Lode Lauwaert

Wij, robots

Lannoo Campus

Lode Lauwaert

Wij, robots

Een filosofische blik op technologie en artificiële intelligentie

Lannoo Campus

Voor Jesse

Inhoud

De techgemeenschap heeft geen zelfinzicht. Wij zijn geen humanisten of filosofen. Wij zijn ingenieurs. Voor Google en Facebook zijn mensen algoritmen.

John Batelle, medeoprichter van het technologietijdschrift *Wired*

Inleiding

Filosoof met de hamer

Mensen worden ontslagen door machines. Stephen Normandin is een van hen. Hij is een legerveteraan die tijdens zijn loopbaan in Arkansas in de Verenigde Staten kookte voor vluchtelingen uit Vietnam. De voorbije vier jaren reed hij in Phoenix rond voor Amazon om pakketjes aan huis af te leveren. Tot voor kort dus. De technologie van het miljardenbedrijf die hem onafgebroken in de gaten hield, kwam tot de conclusie dat hij zijn werk niet naar behoren deed. Normandin kreeg daarop een geautomatiseerd bericht. Hij mocht vertrekken. Dat was het harde oordeel. Het kwam niet van een mens, maar van moderne technologie.[1]

Deze ochtend ben ik wakker geworden. Links van mij lag op het nachtkastje een Samsung smartphone. En ik moet toegeven: het eerste wat ik na het wakker worden heb gedaan, is mijn telefoon aanzetten. Maar ik ben niet de enige. Naar schatting de helft van de mensen die een smartphone heeft, zou ’s ochtends in bed al aan het scrollen gaan. Mijn toestel is overigens dertien centimeter lang. Ook dat is niet uitzonderlijk, want anders dan de eerste mobiele telefoons is een smartphone tegenwoordig ongeveer veertien centimeter groot. Dat is vooral makkelijk als je een man bent. Die afmetingen komen immers overeen met de gemiddelde lengte van de handen van mannen, die over het algemeen groter zijn dan die van vrouwen. Aan de smartphone kleeft een mannelijke blik.[2]

Iets anders. Op de laatste dag van het jaar 2019 ging een rood licht branden in de kamers van het Canadese softwarebedrijf BlueDot. Op basis van de analyse van ontelbare berichten kwam de technologie van het bedrijf tot de conclusie dat in de Chinese stad Wuhan een uitbraak zou plaatsvinden van het coronavirus SARSCoV-2. Toen

duidelijk werd dat we in een pandemie zaten, rolden Chinese bedrijven gezichtsherkenningstechnologie uit om burgers die geen gezichtsmasker droegen te detecteren. In tal van landen ontwikkelde men een app die de burgers gratis konden downloaden en die signaleerde dat zij in nauw contact waren geweest met een besmet persoon. Daarnaast werd technologie ingezet om het virus op te sporen en te bestrijden. Er werd een algoritme ontwikkeld dat in minder dan een minuut met een hoge mate van waarschijnlijkheid kon aangeven of iemand al dan niet besmet is. Google zette vervolgens haar technologie in om uit te vlooien uit welke bestanddelen het virus is opgebouwd, met succes overigens. Het ontdekte dat de zogeheten *spike*-proteïne het geheime wapen was waarmee het virus zich aan lichaamscellen bindt.[3]

Van corona naar kolonialisme. De Indiase stad Bangalore wordt vandaag soms het Silicon Valley van het Oosten genoemd, omwille van de sterke focus op de communicatie- en informatietechnologie. Maar India is al enige tijd een voorloper op het vlak van technologie en dat heeft veel te maken met het twee eeuwen durende koloniale bewind van Engeland vanaf het midden van de achttiende eeuw. In 1819 zag de Indiase bevolking de eerste stoomboot op een binnenlandse rivier varen. Niet dat het daar zelf veel aan had. De boot was speelgoed dat door de Britten aan een prins uit de kolonie was geschonken. Na de stoomvaart zorgden de Britten ook voor de aanleg van een uitgestrekt spoorwegennet. Om een idee te geven: toen India zich in 1947 van de Britten losmaakte, lag er ongeveer zeventigduizend kilometer aan spoorwegrails. Een andere technologie die tijdens het koloniale regime werd geïntroduceerd, was de telegraaf. De aanleg van de lijn startte in het midden van de negentiende eeuw en overspande een afstand van meer dan duizend kilometer tussen Calcutta en Agra. Amper enkele jaren later waren er ook al onderzeese kabels tussen Engeland en India. Het is duidelijk dat die investeringen niet waren bedoeld om de Indiërs te plezieren, maar wel om de macht van de Britten te vergroten. Door de stoomtrein konden de troepen zich sneller verplaatsen, door de telegraaf kon

men goed communiceren met het moederland, en door de scheepvaart konden grondstoffen naar Engeland worden gebracht.[4]

Het sterftecijfer wordt vandaag de dag niet sterk beïnvloed door bacteriële infectieziekten. Dat is ooit anders geweest. Toen er nog geen antibiotica waren – voor Alexander Fleming in 1928 penicilline ontdekte – waren ze de hoofdoorzaken van overlijdens. Het gevaar bestaat dat het aantal dodelijke bacteriële infecties opnieuw zal toenemen. Dat komt doordat veel bacteriën intussen resistent zijn tegen antibiotica, onder meer omdat we de voorbije honderd jaar steeds meer zulke geneesmiddelen hebben geslikt. Omdat de ontwikkeling van nieuwe antibiotica tijdrovend is, is recent onderzoek aan het Massachusetts Institute of Technology (MIT) in dat opzicht veelbelovend. Onderzoekers hebben een AI-systeem gevoed met de structuren van duizenden antibiotica om de chemische structuur te achterhalen van de moleculen die de bacteriën doden. Nadien heeft men het model zesduizend stoffen voorgelegd. Het AI-systeem selecteerde één molecule: halicine, zo genoemd naar de computer HAL uit *2002: A Space Odyssey* uit 1968 van Stanley Kubrick. Het voorspelde dat die molecule antibacteriële activiteiten zou vertonen. Bovendien verschilt de chemische structuur van halicine van die van alle bestaande antibiotica.[5]

Er is een sterke band tussen technologie en ecologie. Sommige technologieën – smartphones bijvoorbeeld – gaan slechts enkele jaren mee, maar dragen tegelijk wel de sporen van een eeuwenoude natuurgeschiedenis. In het midden van de negentiende eeuw genoot een tropische boom de interesse van de techwereld: de *Palaquium gutta* uit Azië. De reden was dat het melksap van die boom een stof bevatte, het zogenaamde guttapercha. Die stof was nuttig omdat men er de trans-Atlantische telegraafkabel op de bodem van de oceaan mee kon isoleren, de kabel waardoor er via morsecode tussen Noord-Amerika en Europa kon worden gecommuniceerd. Het probleem was wel dat de eerste kabel 250 ton guttapercha vereiste en dat voor een ton duizenden stammen van de *Palaquium gutta* nodig waren. In Azië

zijn alleen al omwille van die reden verschillende jungles verdwenen. Ook vandaag spelen zulke zaken.

Voor de productie van iPhones, drones, motorvoertuigen, batterijen en cameralenzen worden tal van mineralen gebruikt die men haalt in onder meer Indonesië, Congo en Mongolië. De voorraad is niet oneindig en als de mineralen niet meer beschikbaar zouden zijn, zou de groei van technologie stagneren. Het zal daarom wellicht niet verbazen dat niet ver van Silver Peak in Nevada een Tesla Gigafactory werd gebouwd. In Silver Peak wordt lithium gedolven, wat wordt gebruikt voor de productie van batterijen; voor de batterijen van de elektrische wagens van Tesla zijn vele grote hoeveelheden lithium nodig.[6]

Tot slot Uber. Ik log in op een app, geef aan waarheen ik me wil verplaatsen en vervolgens matcht een algoritme me aan een chauffeur die vrij is en zich in de buurt bevindt. Snel en goedkoop voor mij, en ook voordelig voor de chauffeur: hij of zij verdient makkelijk geld en ontdekt via klanten verschillende culturen enkel en alleen door wat rond te rijden in de stad. Geweldig toch? Volgens een studie in de Verenigde Staten zouden bedrijven als Uber hebben geleid tot de afname van het gebruik van het openbaar vervoer en tot de toename van het verkeer op de weg, en dus tot meer uitstoot van broeikasgassen. Vrouwelijke bestuurders verdienen gemiddeld 7% minder dan mannen. De auto’s van Uber zijn minder toegankelijk voor mensen met lichamelijke beperkingen dan reguliere taxi’s. In 2019 kondigde het bedrijf aan te zullen werken met een zogeheten zwijgknop. Als een chauffeur volgens een klant te veel praat, kan de klant op een knop drukken die de chauffeur de boodschap geeft om te rijden en zwijgen. Uber gebruikt slimme technologie die chauffeurs in de richting van een bepaald type klanten stuurt: de rijken, zij die misschien meer voor een autorit kunnen betalen. Als dat laatste het geval is, kan het bedrijf die extra opbrengsten volledig voor zichzelf houden, zonder de chauffeur meer te betalen. Alleen al daarom zal wellicht dit cijfer niet verwonderen: in het begin van 2019 werd het

vermogen van de medeoprichter van Uber, libertariër Travis Kalanick, geschat op ongeveer zes miljard Amerikaanse dollars.[7]

Technologie en AI

Dit boek gaat over technologie. Ik zal het over verschillende soorten technologie hebben en besteed in het bijzonder aandacht aan kunstmatige of artificiële intelligentie (AI). Het zal dus gaan over slimme en domme dingen: treinen, stoomboten, apps, vliegtuigen, robots, horloges, fietsen, boormachines, planningssoftware, bruggen, koffieapparaten, vibrators, katoenmachines, kerncentrales, telefoons, sociale media, prikkeldraad, tandenborstels, auto's en computers.

Er zijn een aantal goede redenen om daar aandacht aan te besteden. Ten eerste maakt technologie integraal deel uit van ons leven, in die mate dat er heel weinig of zelfs nauwelijks nog zaken zijn die we doen zonder gebruik te maken van dingen gemaakt door ontwerpers en computerwetenschappers. Ten tweede: uit een recente poll van het Kenniscentrum Data & Maatschappij onder meer dan duizend Vlamingen blijkt dat twee op de drie respondenten niet goed weet wat AI precies is. De helft van de deelnemers weet niet of die technologie aanwezig is in hun leven.[8] Ten derde is technologie verantwoordelijk zowel voor problemen als voor de toename van welvaart en welzijn. De industrialisering zorgde voor zwarte rook en zwarte rivieren, maar medische technologie heeft ertoe geleid dat we ziekten sneller kunnen voorkomen, opsporen en genezen. Ten vierde: de overheden en bedrijven in België en Nederland investeren jaarlijks veel (belasting)geld in het onderzoek naar en de ontwikkeling van technologie. AI is het wel goed voor ogen te houden dat de bedragen in het niets verzinken vergeleken met wat in de Verenigde Staten en China wordt geïnvesteerd in technologie en AI. Ten vijfde is de techindustrie een van de leidende en meest kapitaalkrachtige sectoren van onze samenleving. Naast *Big Finance* en *Big Pharma* bestaat ook *Big Tech*, en dan vooral de *Big Five*: Amazon, Google,

Microsoft, Apple en Facebook. En ten slotte: technologie is zo oud als de menselijke soort zelf. De ontwikkeling van gentechnologie en de zelfrijdende auto werd voorafgegaan door het maken van vishaken en netten, en door het maken van messen op basis van horens en slagtanden.

Sommigen vinden die laatste opmerking wellicht merkwaardig of zelfs onterecht. Zijn vishaken en netten dan technologieën? Naar wat verwijzen mensen met andere woorden met de term ‘technologie’? Wat bedoelen we eigenlijk wanneer we zeggen dat iets technologie of, meer specifiek, AI is? Het is belangrijk om daar bij stil te staan, omdat duidelijk moet zijn waarover ik het precies wel en niet zal hebben, om misverstanden te vermijden.

DE BETEKENIS VAN TECHNOLOGIE

Om te beginnen wil ik erop wijzen dat het woord ‘technologie’ taalkundig verwant is aan tal van woorden die niettemin een andere betekenis hebben. We zeggen dat Nederlander Jaap Stam op *voetbaltechnisch* vlak destijds minder begaafd was dan de intussen overleden Diego Armando Maradona, die tijdens de opwarming voor de wedstrijd tussen Napoli en Bayern München in 1989 het publiek vermaakte met enkele *technische* hoogstandjes. De legendarische Argentijn deed dat op muziek van de band Opus, muziek die niet te vergelijken is met de *technomuziek* van Ken Ishii en die helemaal anders zou klinken zonder de hulp van een *geluidstechnicus*. Andere bekende tegen ‘technologie’ aanleunende woorden zijn ‘techneut’, ‘breitechniek’, ‘technocratie’, ‘techniciteit’, ‘technowetenschap’, ‘technologisering’, ‘technofilie’ en ‘technofobie’. De rode draad door deze lange opsomming is ‘techn-’, dat teruggaat op *tekhn* uit het Sanskriet. Die stam verwees toen, meer dan vierduizend jaar geleden, naar zaken als hout- en timmerwerk.

