technieken zoals *machine learning*, *deep learning* en neurale netwerken die tot doel hebben om computers te laten leren uit hun ervaringen".[3] Technologie waar we dagelijks mee werken – zoals de zoekmachine van Google en de tijdslijn van Facebook – bevat AI-toepassingen. In de eerste industriële robot in de jaren 1960 zat geen AI, maar vandaag zijn robots containers van AI of de *belichaming* van AI.[4] De combinatie van robots en AI leidt tot ongelooflijk veel mogelijkheden.

AI bestrijkt een breed en interdisciplinair veld: filosofen, wiskundigen, neurowetenschappers, psychologen, computerwetenschappers, linguïsten, economen, biologen, juristen, enzovoort. AI kende de voorbije decennia pieken en dalen, de zogenaamde 'AI-winters' met verminderde interesse en afgenomen enthousiasme. Het was ook moeilijk om de hoge verwachtingen in te lossen. Vandaag zitten we in een 'AI-zomer'.

AI is een 'gamechanger' door **1.** nieuwe technologische ontwikkelingen en applicaties, **2.** toegenomen opslagcapaciteit, snelheid en rekenkracht (die, volgens de Wet van Moore, elke achttien maanden verdubbelt), en **3.** big data die gigantische datasets aanleggen waar algoritmen mee getraind worden.[5] Het zijn de datasets die tot nieuwe ethische bezorgdheden leiden: vooringenomen ('biased') datasets, hypernudges[6] en personalisering geven problemen op het vlak van transparantie, (gebruikers)controle, privacy en autonomie.

Filosofen maken vaak het onderscheid tussen sterke en zwakke AI.[7] Bij zwakke AI gaat het om het uitvoeren van specifieke taken en het opleggen van een beperkte doelstelling. Machines handelen *alsof* ze intelligent zijn (simulatie). Spraakassistent Siri (Apple), bijvoorbeeld, werkt binnen een vooraf gedefinieerde set van functies. Sterke AI heeft zelfbewustzijn en een vrije wil en *is* intelligent – die is er nog lang niet en het is maar de vraag in hoeverre die er ooit komt.[8] Een aantal prominente AI-onderzoekers, waaronder Max Tegmark, hanteert de tweedeling zwakke–sterke AI liever niet, omdat AlphaGo (DeepMind) onder zwakke AI valt. De tweedeling onderschat volgens hem waar deze systemen al tot in staat zijn.[9]

Het is cruciaal om een onderscheid te maken tussen wat AI *al* kan, *nog niet* kan en *mogelijk nooit* zal kunnen, ook om een nieuwe 'AI-winter' te vermijden en om angsten bij mensen weg te nemen.[10] Taal en emoties vormen nog een grote uitdaging voor AI[11] en om die reden moet kritisch omgegaan worden met commerciële toepassingen (vele ervan zijn nog onbetrouwbaar[12]) en hard ingezet worden op verder onderzoek (binnen Natural Language Processing (NLP) en 'emotion recognition').

De toepassingsgebieden zijn erg divers: van geneeskunde en de zorgsector, tot het openbare leven, inclusief zelfrijdend transport, om slechts enkele voorbeelden te geven. Big data en het Internet of Things (IoT) dragen bij aan AI-mogelijkheden en staan er zeker niet los van.

Wat geneeskunde betreft:
Esteva et al. (2017) trainden diepe neurale netwerken met een dataset van 129450 klinische beelden van huidletsels ('skin lesions').[13] Hun resultaten tonen dat AI op het niveau van een dermatoloog staat in het voorspellen van huidkanker. Dit is maar één voorbeeld van de vele toepassingsmogelijkheden binnen geneeskunde.[14] Die mogelijkheden zullen tot verschuivingen leiden binnen een aantal artsenspecialisaties (bv. dermatologie, radiologie, …) en de curricula van geneeskundestudenten moeten daar op aangepast worden. Robotchirurgie, zoals de 'Da Vinci robot' die precisieoperaties uitvoert die risicovol zijn voor mensenhanden (omdat die beven), leidde al tot transities binnen de chirurgie. Opleidingen moeten sterker voorbereiden op dataverwerking- en toepassingen. Uiteraard is het essentieel om kritisch te blijven en de mogelijkheden steeds evidence-based te onderzoeken. Verder zijn ook verantwoordelijkheid, vertrouwen en onzekerheid ('uncertainty') sleutelbegrippen. Het gaat

ook niet zomaar over één toepassing, maar over robots in de zorgsector, AI voor diagnostiek en behandeling, digitale self-tracking technologieën[15] en dergelijke meer. Dit zijn uiteenlopende zaken die elk hun eigen uitdagingen meebrengen.[16]

Wat het openbare leven betreft:
De vraag die onvermijdelijk rijst als het publieke leven meer en meer gereguleerd wordt door AI-systemen (en, daarmee samenhangend, IoT en big data – denk maar aan de slimme stad), is in hoeverre hierdoor de aanwezigheid van keuzes, die noodzakelijk zijn voor vrijheid en zelfbepaling, wordt ingeperkt. Een overheid moet een evenwicht vinden tussen openbare veiligheid en individuele privacy: de overheid staat in voor de bescherming van haar burgers, maar mag niet zo ver infiltreren dat die burgers niet meer kunnen genieten van vrijheid en anonimiteit. We mogen de huidige 'paradox van openheid' niet negeren: burgers en consumenten worden zichtbaarder voor respectievelijk overheden en commerciële bedrijven, maar ironisch genoeg moeten zij net transparanter over hun intenties communiceren en hierover in dialoog treden. Openheid is essentieel om vrijheid en privacy te garanderen. Er moet open gecommuniceerd worden over hoe en waarom data verzameld worden en waarvoor ze ingezet worden, op korte en lange termijn. Kortom: de voorwaarden moeten duidelijk zijn.

De problematiek inzake AI & ethiek (en het openbare leven/de slimme stad) komt duidelijk naar voren in de discussie over zelfrijdend transport. Om die reden heb ik in de bijlagen een tekst over de Duitse code voor zowel gedeeltelijk als volautomatisch zelfrijdend transport toegevoegd: de code bestaat uit twintig regels en geeft een interessante inkijk in het ethisch benaderen, analyseren en reguleren van AI-toepassingen.

AI op Nederlandse universiteiten
Mocht dit nog niet gebeurd zijn, dan adviseer ik om de AI-onderzoeksgebieden aan de Nederlandse universiteiten in kaart te brengen en met die van buitenlandse universiteiten te vergelijken. Zelf heb ik er geen gedetailleerde inzage in.

Er zit veel kennis aan de Nederlandse universiteiten en academici moeten meer gestimuleerd worden om het maatschappelijke debat te voeden met die kennis (wetenschapscommunicatie). Vandaag waarschuwen academici[17] onder andere voor commerciële toepassingen van 'emotion recognition algorithms', omdat de onzekerheid, en dus ook de onbetrouwbaarheid, erg hoog is; om die reden wordt geadviseerd om producenten te verplichten om de accuraatheid van het model (algoritmen) er expliciet bij te vermelden. Grote bedrijven kan je in die context ook verplichten om de te verwachten foutmarge, op basis van data die de realiteit weerspiegelen, erbij te vermelden. De modellen worden getraind op door mensen gelabelde data. Daarom is het raadzaam om hen op te leggen om (een deel van) hun data beschikbaar te maken, samen met hun resultaten op die data (die test moet gebeuren op een dataset die enkel dient om de prestatie van de modellen op te evalueren, m.a.w. zeker de test en de resultaten erop moeten vrijgegeven worden). Dit is vooral nog een academische discussie die het brede publiek niet of in elk geval te weinig bereikt.[18]

B. VERDIEPING
Mens en ethiek[19]

Moral disengagement
Een belangrijk aspect van moraliteit is 'moral disengagement', een mechanisme dat we in het dagelijkse leven toepassen om ons eigen bedenkelijke gedrag te rechtvaardigen.[20] Een voorbeeld is dat we door taalgebruik de focus van de handeling proberen weg te leiden. Als in een bedrijf een grote ontslagronde plaatsvindt, dan kun je de focus van het menselijk leed afwenden door te stellen dat je het bedrijf weer 'gezond' maakt of dat je het 'redt' van de ondergang, waardoor het bijna als een heldendaad klinkt. 'Verspreiding van verantwoordelijkheid' is er een andere bekende vorm van: als veel mensen samen verantwoordelijk zijn, bijvoorbeeld voor het ontwerpen van een moreel bedenkelijk product, voelt niemand zich nog echt verantwoordelijk. 'Vervanging van verantwoordelijkheid' is er ook

een vorm van: onderzoek toont aan dat mensen die zich niet verantwoordelijk voelen, omdat ze gehoorzamen aan een autoriteit die de verantwoordelijkheid op zich neemt, tot verregaande negatieve gevolgen kan leiden.[21]

Moral disengagement vormt een 'makkelijke' manier om verantwoordelijkheid af te schuiven op het algoritme of het AI-systeem: 'Het algoritme heeft beslist'. Geen enkele robot of AI-toepassing ontwerpt zichzelf: ménsen maken die keuzes en zijn dus verantwoordelijk. Onderzoekers Deborah Johnson en Mario Verdicchio noemen dit 'sociotechnische blindheid': blindheid voor alle betrokken menselijke actoren en alle beslissingen (genomen door mensen) die nodig zijn om AI-systemen te maken.[22]

Technologie is 'made with morality'

Wie ontwerpt, maakt voortdurend keuzes: zowel functionele als morele. Moraliteit en technologie zijn dus geen afgescheiden domeinen. Techniekontwikkelaars, ingenieurs, computerwetenschappers, programmeurs, …:[23] ze zien zichzelf vaak als neutrale, uitvoerende spelers, werkzaam binnen de exacte wetenschappen. Zonder het zo expliciet te maken of te zien, kunnen ze hun eigen leefwereld, moreel kader of geslacht echter als norm nemen, met als gevolg dat anderen gediscrimineerd of genegeerd worden. Pas door expliciet aandacht te schenken aan ethische aspecten van ontwerp worden die blinde vlekken zichtbaar.

Doordat ze het gedrag van gebruikers kunnen sturen, hebben ontwerpers veel macht en verantwoordelijkheid. Ons gedrag en keuzes zijn in grote mate afhankelijk van de context waarin we handelen, en technologie maakt deel uit van die context.[24]

Op minstens vier niveaus speelt ethiek een cruciale rol: dat van de maatschappij, de technologie zelf die het resultaat is van (morele) keuzes, de maker (incl. onderzoeker en wetenschapper) en gebruiker. Lange tijd werd (techniek)ethiek beschouwd als iets dat *achteraf* komt, nadat een technologie ontwikkeld is. Maar ethische vragen over ontwerp moeten net gesteld worden vóór en tijdens het ontwerpproces: elk product dat door mensen gemaakt is, is 'made with morality'.

Bij AI-, machine- en robotethiek ligt de focus ook op vragen over wat machines zélf kunnen, zoals het nemen van *autonome* beslissingen zonder menselijke tussenkomst. Mensen zijn in deze context vaak misleid door het begrip 'autonomie': dit betekent niét dat het buiten menselijke controle valt. 'Autonomie' betekent in dezen het bereiken van een doelstelling die niet op voorhand al volledig werd vastgelegd door programmeurs.[25] Hoe minder menselijke interventie er is in de werking en hoe breder de reikwijdte van de actie, des te autonomer de machine.[26] Hierdoor stijgt menselijke morele en professionele verantwoordelijkheid: denk maar 'autonome' drones in oorlogsvoering. Zorgrobots en dergelijke meer verdienen onze aandacht, niet omdat ze zelfbewust (cf. sterke AI) handelen, maar wel omdat ze deel zullen uitmaken van onze maatschappij en daardoor onder onze morele evaluatie vallen (en uiteraard ook omdat ze het resultaat zijn van mensenwerk).

Bestaande richtlijnen (internationaal) voor op de werkvloer

Er zijn al praktische, ethische richtlijnen voor algoritmen (AI-toepassingen). Het is raadzaam om te vertrekken van bestaande lijsten en die verder uit te bouwen.
Zie bijvoorbeeld:

1. https://www.fatml.org/resources/principles-for-accountable-algorithms
2. https://ethicalos.org/
3. https://cdt.org/issue/privacy-data/digital-decisions/
4. https://open.canada.ca/data/en/dataset/748a97fb-6714-41ef-9fb8-637a0b8e0da1 (voorbeeld van Algoritmic Impact Assessment, deel van de Canadese AI-strategie)
5. https://data-service-alliance.ch/innovation/ethics

FAT/ML ('Principles for Accountable Algorithms and a Social Impact Statement for Algorithms') staat voor 'fairness', 'accuracy' en 'transparency' bij machine learning. Een groep van academici en onderzoekers (Google Research en Microsoft Research) ontwikkelde deze lijst, bestaande uit principes en richtlijnen, om bewustwording op het vlak van verantwoordelijkheid (en aansprakelijkheid) bij ontwikkelaars te stimuleren en verhogen. De gerichte vragen van FAT/ML kun je bijvoorbeeld in een webformulier gieten en uitbreiden met

specifieke vragen over data, zoals bij de Canadese overheid (zie link 4). Ook de Zwitserse codex (zie link 5) besteedt aandacht aan de onderliggende data.[27] Maar het is aangewezen om ontwikkelaars ook persoonlijk aansprakelijk te stellen voor de algoritmen die ze ontwerpen (cf. infra, punt D).

Van 'wie' leren zelflerende AI-systemen?

Tay, de AI-spraakrobot van Microsoft, werd in maart 2016 losgelaten op Twitter. Tay moest 'leren' uit tweets. Mensen konden op Twitter een conversatie met Tay aangaan, zodat de bot waarheidsgetrouwe antwoorden leerde geven. Het was de bedoeling dat Tay leerde converseren zoals een Amerikaans pubermeisje, maar door de input die de chatbot van Twitter-gebruikers kreeg, werd het taalgebruik al snel gemeen en racistisch.[28] Toen het hopeloos de verkeerde kant uitging, haalde Microsoft Tay van het internet. Microsoft publiceerde nadien een publiek statement waarin ze toegaven dat de uitdagingen niet alleen technologisch, maar ook sociaal en ethisch zijn.[29] Ook bij Zo, de opvolger van Tay, liep het mis.[30]

Een essentiële vraag is: van 'wie' leren AI-systemen, nu en in de toekomst? Hier moet diepgaand over nagedacht worden. Uiteraard zijn kwaliteitsvolle datasets (niet vooringenomen, door meerdere mensen onafhankelijk gelabeld, enzovoort) essentieel, maar ook interdisciplinariteit en diversiteit in ontwerpteams.

Wat wel en wat niet delegeren aan AI-systemen?

Dit is een menselijke keuze en verantwoordelijkheid: "Wie beslist welke AI-systemen worden gebouwd en in een context worden geplaatst? Wie beslist en hoe worden beslissingen genomen over welke taken aan mensen moeten worden overgedragen en welke aan machines? Hoe worden de mensen die werken binnen AI-handelssystemen, zelfrijdende transportsystemen of dronesystemen getraind?"[31] Het is ook belangrijk om een aantal beslissingen er niét aan uit te besteden en het oordeel bij de mens te laten. Norbert Wiener, de Amerikaanse wiskundige die in de jaren 1940 de informatietheorie 'cybernetica' ontwikkelde, waarschuwde daar al voor.[32]

We hebben goede en duidelijke regels nodig, zeker als het machines betreft die zelf beslissingen nemen, in een fractie van een seconde en zonder menselijke supervisie. Niet doden kan één van die regels zijn, dus geen gewapende autonome 'decision-making' (besluitvorming) drones inzetten voor militaire doeleinden. In België is er al een verbod op zogenaamde 'killerrobots': er mag wel onderzoek naar gebeuren, maar de productie ervan is verboden. Stephen Hawking schreef in zijn postuum uitgegeven boek *Brief Answers to the Big Questions*: "The best time to stop the autonomous-weapons arms race is now". Hiervoor is ook een internationale samenwerking en strategie nodig.[33]

Privacy

De privacyproblematiek zal door AI alleen maar toenemen. Privacy is niet enkel een mensenrecht (Art. 12 Universal Declaration of Human Rights), Europees recht (Art. 8 The European Convention on Human Rights) en kinderrecht (Art. 16 Convention on the Rights of the Child), maar ook een ethische waarde die raakt aan andere waarden zoals vrijheid, autonomie en sociale rechtvaardigheid.