Daarnaast leert een blik op de geschiedenis dat het woord ‘technologie’ de voorbije tweeduizend jaar op verschillende manieren is ingevuld. Die meerzinnigheid is niet in het leven geroepen door een

spitsvondige taalwetenschapper of filosoof, maar kenmerkt nu eenmaal het gebruik van de term ‘technologie’ door miljoenen mensen door de eeuwen heen, net zoals dat voor ‘wetenschap’ het geval is overigens, en bijvoorbeeld ook voor *fairness* en ‘verantwoordelijkheid’. Op die laatste zaken ga ik later nog in, nu focus ik op de meerdere betekenissen van ‘technologie’. Zeker twee daarvan worden tegenwoordig niet of nauwelijks meer gebruikt. Welke?

De eerste en tegelijk ook oudste betekenis van ‘technologie’ dateert uit de klassieke oudheid en was zo goed als volledig verdwenen tegen het einde van de negentiende eeuw. Ze luidt dat technologie net zoals psychologie of sociologie een wetenschap is. Meer dan tweeduizend jaar geleden was technologie een studie van *liberal arts* als grammatica en retorica. Later, vanaf ongeveer de achttiende eeuw, werd ook het onderzoek naar *illiberal art*s of handenarbeid ‘technologie’ genoemd. ‘Technologie’ betekende toen ook de studie van bijvoorbeeld het koken of het werken met machines. De tweede betekenis is minder oud. Ze is onder meer terug te vinden in *Das Kapital* uit 1867 van Karl Marx. Op verschillende plaatsen gebruikt hij daar ‘technologie’ in de hiervoor aangehaalde betekenis, namelijk als *Wissenschaft*. Maar hij vindt ook dat er een geschiedenis van de technologie moet worden geschreven en stelt vervolgens deze retorische vraag: ‘Verdient de geschiedenis van het ontstaan van de productieve organen van de mens in de maatschappij, van de materiële basis van elke afzonderlijke maatschappijorganisatie niet evenzeer de aandacht?’[9] Hieruit blijkt dat ‘technologie’ bij de Duitse filosoof en verschillende van zijn tijdgenoten nog een tweede invulling krijgt. Het verwijst namelijk ook naar het productieproces.

Tegenwoordig verwijzen we met ‘technologie’ niet naar zoiets als wetenschap of het productieproces. Meestal heeft men twee andere zaken voor ogen: een manipulatieproces en een ding, dat doorgaans materieel is. We verwijzen met ‘technologie’ naar een manipulatieproces wanneer we het over biotechnologie hebben. De term ‘biotechnologie’ zelf wordt nog niet bijzonder lang gebruikt, maar

refereert wel aan een proces dat al eeuwen oud is. Denk aan het veredelen van dieren en planten, het proces waarbij de gewenste eigenschappen worden geselecteerd en gekruist, met als doel dat een bestaande dieren- en plantensoort in een verbeterde conditie blijft voortbestaan. Ook ‘gentechnologie’ verwijst naar een manipulatieproces. Het gaat in dat geval om een proces waarbij men een gen dat codeert voor een bepaalde eigenschap overbrengt van één organisme naar een ander, opdat het genetisch gemodificeerde organisme ook de gewenste eigenschap bezit. Daarnaast bedoelen we met ‘technologie’ ook dingen, zoals deze die acteur Ewan McGregor opsomt aan het begin van het lied ‘Choose Life’ uit de Britse cultfilm *Trainspotting* uit 1996: televisietoestellen, wasmachines, auto’s, cd-spelers en elektrische blikopeners. Aan dit lijstje kunnen we verder ook broodroosters, strijkijzers, elektrische tandenborstels, computers, fototoestellen, camera’s en medische apparaten als infuuspompen of pacemakers toevoegen.

Wanneer ik het vanaf nu over technologie heb, bedoel ik deze twee zaken: ofwel een manipulatieproces ofwel een ding als een smartphone of auto. Al moet ik daar wel onmiddellijk aan toevoegen dat het doorgaans over dat laatste zal gaan. Ik heb het *nooit* over technologie in de zin van wetenschap of het productieproces, *soms* over biotechnologie of gentechnologie, en *meestal* over dingen.

Ik neem aan dat dit aansluit bij hoe de meeste mensen de term ‘technologie’ doorgaans gebruiken. Niettemin wil ik de aandacht vestigen op het volgende. Iedereen of zo goed als iedereen mag dan wel met ‘technologie’ meestal naar dingen verwijzen, toch is er geen eensgezindheid over welke dingen nu wel en niet precies een technologie zijn. Zeker, deeltjesversnellers en smartphones zijn wellicht voor iedereen technologieën, net zoals computers en defibrillators dat zijn. Maar voor sommigen is een boormachine een technologie terwijl dat voor anderen niet zo is; de een vindt een radio wel technologie, de ander niet. Die verschillen hangen samen met andere verschillen, bijvoorbeeld verschillen in afkomst, beroepsgroep of leeftijd. Ben je bijvoorbeeld een ingenieur, dan is de kans groot dat

je een boormachine geen technologie vindt, maar ben je een antropoloog, dan beschouw je meubels en kledingstukken wellicht wel als technologieën.

Het is dus raadzaam om duidelijk te zijn over wat ik wel en niet als technologie zie. Ik hanteer een brede opvatting. Breed, in de zin dat ik onder ‘technologie’ een hele reeks zaken laat vallen: niet alleen computers, auto’s en AI, maar ook boormachines en meubels. Ik kies voor zo’n brede invulling, omdat er eigenlijk geen goede reden is om bijvoorbeeld een computer wel en een stoel geen technologie te noemen. Natuurlijk zijn er verschillen. De ene technologie is gestoeld op wetenschappelijke kennis, is gemaakt op basis van silicium en heeft knopjes, terwijl dat bij een stoel normaal gezien niet zo is. Deze en andere verschillen zijn echter niet relevant om beide dingen sterk van elkaar te onderscheiden, en om het verhaal dat ik in de komende hoofdstukken uit de doeken zal doen alleen toe te passen op bijvoorbeeld computers en niet op pakweg boormachines.

Daarnaast is het goed om te weten dat er ook regionale verschillen zijn. In Vlaanderen wordt de term ‘techniek’ doorgaans gebruikt om het over een vaardigheid te hebben, iets wat je kunt leren en trainen. In Nederland echter wordt de term ‘techniek’, meer dan in Vlaanderen althans, soms ook gebruikt om te refereren aan de zaken waarover ik het meestal zal hebben: robots, auto’s, radio’s, smartphones en zoveel andere dingen die we dagelijks gebruiken. Men gebruikt ‘technologie’ nooit of zo goed als nooit om over een vaardigheid te spreken, maar omgekeerd wordt ‘techniek’ in ons taalgebied wel af en toe gebruikt om naar hetzelfde te verwijzen waarvoor ik hier de term ‘technologie’ gebruik. Verder noemen sommigen ook de voorwerpen gemaakt door pakweg apen of vogels ‘technologieën’, en kunnen de allernieuwste intelligente robots zelf ook andere technologie op de wereld zetten. Om kort te zijn: om naar dingen als auto’s en apps te verwijzen gebruik ik de term ‘technologie’ en wanneer ik het over technologie heb, gaat het over dingen die zijn gemaakt door dieren die we ‘mensen’ noemen.

EEN BREDE OPVATTING

De voorbije opmerkingen volstaan om een ruwe definitie te geven van ‘technologie’ zoals veel mensen die term normaal gezien gebruiken, en zoals ik het ook straks zal doen, namelijk om te verwijzen naar dingen, en dan niet enkel naar smartphones maar ook naar meer basale zaken. Die definitie luidt dat alle technologie een artificieel karakter heeft, uitgerust is met een functie en materieel van aard is, of nauw verbonden is met iets dat materieel is. Laat ik dat even toelichten.

Wanneer we zeggen dat iets artificieel of kunstmatig is, bedoelen we doorgaans dat het een gevolg van menselijk handelen is – doorgaans, want ‘artefact’ wordt soms ook gebruikt wanneer het over dieren gaat die geen mensen zijn, maar zoals aangegeven is dat hier niet van belang. Die omschrijving is echter nog te algemeen, want ook het vergroten van het gat in de ozonlaag is een effect van ingrijpen door de mens. Mensen zijn dus wel oorzakelijk verbonden met de totstandkoming van een artefact, maar dat volstaat nog niet om iets als een artefact te zien. Vandaar deze verfijning: een artefact is het *bedoelde* effect van menselijk handelen. Stel echter dat ik in mijn tuin een zaadje plant, met als doel dat er een boom in mijn tuin groeit. Wanneer dat daadwerkelijk gebeurt, dan is die boom het bedoelde gevolg van mijn handeling, het in de grond stoppen van de zaadjes. Toch zullen we niet zeggen dat de boom een artefact is. Dat komt omdat dat natuurobject ook zonder mijn toedoen kan groeien. Dingen met een kunstmatig karakter zijn dus zaken die *uitsluitend* of bijna uitsluitend ontstaan op basis van menselijke interventie. Let wel, artefacten verschillen van voorwerpen uit de natuur, maar kunnen wel zijn samengesteld uit die voorwerpen. Een kano is een artefact, ook als die is gemaakt op basis van een boomstam.

Alle technologie is dus kunstmatig, maar is alles wat kunstmatig is ook technologie? Nee. Er zijn wel meer zaken die artificieel zijn, maar waarover wellicht niemand zou zeggen dat het technologie is. Denk aan muziek. Dat is niet uitsluitend geluid, maar het gevolg van het plan

van minstens één iemand om geluid op een bepaalde manier te ordenen. Muziek ontstaat dus niet zoals een boom ontstaat, maar toch beschouwen we het doorgaans niet als technologie. Dat komt omdat technologie naast een artificieel karakter ook altijd een functie heeft. Dat wil zeggen: alle technologie is ontworpen met een doel voor ogen; zonder die doelgerichtheid is een ontwerp geen technologie. Muziek kan vanzelfsprekend wel nuttig zijn, maar muziek die geen doel heeft, is nog steeds muziek. Nut zit met andere woorden niet gebakken in muziek, terwijl dat wel zo is bij technologie. Uiteraard weet ik ook wel dat technologieën soms tijdelijk of definitief niet meer werken, maar dat wil niet zeggen dat het geen technologieën meer zijn. Iets dat van menselijke makelij is, is een technologie als het is ontworpen om een doel te verwezenlijken. Dat blijft zo, ook wanneer blijkt dat het ontwerp al dan niet tijdelijk niet meer doet wat het behoort te doen.

Tot slot zijn zo goed als alle technologieën materieel, in de zin dat ze meetbaar en tastbaar zijn, dat je ze kunt aanraken en wegen. Maar dat is niet noodzakelijkerwijs zo. Chatbots bijvoorbeeld zijn technologieën, maar kun je voelen noch wegen – hoewel hun output wel meetbaar is. Aan de andere kant zijn er geen chatbots zonder dingen die een materiële dichtheid hebben: hardware. Vandaar deze omschrijving: hoewel technologie niet per definitie materieel is, is ze wel altijd verbonden met iets materieels – of die materie nu staal of ijzer is, speelt voor het overige geen enkele rol.

Technologie verschilt overigens in dat opzicht van juridische wetten, pakweg de antidiscriminatiewet. Zulke wet is duidelijk een artefact en bovendien gemaakt met een doel voor ogen: om mensen of groepen te beschermen. Toch is een wet geen technologie: de eerste hangt niet vast aan een fysieke entiteit, de tweede wel. Is geld dan wel een technologie? Geld is gemaakt door mensen, ontworpen met een doel voor ogen, en het bestaat in de vorm van munten. Dat klopt wel, maar bij geld in de vorm van munten is er geen nauwe band tussen het geld en de munten. Dat een munt geld is, is een kwestie van afspraak. Om het even wat zou in principe geld kunnen zijn: koffiebonen, snoepjes, dobbelstenen of andere zaken. In het geval van technologie echter is

er wel een sterke band. Een technologie kan het beoogde doel alleen realiseren op voorwaarde dat het deze bepaalde vorm heeft, dat specifieke gewicht heeft, enzovoort.

ALGEMENE EN SMALLE AI

Rest ons nog die andere vraag: wat bedoelen mensen doorgaans wanneer ze zeggen dat iets AI is? Er zijn minstens drie antwoorden op die vraag. Een aantal mensen gebruiken de term 'AI' net zoals we het woord 'intelligentie' gebruiken in de context van mensen en dieren die geen mensen zijn. In die context zien we intelligentie normaal gesproken als een vermogen dat wordt toegeschreven aan een mens of pakweg een hond. Vanuit dat opzicht is AI dan een vermogen, niet van een organisme, maar van een artificieel systeem. Anderen verwijzen met 'AI' dan weer naar een wetenschapsdomein waarin onderzoek wordt gedaan naar technologie die is uitgerust met intelligentie – heel vaak vindt zulk onderzoek plaats in een academische instelling, al is dat niet noodzakelijk zo. In de meeste gevallen echter wordt met 'AI' geen wetenschap of een vermogen bedoeld. Doorgaans refereert 'AI' aan technologie, zij het een technologie die is uitgerust met een vermogen dat andere technologie niet heeft. Dat is ook de betekenis die ik zal gebruiken.