Veel robots, zoals huishoudelijke maar ook zorgrobots, verzamelen persoonlijke gegevens. Lynx, een slimme huisrobot die is uitgerust met Amazons Alexa, bevat bijvoorbeeld een camera en ingebouwde microfoon. Lynx heeft ook een alarmfunctie, die beweging in huis detecteert als je niet thuis bent en een filmpje doorstuurt. Bij zo'n apparatuur is het niet altijd duidelijk wat er met die data gebeurt. Om zich te kunnen positioneren in je woning, verzamelt de Roomba data. Zulke ontwerpkeuzes zijn nodig voor de functionaliteit van het apparaat, maar ze hebben ook gevolgen op het vlak van privacy. In 2017 bleek dat de CEO die data wilde verkopen aan bedrijven zoals Amazon.[34]

De voorbije maanden berichtten nationale en internationale media dat werknemers van Google (Google Home; Google Assistant)[35], Apple (HomePod)[36] en Amazon (Echo)[37] opgenomen gesprekken beluisteren. Al die data maken bovendien problemen rond personalisering (bv. op maat gemaakte advertenties) en traceerbaarheid alleen maar urgenter. Google, Apple en Amazon vormen samen met Facebook en Microsoft 'the big five'; hun

marktconcentratie blijft problematisch. Om de dominantie tegen te gaan, moet er een rem komen op het gemak waarmee zij andere bedrijven opkopen. Facebook nam WhatsApp en Instagram over. Alphabet onder meer YouTube, Nest Labs en Doubleclick, waardoor je als gebruiker steeds minder keuze krijgt én het overzicht verliest van welke bedrijven allemaal deel zijn van Alphabet. Start-ups die mogelijk een bedreiging vormen, worden preventief opgekocht. Een strenge regulering vereist beleidsmakers die kritisch en onafhankelijk blijven, ongeacht het legertje lobbyisten dat namens Google, Apple, Facebook, enzovoort Brussel afschuimt. Als burger heb je het raden naar de werkwijzen en precieze invloed van lobbyisten.

Toenemende surveillance, maar ook coveillance (peer-to-peer monitoring) staan haaks op wat individuen belangrijk vinden voor een goed leven: autonomie. De menselijke behoefte aan autonomie is uitgebreid bestudeerd in verschillende contexten en culturen, en wordt bevestigd als een universele psychische basisbehoefte.[38]

Doordat zoveel data verzameld worden, staat het anonimiseren ervan sterk onder druk. Hoe hard je ook probeert om bepaalde gegevens, zoals je identiteit, uit data te halen: de kans is reëel dat je door geavanceerde data-analyse en cross-referencing met andere data, zoals een publiek beschikbaar LinkedIn-profiel, kan achterhalen om wie het gaat. AI-systemen worden almaar beter in deanonimisering en re-identificatie. Een recente studie in *Nature* toonde aan dat een AI-model in staat zou zijn om 99.98% van de Amerikanen correct te re-identificeren in eender welke dataset op basis van vijftien demografische kenmerken.[39] Deze problematiek moet grondig verder uitgezocht worden.

C. EUROPESE DIMENSIE

Korte reflectie op ‘mensgerichte, ethische AI’ als onderscheidende focus en beschrijving van het Europese speelveld

Deze webpagina bevat een overzicht van nationale (per land) en internationale (bv. EU; VN) AI-strategieën: https://futureoflife.org/national-international-ai-strategies/[40]

Mocht dit nog niet gebeurd zijn, dan adviseer ik een studie die een diepgaande vergelijking tussen al de bestaande AI-strategieën uitvoert, ook om te kijken hoe Nederland een verschil kan maken en zich kan onderscheiden (de Nederlandse AI-strategie mag niet simpelweg inwisselbaar zijn met reeds bestaande nationale strategieën en moet clichés vermijden, want vrijwel elk land stelt in hun strategie ‘leading’ te willen zijn), maar ook hoe Nederland zich aansluit bij internationale strategieën en andere landen. Denemarken zet bijvoorbeeld hard in op mediawijsheid (digitale competenties in onderwijs). De drieledige AI-aanpak van de Europese Commissie is algemeen maar interessant, omdat ze onder meer inzet op modernisering van onderwijs.[41]

AI moet mensen dienen en niet omgekeerd. De mensgerichte focus is niet nieuw: de eerder vermelde Amerikaanse wiskundige en ingenieur Norbert Wiener benadrukte in 1950 al dat humane waarden steeds voorop moeten staan in ‘the automatic age’, waar mens en maatschappij in toenemende mate afhankelijk zijn van informatie- en communicatietechnologie (ICT).[42] De ‘automatic age’ is, kortom, de tijd waarin wij leven. Hij besteedde daarbij ook aandacht aan de militaire en politieke gevolgen van technologie. Een goede maatschappij plaatst volgens Wiener rechtvaardigheid, vrijheid, gelijkheid en welwillendheid (‘benevolence’) centraal.[43]

D. ORGANISATIE EN BELEID
Succesvol AI-ecosysteem en verdeling van rollen (bedrijfsleven, overheid, kennisinstellingen, andere organisaties)

Ik koos voor vier specifieke en concrete aanbevelingen (in een willekeurige volgorde); uiteraard is deze lijst niet exhaustief.

1) Regelgeving: Maak van ethiek een standaardonderdeel in technologisch aspectenonderzoek (TA)
Doorgaans ligt de klemtoon van 'Risk & Technology Assessments' op de *harde* impact van technologie, zoals schade voor het milieu, gezondheid en veiligheid.[44] Harde impact wordt gezien als objectief, feitelijk en neutraal. Het causale verband tussen de technologie en de harde impact is duidelijk en kan bovendien getest (gemeten) worden. Hierbij wordt geen expliciet gebruik gemaakt van ethische criteria en implicaties: die vallen onder de *zachte* impact van technologie.[45] De zachte impact wordt nog te vaak afgedaan als subjectief en daardoor 'privé'.[46]

'Ethical risk analysis' (ethische risicoanalyse) is pas vrij recent ontstaan en focust op verantwoordelijkheid, rechtvaardigheid, autonomie, welzijn en dergelijke meer in relatie tot risico's.[47] Vaak gaat het om een kwantitatieve benadering zoals een kostenbatenanalyse. Zeker voor wat zogenaamde NEST ('New and Emerging Science and Technology') betreft (denk bv. aan zorgrobots of nanotechnologie), is het essentieel om mogelijke ethische gevolgen in kaart te brengen, vanaf Onderzoek & Ontwikkeling en de introductiefase, door te anticiperen op toekomstig gebruik, toepassingen en sociale en ethische gevolgen.[48] Ethische analyse moet een standaarddeel van TA worden, ook om het overzicht te bewaren waar individuele techniekontwikkelaars en onderzoekers niet altijd zicht op hebben. 'Ethical Technology Assessment'[49] (eTA), 'techno-ethical scenarios approach'[50] en 'Anticipatory Technology Ethics'[51] (ATE) leveren alvast interessante methodes hiervoor.

eTA focust bijvoorbeeld op beleid en ontwikkelaars. Het doel ervan is dat ethici gedurende het hele ontwikkelingsproces in nauw contact staan met de ontwikkelaars, om mogelijke benaderingen (of oplossingen) met hen te bespreken bij problemen die gaandeweg ontstaan. Het is dus niet één assessment, maar herhaaldelijke assessments (iteratief), om het ontwerp te kunnen bijsturen en evalueren. eTA gebeurt aan de hand van een lijst bestaande uit 9 concepten: (1) Dissemination and use of information, (2) Control, influence and power, (3) Impact on social contact patterns, (4) Privacy, (5) Sustainability, (6) Human reproduction, (7) Gender, minorities and justice, (8) International relations, and (9) Impact on human values. Het doel ervan is om in een vroeg stadium een indicatie te krijgen van negatieve ethische implicaties. Een kritiek erop is dat de lijst onvoldoende gedetailleerd is.[52] Zelf zou ik onder meer 'kinderen' toevoegen aan de conceptenlijst. Ethiek is niet binair; het doel van de ethische TA moet daarom liggen op het stellen van de *juiste* vragen[53]. Het mag geen afgesloten ja/nee-systeem worden.

Ook bij ethische TA is er, uiteraard, het probleem van onzekerheid: NEST is nog niet breed in gebruik, dus we weten niet hoe ze ingebed zullen worden en wat de gevolgen zullen of kunnen zijn. Net daarom is het belangrijk om financieel te blijven investeren in TA, ook om *nieuwe* theorieën, methodes, (efficiënte) testen, standaarden en procedures te ontwikkelen.

2) Regelgeving: Ontwerpers van algoritmen aansprakelijk maken
De ingenieurs die de algoritmen trainen, hebben de kennis en dus de sleutel in handen om te weten waar het kan mislopen. Een lijst zoals FAT/ML (cf. supra) mag geen vrijblijvende checklist worden en om die reden is het raadzaam om de aansprakelijkheid bij hen te leggen. Bij individuele verantwoordelijkheid is elk lid van de organisatie verantwoordelijk voor zijn of haar bijdrage. Het voordeel van dit model is dat het in principe iedereen motiveert en dat het moreel rechtvaardig is. Bij hiërarchische verantwoordelijkheid, bijvoorbeeld, is enkel het management verantwoordelijk. Een individueel verantwoordelijkheidsmodel anticipeert op problemen inzake moral disengagement (cf. supra).