Het is duidelijk dat niet alle technologie AI is, maar in mijn ogen is alle AI dus wel technologie. Dat houdt in dat een aantal zaken die daarnet aan bod zijn gekomen ook nu van toepassing zijn. Alle technologieën hebben een materieel karakter of zijn verbonden met iets materieels, en dus AI ook. Slimme systemen hebben hardware nodig; er moeten computers zijn om grote hoeveelheden informatie te verwerken en om snel te kunnen rekenen. Daarnaast is er ook geen eenduidig antwoord op de vraag welke technologie men precies op het oog heeft wanneer men het over AI heeft, net zoals er meerdere betekenissen van 'technologie' zijn. Dat heeft onder meer te maken met het onderscheid tussen deze drie zaken: *superintelligence*, algemene AI en smalle AI. Laat ik dat onderscheid even toelichten.[10]

De term ‘artificiële intelligentie’ werd gemunt door informaticus John McCarthy in de zomer van 1956 tijdens het fameuze Summer Research Project on Artificial Intelligence aan het Dartmouth College in New Hampshire in de Verenigde Staten. Tegenwoordig wordt met die uitdrukking heel soms verwezen naar kunstmatige entiteiten die de vermogens van mensen verregaand overtreffen, en die dat op meerdere vlakken doen: in cognitieve zin, door in een flits patronen in data te herkennen, en op fysiek vlak, omdat ze zich bijzonder rap kunnen verplaatsen, zowel in het water en de lucht als over de grond. Deze vorm van slimme technologie staat bekend als *superintelligence*. Ze spreekt tot de verbeelding, en hoewel het dat type van AI is dat regelmatig in films en populaire media opduikt, bestaat zulke technologie momenteel niet, en is het nog maar de vraag of die ooit zal bestaan, laat staan dat we weten wanneer ze zal worden gemaakt.

Hetzelfde geldt ook voor de tweede vorm: algemene AI (*general AI*). Dat is technologie die nauw verwant is aan hoe wij, mensen, functioneren. Ze wordt ‘algemeen’ genoemd omdat ze in staat is om al dan niet tegelijkertijd meer dan één taak uit te oefenen en om de informatie uit de ene taak ook voor een andere te gebruiken. Algemene AI is zich daarnaast ook bewust van de dingen die ze doet, is in staat om te genieten en om te lijden of pijn te hebben. Maar zoals aangegeven: zulke technologie bestaat momenteel niet, ook al zijn er momenteel wereldwijd tal van ontwerpers en computerwetenschappers die er alles aan doen om zoiets te maken. Algemene AI is zowat de droom van menig technoloog.

Wanneer politici, onderzoekers en anderen vandaag de loftrompet over AI steken of AI kritisch bejegenen, dan wordt zo goed als altijd de derde variant bedoeld: smalle AI (*narrow AI*). Het is ook de enige vorm van AI die momenteel bestaat en wordt gebruikt. Men spreekt hier over ‘smal’ in de zin dat de technologie uitsluitend in staat is om een bepaalde, zeer specifieke opdracht uit te voeren, en dat ze niet kan switchen tussen verschillende soorten opdrachten. Dat laatste is een beperking, maar tegelijk heeft die technologie wel het voordeel dat ze

heel goed is in datgene waarvoor ze is gemaakt, doorgaans beter of minstens sneller dan mensen. Concreet gaat het over de volgende taken: het herkennen van gezichten van mensen en andere dieren, bepalen of een gevangene vervroegd mag vrijkomen, natuurlijke taal verwerken, de kortste weg naar je bestemming uitstippelen, inschatten waar geweldsdelicten zullen worden gepleegd, de uitslagen van sportwedstrijden voorspellen, *fake news* verspreiden, commerciële en politieke advertenties aan de juiste doelgroep tonen, inschatten of iemand kredietwaardig is, terroristen lokaliseren, het risico op de uitbraak van een virus bepalen, kanker opsporen, muziek en series aanbevelen, verzekeringspremies vastleggen, mogelijke vrienden en partners voorstellen, het selecteren van kandidaten voor een job, het volgen en evalueren van werknemers, en zoveel meer andere dingen.

Alles wat met smalle AI te maken heeft, kun je nog verder ordenen door het onderscheid tussen twee soorten te onderstrepen: expertsystemen en *machine learning*. Wat houdt dat onderscheid precies in?

De eerste vorm van (smalle) AI, ook wel *Good Old Fashioned Artificial Intelligence* (GOFAI) genoemd, ontstond in de tweede helft van de twintigste eeuw. Het doel van deze technologie is om op basis van inkomende gegevens een beslissing, een inschatting of aanbeveling te maken. Neem MYCIN, een van de eerste expertsystemen dat in de jaren 1970 werd ontwikkeld aan Stanford University. Het ontving de data van patiënten en vervolgens schatte het de kans dat de patiënt een bloedstollingsziekte of bacteriële infectie had. Cruciaal daarbij is de fase tussen de input en de output: de verwerking van de data tot een conclusie. Dat gebeurt door een algoritme, en dat wil zeggen: een reeks stappen die moeten worden gevolgd en die vertrekkende vanuit een beginsituatie (de input) naar een doel leiden (de output). In het geval van expertsystemen zijn algoritmen regels die de vorm hebben van een ‘als-dan-constructie’. Een voorbeeld van zo’n instructie, hier in het geval van een robotstofzuiger, is de zin ‘als men tegen een hard voorwerp botst, dan moet men rechtsomkeert maken’. De reden

waarom nu technologie met zulke regels een ‘expertsysteem’ wordt genoemd, is dat de regels niet door de technologie zelf worden geschreven, maar afkomstig zijn van mensen, en in het bijzonder experts. Neem opnieuw MYCIN. Voor het formuleren van de regels voor dat systeem baseerde men zich niet op het buikgevoel, maar op de expertise van artsen over bacteriën en infectieziekten. Expertsystemen zijn met andere woorden in hoge mate mensenwerk. Men beroept zich eerst op de gespecialiseerde kennis van een expert en nadien wordt die kennis in de vorm van instructies vertaald in een computerprogramma. Dat laatste is het werk van een programmeur.

Een expertsysteem heeft het voordeel dat het transparant is. Op het moment dat een beslissing is genomen, is het niet moeilijk om te achterhalen waarop die beslissing is gebaseerd. Men kan immers terugvallen op de input en de regels. Maar er zijn ook problemen verbonden aan zulke technologie. Het vastleggen en schrijven van de regels kost veel tijd, en wanneer het systeem in een nieuwe situatie terechtkomt waarvoor geen regels werden geschreven, dan doet het niets. Een oplossing voor die euvels komt van die andere vorm van AI: *machine learning.*

Zoals de naam al zegt, gaat in het geval van machinaal leren de technologie zelf achterhalen hoe een input kan worden verwerkt tot output, gaat het AI-systeem eigenhandig leren om aan binnenkomende informatie een beslissing te koppelen. Het doet dat door in de leerfase heel veel voorbeelden te zien, lees: de technologie wordt getraind met talloze data, met bijvoorbeeld miljoenen foto’s. Anders dan een expertsysteem werkt *machine learning* datagedreven. Je kunt in deze context denken aan de spam in je mailbox. Die is daar terechtgekomen door een AI-systeem dat eerst is getraind met enorm veel voorbeelden: berichten die onwenselijk zijn (de spam) en berichten die we wel graag zouden lezen. Als het systeem is getraind, moet het inkomende berichten (input) juist weten in te schatten: hetzij als wenselijke e-mail, hetzij als spam – die voorspelling is de output. Als het daartoe in staat is, dan is het een voorbeeld van een slimme technologie die functioneert op basis van

een zelflerend algoritme, zonder de hulp van een expert of programmeur. Een ander voorbeeld van *machine learning* is het algoritme van het Zweedse muziekplatform Spotify. Dat houdt alle beslissingen van haar gebruikers op het platform bij en speurt naar gelijkenissen tussen de keuzes van de gebruikers. Als het algoritme die vindt, dan worden de gebruikers met overeenkomstige gedragingen gegroepeerd. Wanneer nu iemand uit die groep naar een album luistert – *Stadium* uit 2018 van Eli Keszler bijvoorbeeld, een geweldig album overigens – dan wordt die plaat als suggestie ook voorgesteld aan de andere leden van diezelfde cluster. Ook nu worden met andere woorden inkomende data (de keuzes op Spotify) aan een output (de suggestie) gekoppeld, en dat opnieuw zonder menselijke tussenkomst.

Het is in deze context ook nuttig om kort het volgende te onderstrepen: *machine learning* betekent niet hetzelfde als *deep learning*. Machinaal leren is niet noodzakelijk *deep learning*, maar alle *deep learning* is wel machinaal leren. *Deep learning* is met andere woorden een specifieke vorm van machinaal leren, een manier om technologieën te leren hoe ze data (input) moeten verwerken tot de beste beslissing (output). Wat *deep learning* van andere vormen van machinaal leren onderscheidt, is dat bij deze technologie de structuur van het menselijke brein is nagebootst. Ons brein bestaat uit neuronen: bijzondere cellen die signalen ontvangen van neuronen en die signalen zenden naar andere neuronen. Technologieën die werken met *deep learning* zijn gemodelleerd naar die structuur via neurale netwerken, op zo'n manier dat er verschillende lagen met kunstmatige neuronen zijn. Vaak hebben neurale netwerken drie tot vier lagen, maar sommige technologieën die leren via neurale netwerken hebben wel tien tot soms zelfs twintig lagen. Het is ook naar die gelaagde neurale structuur dat wordt verwezen met *deep*.

Wanneer over AI wordt gesproken, bedoelt men soms enkel expertsystemen als MYCIN. Zulke technologie wordt ook gebruikt voor bijvoorbeeld spelling- en grammaticacontrole. In andere gevallen gaat het over de combinatie van expertsystemen en *machine learning* – die

combinatie is volgens technologen overigens ook de manier waarop AI hoofdzakelijk zal moeten draaien in de toekomst. Meestal echter bedoelen mensen met ‘slimme systemen’ technologieën die enkel en alleen gestoeld zijn op een zelflerend algoritme. Voorbeelden zijn de software van Netflix die series en films aanbeveelt, en AI-systemen die artsen helpen om kanker op te sporen. Ik volg dat gebruik van de term ‘AI’. In de hoofdstukken die volgen, bedoel ik met ‘AI’ *nooit* superintelligente technologieën of algemene AI. Wanneer ik over AI spreek, heb ik het *doorgaans* over technologie met een zelflerend algoritme, *soms* over machinaal leren in combinatie met expertsystemen, en *nauwelijks* uitsluitend over expertsystemen.

Een filosofische blik

Dit boek biedt een filosofische kijk op technologie en AI. Ook daar is zeker één goede reden voor. Momenteel blijft zo’n kijk nog sterk onder de radar. Jazeker, er zijn boeken en opiniestukken over technologie en AI vanuit het perspectief van de ondernemer of de econoom, over de vraag hoe je die zaken in een bedrijf kunt implementeren zonder het welzijn van de werknemer al te negatief te beïnvloeden, of over welke jobs nu wel en niet op de helling staan door automatisering. En natuurlijk zijn er in ons taalgebied over die thematiek tal van publicaties voor zowel een gespecialiseerd als een lekenpubliek vanuit een meer technisch, ingenieursgericht perspectief. Maar dat neemt niet weg dat globaal genomen weinig of geen aandacht wordt geschonken aan een filosofisch perspectief op technologie en AI. Sommigen malen daar wellicht niet om, maar toch is zo’n kijk naast interessant ook relevant. Ontwerpers, verkopers, gebruikers en producenten van technologie en AI, allemaal worden ze geconfronteerd met thema’s als duurzaamheid, gelijkheid, transparantie, privacy, bias en verantwoordelijkheid. Geen van die zaken is weg te denken uit de wereld van de eenentwintigste eeuw, laat staan wanneer het gaat over twee invloedrijke dingen die we tegenwoordig maken: intelligente en domme dingen. Het zijn een voor

een ethische thema's en morele bezorgdheden, en laat nu net ethiek of moraalfilosofie een van de deeldomeinen van de filosofie zijn. Alleen dat al ondersteunt de relevantie van een filosofische blik.

DE ESSENTIE VAN FILOSOFIE

Wat volgt, is dus geen inleiding op technologie en AI, ook al licht ik op meer dan één plaats voorbeelden uitgebreid toe. Het boek biedt ook geen empirisch-wetenschappelijk kader om over smartphones en stoommachines na te denken. Hoewel ik me vanzelfsprekend laat informeren door empirische studies en dit boek geen voorbeeld is van *armchair philosophy*, laat ik dus een reeks boeiende en relevante vragen uit verschillende domeinen van de wetenschap doelbewust links liggen. Economie: zullen deze en gene jobs verdwijnen? Heeft automatisering een grens? Is AI geen reden om opnieuw na te denken over een basisinkomen voor iedereen?[11] Recht: moeten alle morele problemen worden opgevangen door juridische wetten? Wat is de zin en onzin van rechten voor robots?[12] Is het haalbaar en wenselijk om rechters door AI te vervangen? Psychologie: hoe komt het dat mensen veel technologie gebruiken maar vaak ook terughoudend reageren wanneer een nieuwe ontwikkeling op de markt wordt gebracht? Wat zijn de politieke overtuigingen van de makers van technologie en AI? Hoe moeten we het fenomeen van de *uncanny valley* interpreteren, het feit dat mensen afkerig reageren op robots die heel sterk op mensen lijken? Deze en andere vragen moeten worden beantwoord door economen, juristen en psychologen. Ik zal ze in dit boek niet belichten, niet omdat ze niet uitdagend of relevant zouden zijn, want dat zijn ze minstens evenveel als de filosofische vragen die ik stel (en misschien zelfs nog meer), maar wel omdat het geen vragen zijn die bestemd zijn voor het domein van de filosofie en ik alleen vertrouwd ben met een aantal zaken uit dat domein.

Wat biedt het boek dan wel? Wat houdt een *filosofische* kijk in? Die vraag suggereert dat er iets is dat alles wat we 'filosofie' noemen gemeenschappelijk heeft, dat er iets is dat alle filosofische teksten en

redeneringen delen. Daarnaast lijkt die vraag ook te polsen naar iets dat uniek is voor filosofie, naar een eigenschap die we uitsluitend toeschrijven aan filosofie. Is er echter iets dat kenmerkend is voor *alle* filosofie en *alleen* voor filosofie?[13]

Een populaire opvatting van filosofie is dat het een kritisch licht op de dingen laat schijnen. Grote delen van de hoofdstukken die straks volgen, zijn inderdaad een regelrechte aanval op diepgewortelde overtuigingen, maar daartegenover staat dat kritiek niet het privilege van de filosofie is. Ontwerpers en bakkers, garagehouders en wiskundigen, er is weinig of geen reden om te vermoeden waarom zij minder kritisch zouden zijn dan filosofen, wel integendeel, laat staan dat zij niet kritisch zouden zijn. Een andere mogelijkheid is dat je kijkt naar het instrumentarium. Straks zal ik enkele gedachteexperimenten opzetten – ongewone denkbeeldige situaties – met als doel om sommige van onze intuïties over technologie te toetsen, bijvoorbeeld de overtuiging dat robots niet kunnen lijden. Wanneer je nu de indruk hebt dat dit filosofie afzondert van andere disciplines, dan is dat onterecht. Ook wetenschappers maken gebruik van zulke experimenten – denk aan het valexperiment van wiskundige Giambattista Benedetti uit de zestiende eeuw – en bovendien zijn er ook filosofen die zulke methoden nooit gebruiken. Een andere mogelijke piste is dat je kijkt naar het soort vragen dat in de filosofie wordt gesteld. Die zouden niet alleen moeilijk te beantwoorden zijn, maar daarnaast ook, en vooral, zeer abstract. Dat klopt voor tal van vragen afkomstig van filosofen, maar zeker niet voor alle vragen. Wanneer een ethicus nadenkt over euthanasie, vegetarisme of migratie, dan zijn dat verre van makkelijke thema's. Die zijn echter zeker niet buitengewoon abstract, in ieder geval minder abstract dan sommige vragen die door theoretisch natuurkundigen worden gesteld.

Ik kan me vanzelfsprekend vergissen, maar er is waarschijnlijk geen eenduidig antwoord op de vraag wat filosofie nu precies wel en niet is, althans niet voor zover die vraag verwijst naar een eigenschap die kenmerkend is voor enkel en alle filosofie. Anders geformuleerd: wanneer mensen zeggen dat iets 'filosofie' is, dan kunnen ze daar

verschillende dingen mee bedoelen, net zoals ‘technologie’, ‘AI’ of pakweg ‘duurzaamheid’ meerzinnige termen zijn. Sommigen vinden dat wellicht vervelend, want we hebben de neiging om de dingen te onderscheiden en netjes in een hokje te duwen. Maar laat dit dan een troost zijn: filosofie staat in dat opzicht niet alleen. Er zijn nog andere domeinen dan de filosofie die niet kunnen worden samengebracht op basis van een unieke eigenschap. En wat zou verder bijvoorbeeld de eigenschap zijn die alle en enkel sport, oorlog of religie karakteriseert?

BESCHRIJVEN EN OORDELEN

Op zich hoeft het niet bijzonder problematisch te zijn dat er geen conceptuele eensgezindheid bestaat, dat er onder de noemer van filosofie uiteenlopende zaken vallen die soms weinig met elkaar te maken hebben. Het houdt alleen in dat ik duidelijk moet zijn over wat ik hier in dit boek onder een filosofische blik versta. Het is daarom goed om te beginnen met uit te leggen waarop die blik precies gericht zal zijn.

Mensen hebben diepgewortelde en wijdverspreide overtuigingen over de meest uiteenlopende zaken. Het zijn ideeën die al een tijdje meegaan, soms al enkele decennia of zelfs eeuwen, en die terugkeren in verschillende lagen van de bevolking. Sommige van die overtuigingen zijn vrijblijvend of volkomen oninteressant, andere zijn dan weer niet zonder praktische gevolgen. Enkele voorbeelden: mensen denken of dachten dat ze een ziel hebben (en dat robots zoiets niet hebben), dat geluk het doel van het leven is, dat je vlees moet eten om gezond te zijn, dat succes uitsluitend een eigen verdienste is, dat er een wereld buiten hen bestaat, dat hard werken de toegangspoort tot geluk is, dat het kapitalisme de hemel of hel is, dat de aarde plat is en in het midden van het heelal staat, dat genieten van het leven belangrijk is, dat ze al dan niet door God geschapen zijn, dat ze de kroon van de schepping of bijzonder zijn, dat ze vrij zijn – en een hele resem andere zaken.

Het punt is nu dat er ook over technologie en AI sterke overtuigingen bestaan. Die denkbeelden gaan al lang mee – sommige dateren uit de oudheid – en worden door tal van mensen gedeeld: door Mark Zuckerberg en de oprichters van Google, door ontwerpers en computerwetenschappers, door mensen uit de kringen van *Big Tech*, maar ook door mensen buiten de cloudwereld – politici, filosofen, wetenschappers, opiniemakers. De focus van het boek ligt bij zulke beweringen, met name bij deze drie: technologie is neutraal, AI is disruptieve of ontwrichtende technologie, en technologie moet je in termen van determineren begrijpen. Het spreekt voor zich dat er nog andere gebetonneerde beweringen zijn die ik onder de loep zou kunnen nemen – dat technologie toegepaste wetenschap is bijvoorbeeld, ik kom daar kort nog op terug in het laatste hoofdstuk – maar ik kies deze drie omwille van de volgende twee redenen. Vooreerst zijn deze overtuigingen verreweg de bekendste. Daarnaast hangen ze, anders dan andere beweringen, niet in het luchtledige. Het vasthouden aan die stellingen is met andere woorden vaak niet belangeloos is en heeft soms ook verregaande praktische gevolgen.

Dit boek gaat dus over het al dan niet neutrale, disruptieve en determinerende karakter van technologie. Het vestigt daarmee de aandacht op de verhalen die over technologie en AI routineus worden afgehandeld in workshops en advertenties, de stellingen waarop soms onwenselijke techpraktijken zijn gestoeld, de bekende praatjes van Silicon Valley, de geloofsbelijdenissen van technofielen en technopessimisten, de straffe thesen van cyberutopisten en technoalarmisten, de manier waarop tegenwoordig en sinds lang over technologie wordt gesproken door beleidsmensen en reclamemakers, de vastgeroeste overtuigingen van ons – consumenten en gebruikers –, over de dingen die we voortdurend bij ons hebben: slimme en domme technologie. De populaire verhalen over onder meer zelfrijdende auto’s en telefoons: zie hier wat op het spel staat in dit boek.

Het eerste hoofdstuk zoomt in op de vermeende neutraliteit van technologie, en gaat over technologie in het algemeen, niet enkel over

AI. Ik focus onder meer op een morele waarde als *fairness* en op de vraag of zulke waarde al dan niet in technologie gebakken zit. Daarvoor beroep ik me bijvoorbeeld op bruggen ontworpen met racistische bedoelingen, discriminerende algoritmen, boormachines, *Temptation Island* en genderneutrale videospellen. Het tweede deel, over AI als disruptieve technologie, focust louter op slimme technologie, met bijzondere aandacht voor deze ethische thema's: privacy, bias, morele verantwoordelijkheid, ecologie, transparantie, misbruik en veiligheid. In het derde hoofdstuk belicht ik de stelling dat je technologie, en daarmee bedoel ik in die context niet enkel AI, in termen van determineren moet verstaan. Is technologie een product van de samenleving of wordt de samenleving gedreven door de werken van datawetenschappers, ingenieurs en computerwetenschappers? Tal van voorbeelden passeren de revue: de planningssoftware van Starbucks, kerncentrales, fietsen, sociale media, het QWERTY-toetsenbord en de stoommachine.

In de drie hoofdstukken – die je overigens los van elkaar kunt lezen – zal ik telkens overwegend drie zaken doen, hoewel het gewicht bij de eerste twee zal liggen. In de eerste plaats zal ik in elk hoofdstuk duidelijkheid creëren, orde scheppen. Er is immers iets merkwaardigs aan de hand met de stellingen waarop ik mijn aandacht zal richten. Aan de ene kant gaan ze vaak al lange tijd mee en worden ze door velen gedeeld, maar als je aan de andere kant vraagt wat ze precies wel en niet betekenen, dan blijft men meestal het antwoord schuldig of komt men niet verder dan enkele wollige algemeenheden. Vandaar dat ik eerst en vooral een kraakhelder zicht probeer te krijgen op de bewering dat technologie neutraal is of dat technologie de samenleving determineert. Want wat betekenen 'neutraliteit' en 'determineren' eigenlijk wanneer het over technologie en AI gaat? Wat moeten we onder verantwoordelijkheid verstaan? Kunnen we meerdere interpretaties onderscheiden? Zo ja, staan die los van elkaar of zijn ze nauw met elkaar verbonden? Het streven naar precisie en duidelijkheid, niet door vaagheden te versterken met

andere vaagheden, maar door beweringen en begrippen te ontleden en uit elkaar te houden, dat is wat ik straks zal doen.

Daarnaast zal ik, ten tweede, evalueren. Wanneer we ons een beeld hebben kunnen vormen van de beweringen is het tijd om te oordelen. Wat zijn de argumenten voor en tegen de these dat technologie neutraal is? Welke zijn doorslaggevend? Zijn er redenen die de stelling ondersteunen dat AI disruptief of ontwrichtend is? Of zijn er ook redenen om daaraan te twijfelen of zelfs om die bewering aan de kant te schuiven? De vraag luidt dus: kloppen de populaire vastgeroeste verhalen? Ik geef hier mijn besluit al mee. Sommige technologie is inderdaad neutraal, maar dat is niet noodzakelijk zo – tal van dingen zijn waardegeladen; wat ethiek betreft is er mijns inziens geen reden om te beweren dat AI disruptieve technologie is; de argumenten voor de stellingen die technologie in verband brengen met de idee van determineren schieten schromelijk tekort, sterker nog, alle stellingen zijn onjuist. Straks leg ik in ieder hoofdstuk omstandig uit waarom ik denk dat dit zo is.

Tot slot zal ik in elk deel, naast begrijpen en oordelen, ook kort aangeven wat de relevantie van mijn verhaal is. Hebben mijn redeneringen alleen theoretisch belang? Of zijn ze ook in praktisch opzicht belangrijk? Is het nuttig te weten dat technologie al dan niet neutraal is? Of is dat uitsluitend een interessante maar voor het overige weinig relevante denkoefening? Het moge duidelijk zijn dat die laatste vraag louter retorisch is. Het is bijvoorbeeld goed om te weten dat technologie dikwijls waardegeladen is, omdat die geladenheid soms niet zichtbaar is en je dus een technologie kunt kopen en gebruiken die is geladen met een waarde waarmee je eigenlijk niets te maken wilt hebben. Straks meer over deze en andere zaken.

Vanzelfsprekendheden verhelderen en op de helling zetten, dat is de kern van de filosofische insteek van dit boek. Het is ook wat andere filosofen door de geschiedenis heen als hun voornaamste taak hebben gezien en het ligt in de lijn van wat Friedrich Nietzsche

bedoelde met ‘filosoferen met de hamer’. De keuze voor zo’n benadering heeft wel als gevolg dat enkele andere mogelijke invalshoeken niet centraal staan of nauwelijks een rol spelen, bijvoorbeeld een historisch of meer auteursgericht perspectief. Het is niet dat dit niet mogelijk is. Hoewel de filosofische reflectie over technologie en AI in het licht van de lange geschiedenis van de filosofie niet oud is, bestaat ze toch ook al ongeveer anderhalve eeuw. Het filosofische denken over technologie kwam van de grond in de tweede helft van de negentiende eeuw – met name sinds Ernst Kapp in 1877 zijn *Grundlinien einer Philosophie der Technik* schreef. Bovendien zijn er in de voorbije honderd jaar talloze filosofische studies over technologie verschenen. Ik denk dan voornamelijk aan het werk van Martin Heidegger en Karl Jaspers, en de studies van Bernard Stiegler of Gilbert Simondon. De opzet van mijn reflectie brengt echter met zich mee dat de bladzijden die volgen geen historisch overzicht van de techniekfilosofie schetsen, zelfs niet in grote lijnen, of dat ik geen diepgravende studie van het werk van die auteurs presenteer. Dergelijke filosofie is natuurlijk boeiend, en ook al besteed ik meer dan één bladzijde aan Heidegger en sluiten sommige delen aan bij het werk van een aantal techniekfilosofen, toch is een al dan niet beknopte geschiedenis van de techniekfilosofie of een analyse van het denken van een techniekfilosoof hier niet nodig.

Tot slot houdt mijn benadering ook in dat dit boek zich in het kamp van noch het techniekoptimisme noch het techniekpessimisme laat plaatsen. Nochtans zou het niet verwonderen mocht dat wel zo zijn. Veel studies over technologie zijn een uitgebreide lofzang op de schijnbaar oneindige mogelijkheden van technologie – dat is vaak zo wanneer ondernemers of technologen aan het woord zijn – terwijl een hele reeks andere studies een erg alarmistische toon hebben – dat is dikwijls het geval bij auteurs uit de humane wetenschappen, en daar zijn niet zelden filosofen bij. Natuurlijk laat ik mij in dit boek op meer dan één plaats erg kritisch uit, bijvoorbeeld over de ecologische gevolgen van het gebruik van technologie en AI. En ja, uiteraard ben ik zeer lovend over de technologische vooruitgang op onder meer

medisch vlak. Het boek onttrekt zich echter aan de eenzijdigheid waarmee het denken van technofielen en technofobe alarmisten kampt. Beschrijven en oordelen, daarover gaat het in de eerste plaats, en dan is zowel kritiek als optimisme onvermijdelijk.