De Duitse code (cf. bijlagen) adviseert om relevante scenario's over hoé het zelflerende systeem precies leert en hoé dit de veiligheid verhoogt over te dragen aan een centrale catalogus bij een neutrale instantie, om te komen tot universele standaarden. Dit is een interessante manier om kennis te bundelen, in dit geval inzake internationale verkeersveiligheid, waarvan onderzocht moet worden of die ook in andere domeinen kan toegepast worden.

Tony Fadell (Apple; Nest) pleit, net als James Williams (voormalig Google-medewerker), voor een eed van Hippocrates voor ingenieurs en ontwerpers, analoog aan die voor artsen, die hen ethisch ontwerp oplegt.[54] Zo een eed maakt hen bewuster van hun verantwoordelijkheid en de impact van bepaalde keuzes. Maar een gedragscode gebaseerd op vrijblijvende principes is in veel gevallen te zwak. Er zijn ook al veel internationale richtlijnen en ethische codes, zoals de ACM Code of Ethics.

3) Langetermijnvisies: sociaalondernemerschap en andere verdienmodellen

Het verdienmodel van veel (gratis) platformen is gebaseerd op de verkoop van persoonlijke gegevens. Er wordt gretig gebruik gemaakt van onze mentale zwaktes en basisinstincten. Gekende trucjes zijn: meldingen en updates in het rood aangeven, eindeloos kunnen scrollen door tijdlijnen, filmpjes op basis van persoonlijke interesses aanraden op YouTube ('recommending videos'), autoplay, enzovoort. James Williams noemt dit ontwerp 'distraction by design'.[55] Het doel ervan is de aandacht van de gebruiker te kapen. Je krijgt gratis gebruik in ruil voor aandacht en tijd. Achter de big data gaat een marktmodel van *dataminers* en *databrokers* schuil, mensen die data verzamelen en er vervolgens profielen en patronen in zoeken om doelgericht en op maat van het individu te adverteren.

James Williams, voormalig werknemer bij Google, verliet het bedrijf omdat zijn geweten hem parten begon te spelen. Op de werkvloer verschoof volgens hem de focus van het ontwerpen van producten naar het ontwerpen van *gebruikers* ('designing users').[56] Miljarden dollars worden uitgegeven om uit te vissen hoe je hen langer naar iets kunt laten kijken of kunt verleiden tot impulsaankopen.[57] Onlinegedrag wordt gemonitord en diepgaand geanalyseerd:[58] hoeveel 'page views'? Welke intentie zit er achter zoekopdrachten en surfgedrag? Wat zijn de fysieke locaties en interesses van gebruikers? Deep learning-technieken worden hiervoor ingezet en om die reden vormt 'distraction by design' ook een probleem voor toekomstige AI-systemen.

Zelfcontrole moet terug *beloond* worden in plaats van het te ontmoedigen. Ongeduld, luiheid en gemakzucht worden door dit marktmodel op een voetstuk geplaatst. En dat staat haaks op het goede leven en op gedrag dat we onze kinderen graag aanleren, zoals zelfbeheersing. 'Distraction by design' bevordert het aanscherpen van dergelijke cognitieve vaardigheden natuurlijk niet. Gratis gebruik van de diensten en afhankelijkheid van adverteerders verhoogt de focus op 'distraction by design' alleen maar; het is belangrijk om ook naar die context te kijken. Huidige en opkomende AI-systemen moeten minder afhankelijk van adverteerders zijn.

Een overheid kan onder meer sociaal ondernemerschap stimuleren, waar ook langetermijndoelen een rol spelen en verdergegaan wordt dan 'alles voor de groei' en 'profit before principles'.

4) Onderwijs

Het is essentieel om ingenieurs, programmeurs, techniekontwikkelaars, … de tijd te geven om na te denken over mogelijk gebruik, misbruik en ethische gevolgen. Om hen die denkmethodes eigen te maken, moet techniekethiek een standaardvak zijn voor elke toekomstige ingenieur, ontwerper en computerwetenschapper. De Europese waarden (cf. deel C.) gaan terug op de Verlichting; Verlichtingsfilosoof Immanuel Kant hamerde op het belang van zélf denken.[59] Mensen opleiden tot moreel autonome en kritisch denkende burgers is een kerntaak van de universiteit. Er zijn minstens drie redenen waarom filosofie (en ethiek als onderdeel ervan) belangrijk is voor technologie en voor techniekontwikkelaars: *analytisch* redeneren om tot goede begrippenkaders te komen, *kritisch* denken, om te leren naar de 'achterkant' van dingen te kijken en autonoom te reflecteren en beslissingen te nemen en *richtinggevend* denken, zoals over ethische vraagstukken.[60] De drie redenen zijn nauw met elkaar verbonden.[61]

Ingenieurs, computerwetenschappers en andere ontwikkelaars in spe kunnen zich zo rustig de ethische vragen eigen maken en technieken aanleren om hierop antwoorden te bieden, zoals waardengedreven ontwerp (Value Sensitive Design). Zo kunnen ethische vragen een vanzelfsprekendheid en een vast onderdeel van ontwerp worden. Hoe beter de makers zélf de ethische vraagstukken en maatschappelijke processen begrijpen, zoals de impact van hun ontwerp en het feit dat dat niet zomaar neutraal is, hoe beter de beslissingen zijn die ze kunnen maken.

Motivatie van ethiekonderwijs voor toekomstige techniekontwikkelaars:[62]
- Vaardigheden ontwikkelen om sociale en ethische problemen in ('bij', 'van') ontwerp te herkennen
- Analytisch kijken naar eigen ontwerp, zoals de intentie achter het ontwerp, de eindgebruikers, …
- Een 'toolbox' aanreiken om morele problemen analytisch te ontleden, zoals op het vlak van feiten, stakeholders, waarden, gevolgen, …
- Morele creativiteit: nadenken over oplossingen en mogelijkheden, zeker bij conflicterende waarden (bv. tussen privacy en publieke veiligheid)
- Trainen in morele oordeel- en besluitvorming en kritisch leren kijken naar eigen argumenten (morele autonomie)
- Gevoelig maken voor eigen blinde vlekken (bv. moral disengagement)
- Bewustwording van ethische problemen
- Langetermijnvisies (<> kortetermijndoel)

LEESSUGGESTIES (boeken)

Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence*. Dordrecht: Springer:

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

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

Brey, P., Briggle, A., & Spence, E. (2012). *The Good Life in a Technological Age* (Routledge Studies in Science, Technology, and Society). New York: Routledge.

Brooks, R. A. (1999). *Cambrian Intelligence. The Early History of the New AI*. Cambridge, MA: MIT Press.

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

Brynjolfsson, E., & McAfee, A. (2014/2016). *The Second Machine Age. Work, Progress, and Prosperity in a Time of Brilliant Technologies.* New York / Londen: W. W. Norton & Company.

Gabriels, K. (2016). *Onlife. Hoe de digitale wereld je leven bepaalt*. Tielt: Lannoo.

Gabriels, K. (2019, forthcoming). *Regels voor robots. Ethiek in tijden van A.I.* Brussel: VUBPRESS (verschijnt dit najaar).

Hanks, C. (ed.) (2016). *Technology and Values: Essential Readings*. Chichester, West Sussex: Wiley-Blackwell.

Isaacson, W. (2014). *The Innovators. How a Group of Hackers, Geniuses and Geeks Created the Digital Revolution*. Londen: Simon & Schuster. (ook beschikbaar in het Nederlands, *De uitvinders: Hoe een groep hackers, genieën en nerds de digitale revolutie ontketende*)

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

Lin, P., Abney, K., & Bekey, G. A. (eds.) (2012/2014). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that Will Transform How We Live, Work and Think.* Londen: John Murray Publishers,

O'Neil, C. (2016/2017). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books.

Pitt, J. C., & Shew, A. (eds.) (2018). *Spaces for the Future: A Companion to Philosophy of Technology*. New York / Londen: Routledge.

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

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

Thompson, C. (2019). *De coders. Een kijkje in het hoofd van programmeurs – de machtigste beroepsgroep ter wereld*. Amsterdam: Maven Publishing.

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

Verkerk, M. J., Hoogland, J., van der Stoep, J., & de Vries, M. J. (2007). *Denken, ontwerpen, maken. Basisboek techniekfilosofie*. Amsterdam: Boom.

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

Williams, J. (2018). *Stand Out of Our Light. Freedom and Resistance in the Attention Economy*. Cambridge: Cambridge University Press.

Zuboff, S. (2019). *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*. Londen: Profile Books.

ETHISCHE CODE (DUITSLAND) ZELFRIJDEND TRANSPORT[63]

Om het maatschappelijk bewustzijn over zelfrijdend transport te verhogen, ontwikkelde het Massachusetts Institute of Technology (MIT) een website waarop morele dilemma's met een zelfrijdende wagen visueel worden voorgesteld.[64] Je kunt diverse scenario's beoordelen en er ook zelf ontwerpen. Echte situaties op de weg zijn nooit zo simpel en duidelijk als de dertien scenario's die je voorgeschoteld krijgt. Een crash is ook afhankelijk van meerdere factoren, waarin ook een glad wegdek of een lekke band een rol kunnen spelen. Bovendien zijn algoritmen van zelfrijdende wagens nog lang niet zo verfijnd als de scenario's doen geloven; de huidige technologie kan dit nog helemaal niet.

De resultaten van het onderzoek over de MIT 'Moral Machine' legden culturele verschillen bloot, onder meer tussen westerse en oosterse culturen.[65] Zo zijn Japanners meer geneigd om de levens van oudere mensen te sparen dan inwoners van West-Europese landen, die de voorkeur geven aan jongere mensen. Het onderzoek geeft inzage in het morele denken van verschillende culturen. De grote steekproef is verbluffend, maar de resultaten blijven algemeen, omdat de drijfveren achter de afwegingen die mensen maken, terwijl ze de scenario's evalueren, niet onderzocht werden. De contextafhankelijkheid van morele keuzes is ook hier cruciaal: in theorie vind je misschien dat de jongere persoon eerder gered moet worden dan de oudere, maar wat als die laatste je vader of moeder is? Een voorkeur die verschillende culturen deelden, was het sparen van mensen die de wet respecteren in tegenstelling tot voetgangers die oversteken terwijl het verkeerslicht op rood staat. Maar wat als het jouw kind is dat door het rood loopt? Een andere cultuur overschrijdende voorkeur, is dat het leven van een mens primeert op dat van een dier.

Duitsland heeft een code opgesteld, bestaande uit twintig ethische regels, voor zowel gedeeltelijk als volautomatisch zelfrijdend transport. De code verwerpt de keuzes van scenario's van de MIT 'Moral Machine': de machine mag beslissingen nooit baseren op geslacht, leeftijd, enzovoort. Eén regel keert wel terug: het leven van mensen primeert op dat van dieren. De Duitse filosoof Christoph Lütge maakte deel uit van de commissie die de code opstelde. Hij publiceerde die ethische code in 2017, met extra toelichting erbij.[66] De code maakt duidelijk dat je beslissingen niet zomaar mag overlaten aan programmeurs en de vier eerder vermelde niveaus keren er duidelijk in terug: de maatschappij (waaronder beleid), de technologie zelf, de maker en de gebruiker.

De eerste *algemene* regel van de code stelt dat zelfrijdend transport de veiligheid voor al de weggebruikers moet verbeteren. Regels 2, 3 en 4 vallen vervolgens onder 'algemene ethische voordelen'. Zelfrijdende voertuigen moeten mensen beschermen en schade beperken (regel 2). Ze mogen énkel toegelaten worden als onomstotelijk is aangetoond dat ze minder risicovol zijn dan menselijke bestuurders. De commissie concludeerde dat zelfrijdende wagens ethische voordelen hebben, wat een belangrijk argument is om ze te ontwikkelen. Zo zijn zelfrijdende wagens bijvoorbeeld gunstig voor mindervaliden. De publieke sector is verantwoordelijk voor de officiële licenties en monitoring van automatische systemen (regel 3). Automatisch rijden moet aan de nodige voorwaarden en regels voldoen om ongelukken te vermijden en die kan je niet overlaten aan autofabrikanten. Het doel van overheidsregulering is het beschermen van individuen en hun recht op persoonlijke en vrije ontwikkeling (regel 4). Er moet een balans gezocht worden tussen maximale persoonlijke keuzevrijheid in een vrije, democratische samenleving versus de vrijheid en veiligheid van anderen. Dit is een variant op het schadebeginsel van de liberale denker John Stuart Mill: jouw vrijheid stopt zodra iemand anders er schade van ondervindt.

Regels 5-9 gaan over 'onvermijdbare dilemma's' zoals het trolleydilemma.[67] Ongelukken moeten vermeden worden (regel 5): de technologie moet zo ontwikkeld worden dat kritische situaties in de eerste plaats niet ontstaan, zoals de keuze tussen twee ongewenste uitkomsten, waarbij geen simpele afweging gemaakt kan worden. Een heel spectrum aan technologische oplossingen moet aangewend worden en voortdurend worden bijgestuurd en verbeterd, zoals afremmen, sensoren, signalering voor personen die risico lopen (zoals voetgangers) en intelligente weginfrastructuur die signalen doorstuurt naar de zelfrijdende wagen. Enerzijds is er de ethische plicht om volautomatische zelfrijdende voertuigen te

ontwikkelen als ze zoveel veiliger zijn dan menselijke bestuurders, maar anderzijds is het ethisch problematisch als je mensen daardoor onderwerpt aan een 'technologische imperatief' door van hem of haar een schakeltje in een netwerk te maken (regel 6). Dit is een nogal doorwrochte regel die, refererend aan Immanuel Kant, stelt dat je altijd moet vermijden dat de mens een middel, in plaats van een doel op zichzelf, wordt. De commissie was hierover verdeeld, maar eigenlijk is de regel een soort van waarschuwing dat je nooit zomaar iets mag ontwikkelen zonder verdere reflectie. Als een ongeval onvermijdelijk is, ondanks al de technische voorzorgsmaatregelen, dan is de bescherming van een mensenleven topprioriteit (regel 7). Zoals eerder vermeld, is er consensus dat een mensenleven steeds primeert op dat van dieren.

Echte dilemma's, zoals kiezen tussen twee mensenlevens, zijn afhankelijk van de werkelijke, specifieke situatie, en het is onmogelijk om ze te standaardiseren of programmeren op een ethisch onproblematische wijze (regel 8). Systemen moeten ontworpen worden om ongevallen te vermijden, maar een standaardisering ervan is onmogelijk. Het gaat immers om een uiterst complexe en vaak intuïtieve inschatting van specifieke omstandigheden die niet hapklaar te vertalen valt in een abstracte code. Een intuïtieve, persoonlijke keuze mag daarom niet aan een programmeur overgelaten worden: het is verboden om die erin te programmeren.

Als een ongeluk onvermijdelijk is, dan is het strikt verboden om een verschil te programmeren op basis van persoonlijke kenmerken zoals leeftijd, geslacht, fysieke of mentale eigenschappen (regel 9). Het voertuig zo programmeren dat het aantal fysieke letsels beperkt wordt, is wel te rechtvaardigen. Het is verder ook verboden om partijen bij het ongeluk te betrekken die er in eerste instantie niet bij betrokken waren. Deze regel was controversieel; de discussie ging niet over de persoonlijke eigenschappen - daar waren ze het over eens - maar over het al dan niet toelaten dat een programmeur een code schrijft die het aantal "personal injuries" (persoonlijke letsels) beperkt, op welke wijze dan ook. In de luchtvaart is het in principe toegelaten om een gekaapt vliegtuig neer te halen; in dit geval worden individuen opgeofferd om de levens van anderen te sparen. Maar die individuen zijn op voorhand gekend. Een programma is anoniem en kent op voorhand geen individuele slachtoffers; de regel is te abstract en de precieze gevolgen vallen niet te voorzien.

Regels 10 en 11 hebben betrekking op de vraag 'wie is verantwoordelijk?'; het gaat hier over legale verantwoordelijkheid. Deze regels zijn voor de commissie belangrijker dan regels 5-9, omdat ze grote praktische gevolgen hebben. De verantwoordelijkheid schuift mee op naar de fabrikanten en operatoren van de technologische systemen en naar beleids- en wettelijke organen (regel 10). Er is dus een belangrijke transitie, want de huidige focus[68] ligt op de verantwoordelijkheid van de eigenaar van de wagen. Die transitie heeft een enorme impact op verzekeringsaansprakelijkheid.

De aansprakelijkheid voor schade aangericht door de zelfrijdende systemen moet gereguleerd worden zoals bij andere vormen van aansprakelijkheid bij producten ('product liability') (regel 11). Daaruit volgt dat autofabrikanten en operatoren verplicht zijn om hun systemen voortdurend te optimaliseren en om systemen die al in gebruik genomen zijn te observeren en verbeteren waar nodig. Regel 12 schrijft voor dat het publiek geïnformeerd moet worden over de nieuwe technologieën en hun toepassingen. De regels moeten transparant zijn en beoordeeld worden door onafhankelijke instanties, zoals een ngo die consumentenbelangen verdedigt en om die reden bedrijven kritisch monitort.