Technologie is in feite neutraal. Het is een beetje zoals een hamer. Voor de hamer maakt het niet uit of je er een huis mee bouwt of er iemands schedel mee vermorzelt.

Noam Chomsky, taalkundige en filosoof

1

De neutraliteit van technologie

Texas, de lente van 2018. Op een middelbare school vindt een gruwelijke schietpartij plaats. Naar aanleiding van dat vreselijke drama zei Oliver North, voorzitter van de National Rifle Association (NRA): ‘Guns don’t kill people, Ritalin kills people!’ In de nasleep van de shootings in Ohio en (opnieuw) Texas zei een jaar later toenmalig president van de Verenigde Staten Donald Trump iets soortgelijks: ‘Mental illness and hatred pull the trigger, not the gun.’

Beide uitspraken verwijzen naar de leuze ‘Guns don’t kill people, people kill people!’, die ongeveer een eeuw geleden voor het eerst werd gebruikt. Tijdens de jaren 1920 werden de Verenigde Staten geplaagd door een epidemie van geweld, vaak met een dodelijke afloop. In bars vonden veel gewelddadige confrontaties tussen bendes plaats, die meestal werden beslecht met pistolen. Een van de gevolgen daarvan was dat de American Bar Association in 1922 een pleidooi hield voor de sluiting van fabrieken die wapens voor burgers produceerden. Het is niet verwonderlijk dat dit pleidooi, dat op gespannen voet staat met wat sinds 1791 ‘het tweede amendement’ wordt genoemd, veel gehoor vond bij een groot deel van de bevolking. Evenmin verrassend is dat de wapenindustrie weinig opgetogen was met de zet van de American Bar Association, en dat tal van fabrikanten ertegen protesteerden. Wellicht het bekendste voorbeeld van dat protest was een artikel dat verscheen in 1927 in *The Manufacturer*, een blad voor Amerikaanse fabriekseigenaars, en waarin de zin ‘Guns don’t kill people, people kill people!’ stond. Of je het er nu mee eens bent of niet, die zin is sindsdien zowat dé slogan van de voorstanders van het wapenbezit. Dat blijkt onder meer uit het feit dat de NRA enkele decennia later stickers met daarop diezelfde

woorden liet drukken, stickers die voornamelijk op de bumpers van wagens werden gekleefd.

Tegenwoordig duikt die slogan soms ook in een totaal andere context op. Hij wordt bijvoorbeeld gebruikt als men iets duidelijk wil maken over technologie. Men grijpt erop terug, niet om iets over wapens in het bijzonder te zeggen, maar wel om een stelling over technologie in het algemeen te onderstrepen. Het gaat hier over de populaire stelling dat technologie neutraal is, en die ik vanaf nu 'de neutraliteitsthese' noem. Natuurlijk zijn er tal van zaken die in geen enkel opzicht neutraal zijn. Denk aan ouders, het is voor hen (zo goed als) onmogelijk neutraal te zijn als hen de vraag wordt voorgelegd voor wie ze in een noodgeval kiezen: hun eigen kind of het kind van een vreemde. Maar anders dan een ouder en net zoals scheidsrechters, zo luidt de neutraliteitsthese, is technologie neutraal. *Guns don't kill people, people kill people.*

Wie beweert dat bijvoorbeeld mobiele telefoons en apps neutraal zijn, vindt daarmee het warm water niet uit. Zulke technologieen mogen dan wel relatief nieuw zijn, de bewering dat artefacten neutraal zijn, is dat zeker niet. De neutraliteitsthese is al minstens zo oud als de (westerse) filosofie zelf – ze bestond al in het oude Griekenland. Het waren de stoïcijnen die rond de vierde eeuw voor onze jaartelling zeiden dat technologische artefacten neutraal zijn.

De neutraliteitsthese gaat niet alleen ver terug in de tijd, ze is ook wijdverspreid. Dat is ook wat je leest op de openingsbladzijde van het beruchte essay *Die Frage nach der Technik* uit 1954 van Heidegger: 'Maar het ergst zou zijn dat we aan de techniek zijn overgeleverd als we haar als iets neutraals beschouwen; want die opvatting, *die men vandaag de dag bijzonder graag huldigt*, maakt ons volslagen blind voor het wezen van de techniek.'[14] Heidegger schreef zijn tekst in het midden van de vorige eeuw. Als we naar ons tijdsgewricht kijken, waar vinden we dan de gedachte dat technologie neutraal is nog terug?

Filosoof Joseph Pitt verdedigt de befaamde stelling in zijn tekst met de veelzeggende titel '"Guns don't kill, people kill"; values in and/or around technologies' uit 2014.[15] In het boek *Hello World – How to be Human in the Age of the Machine* van wiskundige en radiomaakster Hannah Fry uit 2018 lezen we het volgende: 'Een voorwerp of een algoritme is op zichzelf nooit goed of slecht. Waar het om gaat is hoe ze gebruikt worden.'[16] Taalwetenschapper en politiek activist Noam Chomsky stelde in 2014 dan weer onomwonden: 'Technologie is in feite neutraal. Het is een beetje zoals een hamer. Voor de hamer maakt het niet uit of je er een huis mee bouwt of er iemands schedel mee vermorzelt.'[17] En tot slot was de neutraliteitsthese ook op de achtergrond aanwezig tijdens de *hearing* van Mark Zuckerberg in de Amerikaanse Senaat in april 2018. Toen senator Ted Cruz het woord kreeg en zich tot de oprichter en CEO van Facebook richtte, vroeg hij Zuckerberg of Facebook een neutraal publiek forum is. Zuckerberg antwoordde dat Facebook een platform voor *alle* ideeën is.

Dat technologie neutraal is, is dus ook vandaag de dag nog in zekere zin een breed gedragen stelling. Maar klopt ze wel? Hoe overtuigend is die bewering? Is die even bekende als vaak terugkerende gedachte over technologie juist of onjuist? In het hiervoor aangehaalde citaat stelt Heidegger niet alleen dat de neutraliteitsthese zeer populair is, hij stelt ook onomwonden dat er van alles aan schort. Maar wat precies? Is technologie dan niet neutraal? Heeft híj gelijk, of heeft Zuckerberg het bij het rechte eind?

Op het eerste gezicht lijkt het vanzelfsprekend wat iemand als Chomsky beweert. Het lijkt erg moeilijk, zo niet onmogelijk, om bijvoorbeeld een hamer *niet* als neutraal te zien. Niettemin zijn er tal van zaken die de neutraliteitsthese minstens op de helling lijken te zetten. Neem het gebruik van AI door rechtbanken om het risico op recidive in te schatten.[18] In 2016 werd een van die systemen, het intussen befaamde COMPAS-algoritme – voluit *Correctional Offender Management Profiling for Alternative Sanctions* – doorgelicht door de onderzoeksjournalisten van de onlinenewsroom ProPublica. Hoewel de ontwerpers van COMPAS geen gewicht toekenden aan afkomst,

bleek dat mensen met een witte en donkere huidskleur verschillend werden behandeld. Er waren beduidend meer personen met een donkere huidskleur van wie onterecht werd gezegd dat ze een groot risico vormden (vals positieven); er waren onevenredig veel mensen met een witte huidskleur van wie men opnieuw onterecht dacht dat ze een laag risico op recidive hadden (vals negatieven). Concreet: een vrouw van kleur die een fiets had gestolen, werd als een ernstigere bedreiging gezien dan een man met een witte huidskleur die al enkele jaren in de gevangenis had gezeten wegens gewapende overvallen. Een ontwerp dat op het eerste gezicht volkomen neutraal lijkt te zijn, blijkt bij nader inzien mensen dus ongelijk te behandelen op basis van een kenmerk (huidskleur) dat op zichzelf genomen moreel volstrekt irrelevant is. Het is een voorbeeld van wat ik ‘algoracisme’ noem. Onrecht is het CO_2 van AI.

Onbedoeld, zeer zeker, en bovendien slechts een neveneffect, maar daarom niet minder schadelijk. Al mag dat de aandacht niet afleiden van de ecologische effecten van de bouw en het gebruik van AIsystemen – meer daarover in het tweede hoofdstuk.

Enkele andere voorbeelden. Uit onderzoek aan de universiteit van Pittsburgh is gebleken dat jongeren die veel tijd doorbrengen op sociale media een grotere kans op depressie hebben.[19] In 2015 maakte het magazine *Glamour* bekend dat de eerste elf afbeeldingen die verschijnen als je in Google ‘CEO’ typt foto’s van mannen zijn. Slechts de twaalfde is van een vrouw, bovendien een foto van Barbie. Googelde je in de zomer van 2018 ‘idiot’, dan zag je eerst een foto van Trump. Economen kwamen tot de bevinding dat Facebook tot discriminatie kan leiden.[20] Op het sociaal netwerk plaatsten ze advertenties voor jobs in de zogeheten STEM-richtingen: *science*, *technology*, *engineering* en *mathematics*. En wat bleek? De algoritmen plaatsten de advertenties eerder in de *feed* van mannen dan die van vrouwen. In 2015 ten slotte postte webontwikkelaar Jacky Alcine, een man met een donkere huidskleur, een bericht op Twitter. Hij liet weten dat Google, dat nota bene de beste algoritmen heeft ontwikkeld voor de organisatie van informatie, zijn foto als ‘gorilla’

labelde. Die zaken zijn niet alleen onwenselijk, ze zijn ook niet verrassend als je weet dat in Silicon Valley vier vijfde van de werknemers mannen met een witte huidskleur zijn. Diversiteit staat in de techwereld, en zeker in de middens van de *Big Five* – Amazon, Apple, Facebook, Google en Microsoft –, vaak niet hoog op de agenda.

Zijn technologieën neutraal? Of zijn de algoritmen waarmee die technologieën zijn uitgerust eerder *weapons of math destruction*?[21] Stenen zijn stenen. Ze zijn gesloten, *en-soi*, aldus Jean-Paul Sartre. Maar geldt dat ook voor technologie? Zit er een opening in die schurende, piepende, krakende, staalharde dingen, een opening waarlangs waarden, normen, ideologieën of stereotypen kunnen binnendringen?

Op de volgende bladzijden denk ik na over de neutraliteitsthese, de stelling dat technologie neutraal is. Zoals eerder in de inleiding aangekondigd, volg ik grofweg drie lijnen – maar ik besteed het meeste aandacht aan de eerste twee punten. Ten eerste probeer ik te begrijpen. Wat houdt de neutraliteitsthese precies wel en niet in? Ik moet daarvoor onder meer inzoomen op de term ‘waarde’. Wat bedoelen we als we zeggen dat bijvoorbeeld privacy en *fairness* waarden zijn? Ten tweede ga ik evalueren. Wat spreekt in het nadeel van de neutraliteitsthese, wat in het voordeel? Ik argumenteer dat de neutraliteitsthese onjuist is, maar dat er toch een zekere waarheid in schuilt: sommige technologie is niet neutraal, andere wel. Ten derde vraag ik me af of mijn analyse praktische gevolgen heeft. Zijn de komende paragrafen relevant voor ingenieurs, computerwetenschappers, AI-ontwikkelaars? Dat is de vraag die ik meer naar het einde toe zal beantwoorden, overigens in positieve zin. Wat ik beweer, heeft gevolgen voor onze morele relatie met technologieën en technologen, met dingen en mensen, dode materie en organismen.

Waarde hebben en zijn

In zekere zin trap je een open deur in als je zegt dat technologie niet neutraal is. Het is immers voor iedereen duidelijk dat geen enkele technologie vanuit een bepaald opzicht neutraal is. Technologische instrumenten bestaan niet voor om het even wat, men maakt ze met een welbepaald doel voor ogen. Er ligt dus een keuze in het artefact besloten, een keuze van de ontwerper om *dit* – en niet *dat* – doel te realiseren aan de hand van de technologie. Iedereen weet dat, en dus ontkent niemand dat technologie in deze zin niet neutraal is. Wat verstaat men dan wel onder de neutraliteitsthese?

Je kunt grofweg twee interpretaties onderscheiden: een brede en een smalle. De eerste is breed omdat ze over tal van zaken gaat: ideologieën, opvattingen, normen, perspectieven, mensbeelden, en *ook* waarden. Ze luidt dat geen enkele technologie is geladen met normen, perspectieven en opvattingen, dat die zaken niet in de technologie gebakken zitten. Mocht dat wel zo zijn, dan zou dat inhouden dat je op voorhand weet dat het gebruik van een technologie welbepaalde normen, wereldbeelden of stereotypen bestendigt, uitdrukt of realiseert. Maar, zo luidt de brede invulling van de neutraliteitsthese, dat is niet het geval; technologie hangt niet vast aan deze of gene opvatting of ideologie. De andere interpretatie noem ik ‘smal’, omdat ze *uitsluitend* over waarden gaat. Ze luidt dat technologie op zich losstaat van waarden, dat ze niet waardegeladen is. Technologie is waardeneutraal of waardevrij, aldus de smalle interpretatie van de neutraliteitsthese. Wanneer je apps of andere instrumenten gebruikt, dan onderstreep of verwezenlijk je niet per definitie ook een waarde als duurzaamheid of autonomie. Dat is althans wat je gelooft wanneer je de smalle invulling van de neutraliteitsthese verdedigt.

In dit hoofdstuk zoom ik uitsluitend in op die laatste invulling. Daar zijn een aantal redenen voor. Ten eerste is de interpretatie die over waarden gaat de bekendste; het is de invulling die het vaakst wordt

gebruikt. Daarnaast gaat ze, anders dan de brede opvatting, over ethiek, en dat is precies een van de expertises van de filosofie. Tot slot is de brede interpretatie niet erg geloofwaardig. Technologie staat duidelijk niet noodzakelijk los van bijvoorbeeld stereotypen, denkbeelden, ideologieën. Dat blijkt al snel wanneer je een aantal voorbeelden van dichterbij bekijkt. Ik begin met een intussen bekend zeeptoestel.

WITTE TECHNOLOGIE

In 2017 tweette de Nigeriaanse man Chukwuemeka Afigbo, zelf werkzaam in de techindustrie, een video met daaronder de volgende tekst: ‘Als je ooit een probleem had om het belang van diversiteit in tech en de impact ervan op de maatschappij te begrijpen, bekijk dan deze video.’ Het filmpje toont een toestel dat zeep geeft als je je hand eronder houdt. Eerst verschijnt de hand van een persoon met een witte huidskleur in beeld, waarop vervolgens de zeep wordt gespoten. Maar als Afigba zijn hand onder het toestel houdt, reageert het toestel niet. Houdt hij er ten slotte een wit doekje onder, dan functioneert het zeeptoestel wel opnieuw naar behoren. Wil dat nu zeggen dat de ontwerpers racistische motieven hadden? Uiteraard niet. Maar het filmpje toont wel dat technologieën niet noodzakelijk losstaan van het perspectief of wereldbeeld van de ontwerper, dat ze bijvoorbeeld een witte stempel kunnen dragen. En in dit geval is dat een erg ongemakkelijke waarheid. In de lijn daarvan ligt ook het volgende. Er zijn wellicht uitzonderingen, maar vaak hebben robots een witte kleur. Dat is zo bij Ash in de film *Alien* uit 1979 van Ridley Scott, maar ook bij Sophia, de robot gemaakt door het in Hong Kong gevestigde bedrijf Hanson Robotics, en die intussen ook opdook in het overigens geweldige programma *The Tonight Show with Jimmy Fallon*. Dat robots door de bank genomen wit zijn, is geen toeval. Het heeft te maken met het feit dat het leeuwendeel wordt ontworpen door witte ingenieurs, computerwetenschappers en filmmakers. Het is met andere woorden tijd om ook technologie te dekoloniseren.[22]

Een ander argument tegen de brede invulling van de neutraliteitsthese is afkomstig uit het domein van de *disability studies*. Er zijn slimme assistenten die het spreken van mensen met amyotrofische laterale sclerose (ALS) niet herkennen; sommige AIsystemen voor gezichtsherkenning werken niet bij mensen met een beperking. Een ander voorbeeld komt van het bedrijf HireVue, dat slimme systemen verkoopt aan bedrijven om zo de beste kandidaten voor een job te selecteren. Er vindt eerst een interview met een kandidaat plaats en vervolgens analyseert AI het gesprek, rekening houdend met onder meer de toon van de stem, de gezichtsbewegingen en de patronen in de spraak. Op basis van de resultaten van die analyse schuift de technologie deze of gene kandidaat naar voren. Hoewel zulk systeem goed is in termen van efficiëntie, is het minder goed op ethisch vlak. In een rapport waarin de technologie van HireVue wordt onderzocht, staat het volgende: 'De methode discrimineert op grote schaal veel mensen met een handicap die gezichtsuitdrukking en stem aanzienlijk beïnvloeden: handicaps zoals doofheid, blindheid, spraakstoornissen, en het overleven van een beroerte.' Ondanks het feit dat hier duidelijk sprake is van *ableism* wil dat niet per se zeggen dat de ontwerpers slechte bedoelingen hebben. Wel kun je op basis van de gegeven voorbeelden besluiten dat sommige technologieën zijn ontworpen vanuit een welbepaald referentiekader – dat van mensen zonder al te grote afwijkingen of beperkingen – en dat die technologieën dus niet losstaan van zulk kader.[23]

In de inleiding van het boek wees ik er al op dat de meeste smartphones zijn gemaakt met de gemiddelde lengte van een mannenhand voor ogen. In het verlengde daarvan ligt ook de software voor stemherkenning die wordt gebruikt voor telefoons. Mijn eigen Samsung is niet uitgerust met zulke technologie, maar mocht dat wel zo zijn, dan is de kans reëel dat die beter zou werken voor het herkennen van de stem van een man dan die van een vrouw. In zekere zin is dat verrassend, want we weten uit onderzoek dat het makkelijker is om een vrouw te verstaan dan een man. Aan de andere kant is het niet erg verrassend. De technologie wordt ontworpen aan

de hand van datasets met opgenomen stemmen. En wat blijkt? Het gros van de stemmen komt van mannen. Of kijk naar sommige digitale assistenten: Siri en Alexa zijn twee meisjesnamen uit respectievelijk Scandinavië en Griekenland. Ook op basis van deze voorbeelden kun je dus zeggen dat technologie kan voortvloeien uit een mannelijke blik op de wereld, en dus dat aan die technologie een bepaalde in casu masculiene kijk kan kleven.[24]

Sommige producenten van scheermesjes maken het volgende onderscheid tussen de ontwerpen bedoeld voor mannen en de mesjes die zijn gemaakt voor vrouwen. De scheermesjes voor vrouwen kunnen niet worden opengemaakt, wat met zich meebrengt dat ze niet hersteld kunnen worden mocht dat nodig zijn. De mesjes voor mannen kun je daarentegen wel openen, bevatten een bijsluiter en mogelijk ook borsteltjes om ze te poetsen. Dat onderscheid toont dat ook een vrij basale technologie niet per se losstaat van iets als een denkbeeld, in dit geval een erg traditionele opvatting. Het onderscheid tussen de scheermesjes is een belichaming van het denkbeeld dat vrouwen niet in staat zijn om scheermesjes eigenhandig te herstellen, dat ze daar geen interesse voor hebben of dat het repareren van technologie niet iets is dat door vrouwen moet worden gedaan. Dingen fiksen, dat is iets voor mannen, en overigens niet voor om het even welke man, het is iets voor échte mannen. Dat is althans het oubollige wereldbeeld dat in sommige scheermesjes besloten ligt en dat je dus als argument kunt gebruiken tegen zij die menen dat alle technologie neutraal is, dat aan geen enkele technologie niet-technologische zaken als seksistische opvattingen vasthangen.

HET HORLOGE VAN MIJN GROOTVADER

In dit hoofdstuk zal ik me dus niet op de brede maar op de smalle interpretatie van de neutraliteitsthese richten. Ik vermeldde al dat die wil zeggen dat technologie losstaat van waarden. Maar wat houdt dat dan precies in? Het probleem is nu niet zozeer wat we onder

‘technologie’ moeten verstaan, de moeilijkheid heeft eerder te maken met ‘waarde’. Want hoewel de betekenis van die term op het eerste gezicht min of meer duidelijk is, blijkt bij nader inzien toch dat het niet glashelder is wat daaronder moet worden verstaan. Naar wat werd bijvoorbeeld verwezen toen in het debat over de corona-app in de lente van 2020 werd gezegd dat ook waarden in het geding zijn, een waarde als privacy of *fairness* bijvoorbeeld? Laten we daarom de contouren van het begrip ‘waarde’ schetsen.[25]

Wanneer je wilt weten wat de term ‘waarde’ betekent, dan is het aangewezen voor ogen te houden dat dit woord, net zoals vele andere woorden, meerdere betekenissen heeft. De term ‘waarde’ gebruiken we in het dagelijks leven doorgaans op twee manieren. Je kunt zeggen dat iets een waarde *is* of dat iets waarde *heeft*. Zie hier enkele voorbeelden. Aan de ene kant heb je zinnen als ‘Het AIsysteem voor rekrutering heeft waarde’ en ‘Het horloge van mijn overleden grootvader heeft waarde voor mij’; aan de andere kant kan worden gezegd dat psychiaters een waarde als integriteit belangrijk vinden, of dat autonomie sinds een aantal decennia centraal staat in de gezondheidszorg. Hier zie je onmiddellijk dat ‘waarde’ in de eerste twee zinnen een ietwat andere invulling krijgt dan daarna. Maar wat precies is het verschil?

Kijk naar de zinnen over het AI-systeem en het horloge van de grootouder. Daar wordt beweerd dat die technologieën waarde *hebben*. Wanneer het woord ‘waarde’ in deze zin wordt gebruikt, dan bedoelt men daar een van deze twee zaken mee: ofwel dat iets instrumentele waarde heeft, ofwel dat iets niet-instrumentele waarde heeft, en dus op zich waarde heeft. Als je zegt dat iets instrumentele waarde heeft, betekent dat dat het een middel of instrument is dat een doel dient, dat het een rol vervult in de realisering van een doel. Dat kan over tal van zaken worden gezegd, artefacten en daden bijvoorbeeld. Een hamer en telefoon hebben zulke waarde, omdat je ermee respectievelijk een spijker in de muur kunt slaan en kunt bellen; studeren heeft waarde, omdat het je helpt te slagen voor een examen.

Daarnaast kun je instrumentele waarde toeschrijven aan niet-artificiële zaken. Ik kan een boomstam gebruiken om een stromende rivier af te varen of een steen om vuur te maken.

Je kunt echter van iets ook zeggen dat het waarde heeft, terwijl die waarde geen instrumenteel karakter heeft – zoals in het geval van het horloge van mijn overleden grootvader. Dat wil zeggen: iets kan ook waarde hebben, los van het nut; dingen kunnen op zich waarde hebben. Hoewel nut voldoende kan zijn voor het hebben van waarde, is nut dus niet noodzakelijk. Nutteloze dingen kunnen ook waarde hebben. Wanneer je vindt dat iets zulke niet-instrumentele waarde heeft, dan druk je met zo'n uitspraak normaal gezien een positief oordeel uit. Je geeft dan te kennen dat je positief staat tegenover het ding dat zo'n waarde heeft. Zulk oordeel kan op tal van zaken gebaseerd zijn. Je kunt vinden dat iets mooi is en daarom in esthetische zin op zich waarde heeft. Het horloge van je grootvader kan je dan weer herinneren aan de goede band die je met hem had, waardoor het voorwerp in emotionele zin waarde heeft. Maar wat precies ook de bron van de waarde is, als je zegt dat iets op zich waarde heeft, dan is dat vanzelfsprekend ook een reden voor een positieve omgang met wat die waarde heeft. Omdat het horloge van mijn grootvader waarde heeft, is dat een reden om ervoor te zorgen, om het te beschermen of er met liefde over te spreken.

MOREEL EN NIET MOREEL

Maar er is dus nog een tweede betekenis van 'waarde', die aan het licht komt in de andere zinnen van daarnet, de zinnen waarin het gaat over een waarde als autonomie of respect. In dat geval gaat het niet over iets dat een waarde *heeft*, maar iets dat een waarde *is*. We bedoelen normaal gezien niet zozeer dat autonomie waarde heeft, maar wel dat het een waarde is. Indien je 'waarde' in deze zin gebruikt, dan druk je daarmee geen oordeel uit (zoals bij 'waarde hebben'), maar verwijs je naar een bepaalde toestand. Over 'vrijheid' zeggen we bijvoorbeeld dat het een waarde is. Je verwijst daarmee naar een

welbepaalde stand van zaken, en meer in het bijzonder refereer je dan wellicht ofwel naar de toestand waarin je als persoon of groep niet wordt gedwongen of beperkt door anderen, ofwel naar de toestand waarin je de mogelijkheid hebt om te kiezen uit meerdere opties. Het is nuttig om deze zaken nog verder uit te diepen.

Er zijn tal van waarden die in de lijn van vrijheid liggen. Voorbeelden zijn geluk en duurzaamheid, maar ook loyaliteit en respect. Die laatste twee hebben met vrijheid, geluk en duurzaamheid gemeen dat ze een moreel karakter hebben. Al deze waarden gaan over de gewenste verhouding tussen mensen onderling, tussen mensen en niet-menselijke dieren, tussen mens en natuur, en ook tussen mens en technologie. Toch zijn er ook waarden die een niet-moreel karakter hebben. Denk aan de kunstwereld, waar vaak, maar niet per se, een waarde als schoonheid of elegantie centraal staat. Of neem de ingenieurswereld, waar een waarde als efficiëntie cruciaal is. Schoonheid en efficiëntie komen dus overeen met duurzaamheid omdat het ook waarden zijn, maar verschillen er wel van omdat ze geen morele inhoud hebben. Beide zijn voorbeelden van waarden met een moreel neutraal karakter.

De overeenkomst tussen pakweg schoonheid en duurzaamheid moeten we nog verfijnen. Waarden zijn immers niet louter toestanden in de werkelijkheid. Het zijn standen van zaken die we ook belangrijk vinden. Het is tegenstrijdig om enerzijds van iets te zeggen dat het een waarde is en anderzijds te menen dat het er niet toe doet. Sterker nog, met ‘waarde’ verwijzen we altijd naar een toestand waarvan we vinden dat die zó belangrijk is dat die er *moet* zijn of dat je er rekening mee *moet* houden. Als je bijvoorbeeld zegt dat autonomie in de gezondheidszorg een waarde is, dan bedoel je niet dat mensen er zelf feitelijk over hun leven beschikken. Noch wil je daarmee uitsluitend zeggen dat het belangrijk is dat mensen zelf vorm mogen geven aan hun leven. Nee, wie zegt dat autonomie in de gezondheidssector een waarde is, doet meer dan beschrijven of het belang ervan aanstippen. Men wil in dat geval zeggen dat er zelfbeschikking *moet* zijn.