Regels 13 en 14 gaan over 'veiligheid'. Bij vliegtuigen en spoorwegen is er een centrale controle; het is nog onduidelijk of er op termijn ook een gecentraliseerde controle komt voor zelfrijdend transport (regel 13). Volledige connectiviteit en centrale controle van *alle* gemotoriseerde voertuigen bij digitale transportinfrastructuur is ethisch problematisch als daar een totale surveillance van weggebruikers mee gepaard gaat. Automatisch rijden is enkel te rechtvaardigen zolang mogelijke (cyber)aanvallen of interne zwaktes van het systeem niet tot zo grote schade leiden dat die het vertrouwen in transport beschadigt (regel 14). Dit is een reële discussie, ook over mogelijke hacking van zulke wagens die vervolgens als wapens ingezet kunnen worden, zoals bij de aanvallen van 11 september 2001 ook vliegtuigen als wapens werden gebruikt.

Regel 15 gaat over 'databescherming: de eigenaar en gebruikers van het voertuig beslissen over de data die het voertuig genereert. Enkel na toestemming, op vrijwillige basis,

kunnen data gedeeld worden met bedrijven, maar er moeten ook voldoende alternatieven zijn, zodat het geen valse vrije keuze is. De richtlijn schrijft 'privacy by design' voor (computertechnologie die structureel inzet op onder meer beveiliging en versleuteling) en fabrikanten mogen argumenten zoals 'verbeterd comfort' niet inzetten om gebrek aan privacy of data-eigendom bij gebruikers te rechtvaardigen.

Regels 16 en 17 focussen op de 'mens-machine interface': Het verschil tussen wanneer het zelfrijdende systeem rijdt en wanneer een chauffeur overneemt moet *altijd* duidelijk zijn (regel 16). In het geval van het laatste moet de mens-machine interface zo ontworpen zijn dat het te allen tijde duidelijk is waar (bij wie) de individuele verantwoordelijkheid ligt, voornamelijk op het vlak van controle ('wie bestuurt de wagen?'). Tijdstippen moeten gelogd worden, ook omwille van *internationale* procedures (internationaal wegverkeer), wanneer de mens de controle overhandigt aan de machine. Die procedures moeten eenduidig en helder zijn, zodat zonder discussie duidelijk is wie de controle heeft: mens of machine. De mens moet altijd de optie krijgen om het systeem over te nemen en zelf te rijden, ook al leidt dit mogelijk tot extra risico's. Het systeem (zoals software) moet op maat van de mens ontworpen worden en niet omgekeerd: het moet zo min mogelijk aanpassing van de mens vragen (regel 17). De controle moet dus makkelijk over te nemen zijn door mensen.

Regels 18 en 19 beslaan 'leersystemen': Zelflerende systemen zoals 'machine learning' (de algoritmen leren zelf uit al de data die ze via sensoren binnenkrijgen, zoals patroonherkenning) zijn ethisch toegelaten als ze bijdragen tot verhoogde veiligheid (regel 18). De code adviseert om relevante scenario's over hoé het systeem precies leert en hoé dit de veiligheid verhoogt over te dragen aan een centrale catalogus bij een neutrale instantie, om te komen tot universele standaarden. In urgente situaties moet het voertuig automatisch, dus zonder menselijke begeleiding, schakelen naar een 'safe condition', een veiligheidsmodus (regel 19). Het probleem is dat er voorlopig geen consensus over het concept bestaat, dus die moet zo snel mogelijk bereikt worden. Betekent dit dat het voertuig in het midden van de baan stopt of dat het veilig langs de kant van de weg parkeert? Lütge geeft logischerwijze aan dat het laatste zinvoller is.

Regel 20, ten slotte, gaat over de opleiding van de chauffeur: het goede en juiste gebruik van de automatische systemen is onderdeel van een algemene digitale opleiding. Uiteraard moet een rijopleiding deze zaken aanleren, met diverse proefritten en testen. Op dit vlak is de code nog vrij algemeen, maar de commissieleden geven aan dat dit later gespecifieerd moet worden.

Hoewel deze code enkel in Duitsland geldt, geeft ze goed weer hoe ethische vraagstukken over machines in de praktijk kunnen leiden tot regels. Hoewel er veel discussie in de commissie was, kwamen ze uiteindelijk tot een consensus op alle punten. Een moreel handelende machine die op een intuïtieve, contextafhankelijke manier moreel kan redeneren zoals de mens is nog verre, verre toekomstmuziek en misschien wordt die wel nooit ontwikkeld. Uiteraard zijn er voor de totstandkoming van zo'n machine duidelijke regels nodig, maar die focussen meer op algemene principes zoals rechtvaardigheid, transparantie, privacy by design en het vermijden van vooringenomen datasets.

EINDNOTEN

[1] Brooks, R. A. (1999). *Cambrian Intelligence. The Early History of the New AI*. Cambridge, MA: MIT Press, p. 80.
[2] Tegmark, M. (2017). *Life 3.0. Mens zijn in het tijdperk van kunstmatige intelligentie*. Amsterdam: Maven Publishing, p. 80.
[3] Jacobs, A., Tytgat, L., Maus, M., Meeusen, M. & Vanderborght, B. (red.) (2019). *Homo Roboticus. 30 vragen en antwoorden over mens, robot & artificiële intelligentie*. Brussel: VUBPRESS, p. 14.
[4] Lin, P., Abney, K., & Bekey, G. A. (eds.) (2014). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press, p. 4.
[5] Russell, S. J., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach (Third Edition)*. Essex: Pearson Education Limited, p. 29.
[6] Technologie vormt mee de context waarin we ons gedragen en keuzes maken: de omgeving stuurt ons. Nudging - mensen subtiel een duwtje in de juiste richting geven – speelt daarop in. Het bekendste voorbeeld ervan is de nepvlieg in het mannenurinoir, die bedacht werd op de luchthaven van Schiphol. Nudging helpt mensen bij het nemen van een keuze, weliswaar in de 'goede' richting. Thaler en Sunstein noemen het 'libertair paternalisme'. Paternalistisch, omdat je het vanuit een zorgende rol doet: je wilt gedrag beïnvloeden op een gunstige manier, zodat mensen beter en gezonder gaan leven. Libertair omdat er geen sprake is van dwang: mensen hebben de vrije keuze om zich anders te gedragen. Er is geen keuze die geblokkeerd wordt.
Op onlineplatformen worden we doelgericht benaderd met gepersonaliseerde advertenties. Op basis van onder meer onze 'vind-ik-leuks' (*likes*) en cookies worden onze gegevens in profielen gegoten. Onderzoekster Karen Yeung noemt het nudgen van mensen in een digitale omgeving 'hypernudging'. Hypernudges komen tot stand door verregaande personalisering. Nudges zijn meestal statisch (en gelijk) voor wie ermee in aanraking komt, maar hypernudges zijn geïndividualiseerd, op maat gemaakt en bovendien onderhevig aan verandering door real time (instant) feedback, om het systeem te verfijnen. Ons gedrag kan erdoor beïnvloed worden, maar de werkwijzen erachter zijn niet transparant. Een ander voorbeeld van onwenselijke nudging online is wanneer onze toestemming gevraag wordt: er verschijnt een grote 'YES' in kleur en dan in kleine letters daaronder of ernaast 'no' tegen de witte achtergrond. Een hele krachtige nudge, ook besproken door Thaler & Sunstein, is de standaardinstelling (default setting): die speelt in op de gemakzucht van mensen en hun voorkeur voor de 'weg van de minste weerstand'. Dat is ook de reden waarom digitale platformen 'public by default' verkiezen, omdat data dan onmiddellijk beschikbaar zijn. De Europese Commissie had eerder kunnen ingrijpen inzake privacy by default (nu opgenomen in de GDPR/AVG).
Thaler, R. H., & Sunstein, C. R. (2016). *Nudge. Naar betere beslissingen over gezondheid, geluk en welvaart*. Amsterdam / Antwerpen: Uitgeverij Business Contact. Oorspronkelijke titel (2008): *Nudge. Improving Decisions about Health, Wealth and Happiness.*
Yeung, K. (2016). 'Hypernudge': Big Data as a Mode of Regulation by Design. *Information, Communication & Society 20* (1), pp. 118-136.
[7] Zie ook Russell & Norvig, 2016, p. 1020.
[8] Dit raakt ook aan debatten over 'superintelligentie' en 'singularity', zie:
Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
Kurzweil, R. (2005). *The Singularity is Near. When Humans Transcend Biology*. New York: Penguin Books.
[9] Tegmark, 2017, p. 80.
[10] Geïnspireerd door het zachte weefsel waaruit wij en vele andere organismen bestaan, worden 'soft robots' gemaakt uit flexibele materialen. Door hun flexibiliteit kunnen soft robots ingezet worden voor talloze toepassingen. Zo worden ze gebruikt om delicate en malse voorwerpen te grijpen in de voedselindustrie of in minimaal-invasieve chirurgie. De zachte materialen maken hen echter gevoelig voor schade door scherpe voorwerpen of overmatige druk. Eens beschadigd, moeten componenten vervangen worden of belandt de robot bij het vuil. Als wij een wonde hebben, dan herstelt die vanzelf. Robotonderzoekers van de Vrije Universiteit Brussel maken robotonderdelen van rubberachtige polymeren die zelfherstellende eigenschappen hebben: als die beschadigd worden, dan hecht de 'wonde' vanzelf terug aan elkaar. Dit materiaal is bovendien duurzaam, omdat het gerecycleerd kan worden.
Maar bij robots is morele paniek nooit ver weg. Toen de Britse tabloid *The Sun* schreef over dit onderzoek, wekten ze de indruk dat als mensen in een gewapende strijd robots neerschieten, deze terminators zichzelf zullen herstellen en dus onklopbaar zijn. Met andere woorden: we are doomed. Media praten ons angsten aan: wanneer komen robots ons vermoorden?
Zie Vanderborght, B. (2018). Worden robots binnenkort onverwoestbaar? *Knack.be* (21 maart 2018). https://www.knack.be/nieuws/wetenschap/worden-robots-binnenkort-onverwoestbaar/article-opinion-981963.html