Een nog preciezere omschrijving van waarden is dat ze vaak van niet-instrumentele aard zijn, dat ze met andere woorden verwijzen naar zaken die we op zich belangrijk vinden. Dat betekent dat men naar pakweg schoonheid en autonomie streeft, niet omdat men daarmee nog iets anders beoogt, maar omwille van die waarden zelf. Waarden zijn dan een doel op zich, en geen middel om een ander doel te bereiken. Wellicht het bekendste voorbeeld daarvan is geluk. Dat is zonder twijfel een waarde, maar bovendien een waarde die een doel op zich is. Je zegt normaal gesproken niet dat je gelukkig wilt worden, om daarmee dit of dat te bereiken. Nee, je wilt gelukkig worden. Punt.

Wil dat nu zeggen dat waarden per se *uitsluitend* van nietinstrumentele aard zijn? Nee. Hoewel het eigenaardig zou zijn om te beweren dat je gelukkig wilt worden om daarmee iets anders te bereiken, kunnen waarden naast een niet-instrumenteel ook een instrumenteel karakter hebben. Ik bedoel daarmee dat we sommige waarden naar voren schuiven met een ander doel voor ogen. Dat klinkt op het eerste gezicht misschien wat vreemd, maar toch stemt het overeen met de rol die waarden in het alledaagse leven spelen. Neem privacy. Uiteraard vinden we het op zich verwerpelijk als onze persoonsgegevens zonder onze toestemming te grabbel worden gegooid, los van de gevolgen. Maar de bescherming van de privacy via onder meer de *General Data Protection Regulation* (GDPR) dient ook om de burger te beschermen tegen misbruik door bedrijven en overheden. Lees: privacy is ook een instrument om andere doelen te realiseren.

Tot slot zijn waarden niet per definitie universeel, in de zin dat ze niet door iedereen worden gedeeld en dat niet iedereen ze in de dezelfde mate belangrijk vindt. Neem de discussie over het gebruik van een app om de verspreiding van SARS-CoV-2 tegen te gaan. In Europa ging een flink deel van het debat over het risico dat de app de privacy van de gebruiker niet zou respecteren. In de Aziatische wereld speelden zulke zaken een flink stuk minder. De verschillen kunnen zeer groot zijn. Wat jij bijvoorbeeld als een waarde beschouwt, kan door mij namelijk als volstrekt onwenselijk worden gezien. Ik vind

inclusiviteit bijvoorbeeld belangrijk, en ik neem aan dat dat voor veel mensen zo is, alleen vinden sommigen dat jammer genoeg verwerpelijk. Zij schuiven niet inclusiviteit maar het tegendeel, namelijk segregatie, naar voren als de toestand die we moeten proberen te realiseren. Segregatie is voor hen een waarde, wat ik dan weer – en vele anderen met mij – afkeurenswaardig en weerzinwekkend vind.

Om af te ronden zet ik het belangrijkste op een rijtje. De term ‘waarde’ kun je op twee manieren gebruiken. We kunnen zeggen dat iets, duurzaamheid bijvoorbeeld, een waarde *is*, en we kunnen zeggen dat iets, een horloge of een huis, waarde *heeft*. Telkens kunnen we een instrumentele en niet-instrumentele invulling onderscheiden. Een waarde als geluk is op zich belangrijk, terwijl een waarde als privacy ook andere doelen dient, financiële veiligheid bijvoorbeeld. En wat betreft ‘waarde hebben’: een horloge helpt je de tijd niet uit het oog te verliezen, en kan ook op zich waarde hebben, omdat het je herinnert aan je grootouder.

Technologie en ethiek

Zoals eerder gezegd is technologie – volgens de smalle interpretatie van de neutraliteitsthese – waardeneutraal. Dat is althans wat nogal wat mensen beweren die nauw betrokken zijn bij het maken en verspreiden van technologie en AI. Twee termen staan dus centraal: technologie en waarde. Onder ‘technologie’ verstaan we globaal genomen twee zaken. Meestal wordt met die term verwezen naar al dan niet materiële artefacten met een functie (zoals een laptop), en soms ook naar een manipulatieproces (zoals in ‘biotechnologie’) – ik heb het doorgaans over het eerste. Dat weten we al uit de inleiding. Op basis van de voorbije paragrafen zijn we nu ook in staat om te achterhalen wat we in deze context precies onder ‘waarde’ moeten verstaan. De neutraliteitsthese, zo laat ik dadelijk zien, stelt dat

technologie op zich losstaat van morele waarden als privacy en duurzaamheid.

WENEN OM EEN ROBOT

Tussen oktober 2017 en april 2018 zoomde de tentoonstelling ‘Hello Robot’ in het Design Museum in de Belgische stad Gent in op de verhouding tussen mens en technologie. De bezoeker kon een blik werpen op een scène uit de film *Her* uit 2013 waarin een man (Joaquin Phoenix) verliefd wordt op een besturingssysteem, en wie opgroeide in de jaren 1990 werd terug naar haar of zijn jeugdjaren gekatapulteerd door de Tamagotchi: een virtueel diertje in de vorm van een ei. Op de tentoonstelling kon je ook in een glazen kast een robot zien. Hoewel die visueel niet erg aantrekkelijk was (de robot leek eerder op een ouderwetse stofzuiger met handen en voeten), was de robot in de film wel de publiekstrekker. De reden was dat het ging over R2-D2, de droid die telkens terugkeert in de *Star Wars*films. Ook het feit dat de robot alleen mocht worden in- en uitgepakt door een vertrouwelinge van *Star Wars*bedenker Georges Lucas had daarmee te maken.

Fallujah, Irak, november 2005. Soldaten van het Amerikaanse leger organiseren een begrafenis. In de kist worden twee eremedailles gelegd, de zogenoemde Purple Heart en Bronze Star Medal, gevolgd door 21 geweerschoten. Dat allemaal om het slachtoffer te bedanken voor de vele levens die hij op het terrein heeft gered. Het ging dus over een ‘hij’, genaamd Boomer. Het slachtoffer was echter geen mens of niet-menselijk dier, maar een robot. Het was een Marcbot, een machine die naar gebieden wordt gestuurd om explosieven te ontmantelen. Hier was hij echter vernietigd door vijandige troepen. De robot leek overigens niet op een mens of pakweg een hond, zoals de humanoïde robot Atlas van het Amerikaanse roboticabedrijf Boston Dynamics. Boomer, in kaki kleur, leek eerder op een kleine tank op vier wielen. Maar dat betekende niet dat het artefact de soldaten koud liet. Hoe eigenaardig dat ook lijkt, het ligt in de lijn van wat iedereen

weet: mensen hechten zich niet alleen aan elkaar, dieren die geen mensen zijn of plaatsen, maar ook aan dingen.[26]

Deze voorbeelden brengen in herinnering dat technologieën waarde op zich kunnen hebben, dat je aan technologie waarde in nietinstrumentele zin kunt toeschrijven. Boomer werd vernietigd en heeft dus geen nut meer, hoewel hij niet waardeloos is. De soldaten hebben er zich aan gehecht, en schrijven daarom emotionele waarde toe aan de technologie. R2-D2 wordt niet meer gebruikt op de set, hoewel dat wellicht wel nog zou kunnen. De vaststelling echter dat bezoekers zoveel belangstelling voor de robot toonden, geeft aan dat de droid waarde op zich heeft, zij het dan op symbolisch vlak. Men schrijft R2-D2 zulke waarde toe, omdat hij meespeelde in de *Star Wars*-films van Lucas en daadwerkelijk op de filmset aanwezig was.

Mocht de neutraliteitsthese betekenen dat technologie geen waarde heeft die losstaat van nut en dienstbaarheid, dan zouden de net gegeven voorbeelden de neutraliteitsthese onderuithalen. Niemand echter, ook niet zij die de neutraliteitsthese verdedigen, beweert dat technologie op zich geen waarde kan hebben. Louter het feit dat technologie doorgaans materieel is, volstaat om te beseffen dat technologie symbolisch of emotioneel geladen kan zijn. Door dat tastbare karakter kun je het erven van iemand die jou dierbaar is of kun je je eraan hechten in emotionele zin. Wie beweert dat technologie waardevrij is, meent dus niet dat technologie geen waarde op zich kan hebben, en kun je dus ook niet van antwoord dienen door te verwijzen naar zaken zoals R2-D2 of Boomer.

Nu mag je op basis daarvan niet concluderen dat de neutraliteitsthese over die andere waarde gaat, over instrumentele waarde. De reden is dat er een onlosmakelijk verband bestaat tussen zulke soort waarde en technologie. De link wordt gelegd door de eigenschap die per definitie tot technologie behoort: functionaliteit. Alle technologie, zo zagen we eerder al, heeft noodzakelijkerwijs een functie, in de zin dat het een effect moet realiseren, dat ze is gemaakt met een welbepaald doel voor ogen. Anders geformuleerd: alle artefacten moeten

doelgericht zijn, opdat je ze als technologie kunt zien. Als iets door mensenhanden is gemaakt maar geen functie heeft, kun je niet zeggen dat het technologie is. Wanneer we nu aannemen dat alle technologie naar behoren werkt, dan maakt die herformulering duidelijk dat technologie in alle gevallen instrumentele waarde heeft. Als je namelijk van een artefact niet kunt zeggen dat het technologie is als het geen doel dient, dan volgt daaruit dat alle technologie noodzakelijkerwijs instrumentele waarde heeft. Kortom, het is onmogelijk dat een technologie doet waarvoor het is ontworpen en toch geen instrumentele waarde heeft. Die onvermijdelijke band tussen technologie en instrumentele waarde is de reden waarom niemand de volgende invulling aan de neutraliteitsthese geeft: technologie is vrij van instrumentele waarde.

ZONDER MORELE WAARDE

We kunnen nu stilaan beginnen zien wat de neutraliteitsthese wél betekent. Als het voor zich spreekt dat technologie waarde op zich kan hebben en dat het noodzakelijk instrumentele waarde heeft, dan moet het wel zo zijn dat de neutraliteitsthese alles te maken heeft met bijvoorbeeld privacy en *fairness*, zaken waarover we niet zozeer zeggen dat ze waarde *hebben*, maar dat ze waarden *zijn*. Om daar een goed beeld van te krijgen, sta ik nog even stil bij de instrumentele waarde van technologie.

Technologie kun je voor verschillende soorten doelen gebruiken. Twee voor de hand liggende doelen zijn artefacten en vermogens op ofwel lichamelijk ofwel mentaal vlak. Een boormachine kun je gebruiken om een kast in elkaar te zetten, een exoskelet kun je in zware werksituaties gebruiken, waardoor je lichaam minder wordt belast. Daarnaast kunnen ook waarden doelen zijn – hoewel niet alle doelen waarden zijn. Je kunt er met behulp van technologie naar streven om een toestand te realiseren waarvan je vindt dat die zo belangrijk is dat die *moet* worden gerealiseerd. Natuurlijk is het leven geen ponykamp, en natuurlijk ben ik geen naïeve techno-optimist,

maar het aantal technologieën dat bijdraagt aan de realisering van waarden is schier eindeloos. Verkeerslichten leiden tot meer veiligheid, zonnepanelen hebben een positief effect op duurzaamheid, rolstoelen maken de samenleving inclusiever, en medische technologie helpt bij het voorkomen, opsporen en genezen van ziekten. Of neem het gebruik van algoritmen in de rechtspraak. Hoewel ik eerder heb vermeld dat dit niet zonder problemen is, kan het toch ook zorgen voor een gelijkere behandeling van zaken. Laat je juridische oordelen uitsluitend aan rechters over, dan is de kans reëel dat twee identieke situaties verschillend worden beoordeeld, afhankelijk van bijvoorbeeld het uur van de dag waarop het vonnis wordt geveld. Het tijdstip kan namelijk een rol spelen. Onderzoek wijst uit dat rechters strenger oordelen als ze honger hebben, net voor ze gaan eten. Algoritmen daarentegen zijn althans in dit opzicht een goede zaak dat ze geen last hebben van zaken als honger of vermoeidheid.

Ik heb die voorbeelden gekozen om zo uiteindelijk de neutraliteitsthese scherper in beeld te krijgen. Die these gaat namelijk niet over hoe het gebruik van technologie leidt tot pakweg een rechtvaardigere wereld, ze heeft te maken met technologie op zich, los van en voorafgaand aan het gebruik. Omdat dit de kern van de neutraliteitsthese is, lijkt het aangewezen om dat nog ietwat toe te lichten.