Murphy, M. (2017). Terminator-Style Robot with Self-Healing 'Flesh'. *The Sun* (17 augustus 2017). https://www.thesun.co.uk/tech/4262775/scientists-create-terminator-style-immortal-robot-with-self-healing-flesh/

[11] Boden, M. A. (2016). *AI. Its Nature and Future*. Oxford: Oxford University Press, p. 57.

[12] Zie o.a. het onderzoek van Agnieszka Landowska, bv. Landowska, A. (2019, forthcoming). Uncertainty in Emotion Recognition. *Journal of Information, Communication and Ethics in Society*.

[13] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature 542*, pp. 115-118. DOI: 10.1038/nature21056.

[14] Voor een toegankelijk artikel over mogelijkheden, zie bijvoorbeeld De Cleene, D. (2018). Big data zijn een big deal in de geneeskunde. Worden die beloften ingelost? *EOS Wetenschap* (20 september 2018). https://www.eoswetenschap.eu/gezondheid/big-data-zijn-een-big-deal-de-genees-kunde-worden-die-beloften-ingelost

[15] Voor een studie over self-tracking technologieën in de professionele geneeskunde, zie Gabriels, K., & Moerenhout, T. (2018). Exploring Entertainment Medicine and Professionalization of Self-Care: Interview Study Among Doctors on the Potential Effects of Digital Self-Tracking. *Journal of Medical Internet Research 20* (1), e433. DOI: 10.2196/jmir.8040

[16] Voor mogelijkheden en valkuilen met betrekking tot digitalisering en IoT in de geneeskunde en zorgsector, zie het zesde hoofdstuk in Gabriels, K. (2016). *Onlife. Hoe de digitale wereld je leven bepaalt*. Tielt: Lannoo.

[17] Zie bijvoorbeeld het bovenvermelde onderzoek van A. Landowska.

[18] Een toegankelijk (en niet-academisch) boek over onbetrouwbare computermodellen is O'Neil, C. (2017). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books.

[19] Vaak worden ethiek en moraliteit als synoniemen beschouwd, maar dat klopt eigenlijk niet. Ethiek verwijst naar een specifieke filosofische discipline die een systematische (bv. aan de hand van ethische theorieën) en kritische reflectie op moraliteit beoogt. Moraliteit is de individuele reflectie van normen, gewoontes en gebruiken binnen een specifieke culturele context, zodat we ook spreken over de 'moraal' of zeden van een bevolking of cultuur. Moraliteit hoort bij mensen zoals bijvoorbeeld taal. Mensen zijn 'morele wezens' die in staat zijn tot morele reflectie over hun eigen gedrag, principes, oordelen, normen en waarden. Mensen zijn ook verantwoordelijk voor hun gedrag, zolang ze handelingsbekwaam zijn en uit vrije wil handelen.

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

[21] Een van de bekendste voorbeelden hiervan is de serie experimenten van sociaalpsycholoog Stanley Milgram in de jaren 1960, zie Milgram, S. (1975). *Obedience to Authority: An Experimental View*. New York: Harper & Row.

[22] Johnson, D. G., & Verdicchio, M. (2017). Reframing AI Discourse. *Minds & Machines 27*, pp. 575-590. Dit is een interessant artikel dat ook ingaat op het publieke discours over AI en hoe dit mensen misleidt over 'autonome' machines.
Engelstalige definitie van sociotechnical blindness die Johnson & Verdicchio geven: "blindness to all of the human actors involved and all of the decisions necessary to make AI systems, allows AI researchers to believe that AI systems got to be the way they are without human intervention."

[23] Ik gebruik deze termen door elkaar, maar daarbij wordt telkens naar de volledige groep verwezen.

[24] Uiteraard gaat de problematiek verder dan het ontwerp van een product: zo heeft de omgeving waarin je woont ook gevolgen voor je gezondheid. Je zal minder wandelen als je in een omgeving woont zonder fatsoenlijke voetpaden of als de buurt onveilig is. Wie dichtbij zogenaamde 'takeaways' woont en werkt, heeft dubbel zoveel kans op obesitas. Zie Burgoine, T., Forouhi, N. G., Griffin, S. J., Wareham, N. J., & Monsivai, P. (2014). Associations between Exposure to Takeaway Food Outlets, Takeaway Food Consumption, and Body Weight in Cambridgeshire, UK: Population Based, Cross Sectional Study. *British Medical Journal 348*, g1464.

[25] Zie Johnson & Verdicchio, 2017, p. 576.

[26] Oorspronkelijk citaat: "The less intervention needed by humans in its operation and the wider its scope of action, the more autonomous the artefact". Zie Johnson & Verdicchio, 2017, p. 580.

[27] Met dank aan Dr. Pieter Buteneers, gespecialiseerd in machine learning en huidig CTO van Chatlayer, voor zijn suggesties.

[28] Zie bijvoorbeeld Price, R. (2016). Microsoft Is Deleting Its AI Chatbot's Incredibly Racist Tweets, *Business Insider UK*, 24 maart 2016. http://uk.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3?r=US&IR=T

[29] Lee, P. (2016). Learning from Tay's introduction. *Official Microsoft Blog* (25 maart 2016). https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

[30] Zie Stuart-Ulin, R. C. (2018).Microsoft's Politically Correct Chatbot is even worse than its racist one. *Quartz* (31 juli 2018). https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racist-one/
[31] Johnson & Verdicchio, 2017, p. 589.
[32] Zie Wiener, N. (1950/1954). *The Human Use of Human Beings*: *Cybernetics and Society*. Cambridge: The Riverside Press.
Ook computerwetenschapper Joseph Weizenbaum waarschuwde hiervoor, zie Weizenbaum, J. (1976). *Computer Power and Human Reason. From Judgment to Calculation*. New York / San Francisco: W. H. Freeman and Company.
[33] Zie ook de 'Campaign to Stop Killer Robots', https://www.stopkillerrobots.org/
[34] Zierse, M. (2017). Fabrikant van slimme stofzuigers wil de plattegrond van uw huis doorverkopen. *Trouw* (25 juli 2017). https://www.trouw.nl/home/fabrikant-van-slimme-stofzuigers-wil-de-plattegrond-van-uw-huis-doorverkopen~acd9e66d/
[35] Van Hee, L., Verheyden, T., Van Den Heuvel, R., & Baert, D. (2019). Google-medewerkers luisteren mee naar uw gesprekken, ook in uw huiskamer. *VRTNWS* (10 juli 2019). https://www.vrt.be/vrtnws/nl/2019/07/10/google-luistert-mee/
[36] Hern, A. (2019). Apple Apologises for Allowing Workers to Listen to Siri Recordings. *The Guardian* (29 augustus 2019). https://www.theguardian.com/technology/2019/aug/29/apple-apologises-listen-siri-recordings
[37] Baert, D. (2019). De limme speaker luistert, duizenden Amazon-werknemers luisteren stiekem mee. *VRTNWS* (11 april 2019). https://www.vrt.be/vrtnws/nl/2019/04/11/de-slimme-speaker-luistert-duizenden-amazon-werknemers-luistere/
[38] van den Broeck, A., de Witte, H., Vansteenkiste, M., Lens, W., & Andriessen, M. (2009). De Zelf-Determinatie Theorie: kwalitatief goed motiveren op de werkvloer. *Gedrag & Organisatie 22* (4), pp. 316-334, p. 320.
[39] Rocher, L., Hendrickx, J. M., & de Montjoye (2019). Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models. *Nature Communications 10*. https://www.nature.com/articles/s41467-019-10933-3
New Scientist Nederland berichtte ook over deze studie, zie Schenk, D. (2019). Kunstmatige intelligentie vindt je, zelfs als je gegevens anoniem verwerkt zijn. *New Scientist* (29 juli 2019). https://newscientist.nl/nieuws/ai-vindt-je-zelfs-als-je-gegevens-anoniem-verwerkt-zijn/
[40] Voor bijkomende informatie, zie bijvoorbeeld ook:
-Research Priorities for Robust and Beneficial Artificial Intelligence:
Russell, S., Dewey, Daniel & Tegmark, Max (2015) Research Priorities for Robust and Beneficial Artificial Intelligence, *AI Magazine* 36 (4), pp 105-114, Association for the Advancement of Artificial Intelligence. https://futureoflife.org/data/documents/research_priorities.pdf?x40372 -
Onderzoekscentra:
Cambridge: Centre for the Study of Existential Risk (CSER)
Oxford: Future of Humanity Institute (FHI)
USA/Boston: Future of Life Institute (FLI)
Berkeley: Machine Intelligence Research Institute (MIRI)
-Twee Vlaamse rapporten:
https://www.kvab.be/nl/standpunten/artifici%C3%ABle-intelligentie
https://www.vario.be/nl/publicaties/advies-5-vlaamse-beleidsagenda-artifici%C3%ABle-intelligentie
[41] Jong, R. de, Kool, L. & van Est, R. (2019). *Zo brengen we AI in de praktijk vanuit Europese waarden*. Den Haag: Rathenau Instituut, p. 11.
[42] Wiener, N. (1950). *The Human Use of Human Beings: Cybernetics and Society*. Cambridge: The Riverside Press.
[43] Zie ook Bynum, T. W. (2004). Ethical Challenges to Citizens of 'The Automatic Age': Norbert Wiener on the Information Society. *Journal of Information, Communication & Ethics in Society 2*, pp. 65-74.
[44] Swierstra, T., & te Molder, H. (2011). Risk and Soft Impacts. In S. Roeser (ed.) (2011). *Handbook of Risk Theory*. Springer, Dordrecht, pp. 1050-1066, p. 1050.
[45] Brey, P. (2017). Ethics of Emerging Technologies. In S. O. Hansson (ed.) (2017). *Methods for the Ethics of Technology*. Rowman and Littlefield International.
Risicoanalyse kan nog verder opgedeeld worden in 'risk assessment' en 'risk management': het eerste focust op de identificatie, evaluatie en het meten van de waarschijnlijkheid en ernst van de risico's en het tweede op het nemen van beslissingen over risico's, zie ook Brey 2017.
[46] Swierstra & te Molder, 2011, p. 1050.
[47] Brey, 2017. Zie ook Asveld, L., & Roeser, S. (eds.) (2009). *The Ethics of Technological Risk*. Londen: Earthscan Publishers.
[48] Brey, P. A. E. (2012). Anticipatory Ethics for Emerging Technologies. *Nanoethics 6*, pp. 1-13, p. 1.

[49] Palm, E., & Hansson, S. O. (2006). The Case for Ethical Technology Assessment (eTA). *Technological Forecasting & Social Change 73* (5), pp. 543–558.
[50] Boenink, M., Swierstra, T., & Stemerding, D. (2010). Anticipating the Interaction Between Technology and Morality: A Scenario Study of Experimenting with Humans in Bionanotechnology. *Studies in Ethics, Law, and Technology 4* (2).
Swierstra, T., & Rip, A. (2007). Nano-Ethics as NEST-Ethics: Patterns of Moral Argumentation about New and Emerging Science and Technology. *Nanoethics 1*, pp. 3-20.
Swierstra, T. (2016). The Ethics of New and Emerging Science and Technology. An Introduction. In R. Nakatsu, M. Rauterberg, & P. Ciancarini (eds.) (2016). *Handbook of Digital Games and Entertainment Technologies*. Dordrecht: Springer.
[51] Brey, 2012.
[52] Zie Brey, 2012, p. 4.
[53] Zie ook Swierstra, 2016.
[54] Schwab, K. (2017). Nest Founder: I Wake Up in Cold Sweats Thinking, What Did We Bring to the World?. *Fastcompany* (17 juli 2017). https://www.fastcompany.com/90132364/nest-founder-i-wake-up-in-cold-sweats-thinking-what-did-we-bring-to-the-world
Williams, J. (2018). *Stand Out of Our Light. Freedom and Resistance in the Attention Economy*. Cambridge: Cambridge University Press, p. 5.
[55] Williams, J. (2018). *Stand Out of Our Light. Freedom and Resistance in the Attention Economy*. Cambridge: Cambridge University Press, p. 5.
[56] Zie ook Williams, 2018, p. 10.
[57] Williams, 2018, p. 33.
[58] Williams, 2018, p. 31.
[59] De oorspronkelijke tekst van Immanuel Kant zijn 'Beantwortung der Frage: Was ist Aufklärung?' (1784) en de vertalingen ervan zijn online beschikbaar.
[60] Verkerk, M. J., Hoogland, J., van der Stoep, J., & de Vries, M. J. (2007). *Denken, ontwerpen, maken. Basisboek techniekfilosofie*. Amsterdam: Boom, p. 13 en pp. 15-19.
[61] Verkerk et al., 2007, p. 17.
[62] Gebaseerd op van de Poel, I., & Royakkers, L. (2011). *Ethics, Technology, and Engineering: An Introduction*. Chichester, West Sussex: Wiley-Blackwell, p. 2.
[63] Dit komt uit Gabriels, K. (2019, forthcoming). *Regels voor robots. Ethiek in tijden van A.I*. Brussel: VUBPRESS.
[64] Zie http://moralmachine.mit.edu
[65] Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine Experiment. *Nature 563*, pp. 59-64.
[66] Luetge, C. (2017). The German Ethics Code for Automated and Connected Driving. *Philosophy & Technology*. DOI: 10.1007/s13347-017-0284-0.
[67] Het trolleydilemma is een filosofisch gedachte-experiment van de Britse filosofe Philippa Foot. De meest gangbare versie ervan is dat je ziet hoe een treintje (trolley) op hol slaat, richting vijf personen die op de rails liggen en die niet in staat zijn om zelf te ontkomen, omdat ze bijvoorbeeld vastgebonden zijn. Je merkt een hendel op: als je die omgooit, wijkt het op hol geslagen treintje af, in de richting van één persoon die (vastgebonden) op de sporen ligt. Er zijn dus twee mogelijkheden: nietsdoen, waardoor vijf personen sterven, of de hendel omgooien waardoor je één persoon doodt, maar vijf andere levens redt. Studies tonen aan dat het merendeel kiest voor de consequentialistische of utilitaristische overweging: zoveel mogelijk levens redden. Maar als die ene persoon die zou sterven je familielid of je partner is, of veel jonger is dan de vijf personen die gered zouden worden, dan kiezen minder mensen ervoor om die op te offeren, wat een weinig verrassend resultaat is. Door de jaren heen ontstonden diverse varianten op het trolleydilemma. Het gedachte-experiment werd ook gesimuleerd in een virtuele omgeving, waarbij de proefpersonen de gevolgen van hun keuze zagen gebeuren. De resultaten liggen in de lijn van de andere studies: het overgrote merendeel van de proefpersonen (9 op de 10) koos ervoor om zoveel mogelijk levens te sparen.
Bleske-Rechek, A., Nelson, L. A., Baker, J. P., Remiker, M. W., & Brandt, S. J. (2010). Evolution and the Trolley Problem: People Save Five over One unless the One is Young, Genetically Related, or a Romantic Partner. *Journal of Social, Evolutionary, and Cultural Psychology 4* (3), pp. 115-127.
Navarrete, C. D., McDonald, M. M., Mott, M. L., & Asher, B. (2011). Virtual Morality: Emotion and Action in a Simulated Three-Dimensional "Trolley Problem". *Emotion 12* (2), pp. 365-370. DOI: 10.1037/a0025561.
[68] Zoals opgenomen in het Verdrag van Genève nopens het wegverkeer (1949) en het Verdrag van Wenen inzake het wegverkeer (1968). Zie ook Luetge, 2017.