In de voorbeelden van daarnet wordt technologie aan een waarde als privacy of *fairness* gekoppeld. Het is nu van belang om te weten dat die koppeling in een welbepaalde context tot stand komt. De AIsystemen, rolstoelen en medische apparatuur waarover ik het net had, worden met waarden verbonden wanneer ze worden *gebruikt*. De koppeling vindt plaats wanneer men de technologie in de hand neemt, ermee aan de slag gaat en inzet voor een doel. Welnu, de neutraliteitsthese gaat ook wel over waarden als privacy en *fairness* – dat wisten we al –, maar niet over de gebruiksfase. Ze gaat over technologie op zich, dat wil zeggen: los van het gebruik. Meer in het bijzonder luidt de stelling dat aan technologie, wanneer je die

beschouwt zonder naar de gevolgen van het gebruik te kijken, geen waarden vasthangen of kleven. Waarden kunnen wel in goede en slechte zin worden beïnvloed door technologie, maar omgekeerd staat technologie zelf los van waarden. Technologie is neutraal: ze is niet gebonden aan deze of gene waarde, ze is niet voor of tegen een of andere waarde. Als je technologie met andere woorden gebruikt zoals het moet, dan houdt dat niet noodzakelijk in dat ze een waarde realiseert, uitdrukt, onderstreept of bestendigt. Natuurlijk, het is goed mogelijk dat je een technologie gebruikt op een manier die een waarde verwezenlijkt, maar dat volgt niet uit de bouw van de technologie, uit de bedoeling van de ontwerper. De waarde zit niet gebakken in de technologie, de technologie is niet innig vervlochten met deze of gene waarde. Zie hier de uitleg van de populaire neutraliteitsthese, althans van de smalle interpretatie.

Omdat die stelling het centrale thema van dit hoofdstuk is, is het belangrijk dat er geen misverstanden zijn. Laat ik daarom nog een drietal punten aanstippen. Ten eerste is de bewering niet louter dat alle technologie waardevrij is. Dat zou immers nog kunnen betekenen dat het in theorie wel mogelijk is dat technologie waardegeladen is, maar dat dat de facto niet zo is. De claim is sterker. De neutraliteitsthese luidt dat technologie niet waardegeladen is, omdat ze niet waardegeladen *kan* zijn. Technologie is met andere woorden per definitie waardevrij, aldus iemand als Chomsky of Pitt.

Ten tweede is het van belang te wijzen op de termen die vaak worden gebruikt om de neutraliteitsthese uit te leggen: ‘kleven’, ‘geladen’, ‘vasthangen’. Het spreekt voor zich dat die termen hier uitsluitend in figuurlijke zin worden gebruikt. Niemand, ook niet degene die de neutraliteitsthese verwerpt, beweert dat waarden letterlijk *in* de technologie zitten en er ook daadwerkelijk aan vasthangen. Met de term ‘waarde’ verwijzen we immers naar een stand van zaken in de werkelijkheid buiten de technologie. Rechtvaardigheid en respect bijvoorbeeld hebben te maken met de relatie tussen personen, net zoals duurzaamheid refereert aan onder meer de relatie tussen organismen; die waarden zitten niet in de technologie en staan er los

van. Waarnaar verwijst de term ‘waardegeladenheid’ dan wel? De figuurlijke invulling van de term luidt dat het ontwerp van een technologie refereert aan een waarde, in die mate dat je de technologie niet ten volle kunt begrijpen zonder naar die waarde te verwijzen. In dat geval ‘plakt’ de waarde aan de technologie en is zij dus als het ware met waarde geladen. Volgens de neutraliteitsthese is dat bij geen enkele technologie het geval. Straks onderzoek ik of dat klopt.

Ten derde moet je in gedachten houden dat de neutraliteitsthese niet over om het even welke waarde gaat. In zekere zin ligt dat voor de hand, want wie zou nu beweren dat er geen waarde als accuraatheid of efficiëntie aan technologie kan vasthangen? De smalle invulling van de bewering dat technologie neutraal is, gaat over een welbepaalde soort waarde, namelijk waarden met een moreel karakter, waarden die niet moreel neutraal zijn. Wie dus meent dat technologie niet met waarde is geladen, bedoelt daarmee dat aan technologie geen waarden als privacy, duurzaamheid of *fairness* vastzitten. Kortom, de neutraliteitsthese verwijst naar waarden die te maken hebben met het goede leven, en niet over pakweg wetenschappelijke of technologische waarden. Er zit geen ethiek in de dingen, dat is het punt.

KAPITALISME EN RELIGIE

We zijn bijna toe aan een meer evaluatieve blik op de neutraliteitsthese. Klopt de eeuwenoude en wijdverspreide bewering dat technologie neutraal is? Wat zijn de argumenten voor en tegen? Het voorgaande leert ons dat er zeker twee zaken zijn die we niet kunnen gebruiken als argument tegen die these: ten eerste theorieën over de oorzakelijke band tussen technologie en zaken die geen technologie zijn; ten tweede voorbeelden van technologieën die ‘gekleurd’ zijn in de brede zin van het woord. Het is nuttig om daar nu nog kort even de aandacht op te vestigen.

In zijn studie *Platform Capitalism* uit 2020 schetst Nick Srnicek de volgende band tussen kapitalisme en technologie.[27] In een prekapitalistische samenleving was er geen grote afstand tussen mensen en de middelen om in levensonderhoud te voorzien. Mensen zorgden zelf voor hun eigen voedsel en onderhoud. Dat veranderde in het kapitalisme. Wij, mensen, zijn niet meer de producent van ons eigen voedsel en zijn nu afhankelijk van anderen die de dingen maken om in leven te blijven. We moeten ons op de markt begeven om geld te verdienen, geld dat ons in staat stelt om de levensnoodzakelijke dingen te kopen die door anderen geproduceerd worden. Het probleem echter is dat er op de markt veel mensen zijn die precies dezelfde diensten of producten leveren. Er is dus stevige concurrentie waardoor je als speler het onderspit dreigt te delven ten aanzien van anderen die goedkopere producten of diensten te bieden hebben. Die strijd drijft de marktspelers ertoe om allerlei methoden te bedenken om de arbeidskosten naar beneden te halen en om de productiviteit de hoogte in te duwen. Kinderarbeid, onderbetaling, de verplaatsing van het productieproces naar lageloonlanden en nachtwerk moet je vanuit die optiek begrijpen. Maar, en dat is nu het punt, hetzelfde geldt ook voor technologie. Denk aan de lopende band, voor de Eerste Wereldoorlog nog geïntroduceerd door Henry Ford. Dat is een technologie die werd ingevoerd om de productiecapaciteit op te drijven. Of neem planningssoftware, die ook nog in het derde hoofdstuk terugkeert. Vandaag de dag worden in koffiebars en restaurants de weekroosters voor het personeel vaak niet meer door mensen maar door AI-systemen samengesteld. Die technologie analyseert eerst het komen en gaan van klanten in het verleden en de daarbij horende inkomsten. Op basis daarvan wordt ingeschat welke momenten in de toekomst het drukst en welke momenten het rustigst zullen zijn. Het doel is een efficiënte inzet van het personeel. Er moeten werkkrachten zijn wanneer het echt nodig is, er mag niet te veel personeel zijn wanneer het onnodig is. Kortom, ook technologie is het resultaat van een kapitalistische logica – sommige technologie althans, want niet alle technologie is ingebed in een systeem dat gericht is op steeds meer winst.

Een andere theorie gaat niet over de oorzaak van technologie, bijvoorbeeld winsthonger, maar over het effect van technologie, met name op religie. Die theorie gaat als volgt. Door het gebruik van technologie kunnen mensen beter en sneller armoede, honger en ziekte opsporen, voorkomen en overwinnen. Dat heeft ertoe geleid dat de behoefte is afgenomen om zich te beroepen op bovennatuurlijke krachten, religieuze voorwerpen en heilige plaatsen. Technologie leidt dus niet alleen tot onder meer ingrijpende veranderingen op de arbeidsmarkt, maar ook, zo luidt de redenering, tot de desacralisering van de samenleving.

Ik wil hier nu niet uitvlooien of die verklaringen allemaal even juist of plausibel zijn. Ik neem aan dat ze kloppen en vermeld ze enkel omdat ik de aandacht wil vestigen op het verschil tussen dit soort van theorieën en de neutraliteitsthese. Terwijl de neutraliteitsthese over technologie zelf gaat, is dat niet het geval bij de geschetste theorieën. Die gaan over de oorzakelijke band tussen technologie en iets anders, hier zijn dat kapitalisme en religie. Ze focussen niet op de technologie, maar op de oorzaak (kapitalisme) of het gevolg
(desacralisering) van technologie. De verdediging van de stelling dat technologie waardeneutraal is, sluit dus niet uit, althans in principe niet, dat je ook beweert dat technologie een product van het kapitalisme is of dat meer technologie tot minder religie leidt. De aard van beide uitspraken is namelijk verschillend. Zeker, wie oppert dat technologie een product van het kapitalisme is of de kracht van religie inbindt, meent wellicht ook dat dat niet zonder gevolgen is voor de opvatting van technologie. En ja, het kan dat de oorzakelijke band tussen technologie en bijvoorbeeld kapitalisme een effect heeft op de vraag of technologie moreel neutraal is. Maar omdat het over twee andersoortige beweringen gaat, is de bewering dat technologie een kind van het kapitalisme of de motor achter het verval van religie is op zich geen argument tegen de neutraliteitsthese. Dat is zo voor deze twee beweringen, maar ook voor andere theorieën die niet gaan over kapitalisme en religie, maar wel over een oorzaakgevolgrelatie waar technologie deel van is.

Daarnaast wil ik nogmaals benadrukken dat mijn verhaal niet gaat over de brede interpretatie van de neutraliteitsthese, die op onder meer perspectieven en ideologieën focust. Ik concentreer me op de smalle invulling, over technologie en morele waarde. Ik kan dus geen gebruik maken van de volgende casus om die interpretatie te bekritiseren.[28] Hoewel enkele jaren terug door SEAT een auto voor vrouwen werd ontworpen, worden sinds lang de meeste personenauto's ontworpen met de lichaamsbouw van een man voor ogen. Dat lijkt op het eerste gezicht misschien redelijk onschuldig, maar toch is dat niet zonder gevolgen. Omdat bijvoorbeeld de zitplaatsen zijn gemaakt met het gemiddelde gewicht van mannen voor ogen en omdat vrouwen doorgaans minder wegen, bieden de stoelen vaak niet genoeg bescherming voor vrouwen. Gemiddeld genomen hebben vrouwen ook meer kans op verwondingen en een whiplash bij respectievelijk een frontale botsing en een kopstaartbotsing. En dat heeft alles of zo goed als alles te maken met het feit dat auto's zijn ontworpen vanuit het perspectief van de man.

Hoewel er geen redenen zijn om te vermoeden dat de ontwerpers van auto's seksistische bedoelingen hebben, zijn deze technologieën 'gekleurd', niet neutraal. Ze zijn ontworpen vanuit een welbepaald referentiekader: dat van mannen. Toch kun je deze casus niet gebruiken als argument tegen de neutraliteitsthese, althans niet tegen de smalle interpretatie ervan. Die luidt namelijk dat technologie niet met morele waarden is geladen, terwijl de gegeven casus niet over morele waarden gaat maar over een perspectief van waaruit technologie is ontwikkeld. Uiteraard, aan de auto's 'kleeft' een referentiekader (de mannelijke blik), maar kaders zijn geen morele waarden. Personenauto's kunnen daarom geen reden zijn om de neutraliteitsthese te verwerpen, net zoals de tweet van de Nigeriaanse man Afigbo over de 'witte' zeepdispenser geen tegenargument is, en net zoals je het feit dat scharen meestal zijn ontworpen voor rechtshandigen niet kunt gebruiken om de smalle betekenis van de neutraliteitsthese aan te vallen.

Technologie is neutraal

Het is tijd om van perspectief te veranderen. Ik ruil een beschrijvende blik in voor een meer evaluatieve benadering. Dat technologie neutraal is, is naast de bewering dat AI disruptieve technologie is wel een veelgehoorde stelling, maar klopt ze ook? Zijn er redenen om daarmee akkoord te gaan? Op de volgende bladzijden geef ik twee bekende argumenten voor de neutraliteitsthese: het zichtbaarheidsargument en dubbelegebruiksargument. Ik begin echter met te wijzen op iets herkenbaars dat geen technologie is en ook niet waardegeladen is. Dat is geen argument voor de neutraliteitsthese, maar het laat wel het volgende zien. Als het klopt dat technologie waardeneutraal is, dan ligt technologie in het verlengde van andere zaken die niets met technologie te maken hebben; als blijkt dat de neutraliteitsthese onjuist is, dan is die bewering toch ook niet geheel bij de haren getrokken, aangezien er andere zaken dan technologie zijn waaraan geen morele waarden kleven.

KIJKEN NAAR *TEMPTATION ISLAND*

Veel dingen die we doen, zijn waardegeladen, al dan niet in morele zin. Ouders staan in voor het welzijn van hun kinderen, nietgouvernementele organisaties streven naar een rechtvaardigere samenleving, en de handelingen van artsen zijn afgestemd op de waarde ‘gezondheid’. Maar hangen aan handelingen ook *noodzakelijk* waarden vast?

Stel, je kijkt naar een aflevering van *Temptation Island* om je te ontspannen. Als ontspanning een waarde voor je is, kun je besluiten dat het kijken naar het televisieprogramma met waarde is geladen. Maar je kunt ook kijken om een andere reden dan om je te ontspannen. Je zit bijvoorbeeld ’s avonds in de sofa op je laptop te werken en je besluit om te kijken, zij het slechts zijdelings. Het kijken naar het programma ontspant je niet, integendeel, je ergert je blauw aan de deelnemers. Toch is de drang om te kijken te groot, en dat

heeft alles te maken met je verlangen naar sensatie. Je kijkt, eenvoudigweg om te zien of er nieuwe perikelen of gebeurtenissen zijn. Geeft Elke Arda nog een kans? Vindt Simone Zach echt een baby? (Ik heb het over het twaalfde seizoen, uitgezonden in 2020.) Hoewel sensatie voor nogal wat mensen erg leuk is, is kijken naar *Temptation Island* voor velen toch niet waardegeladen. Reden? In het begin van dit hoofdstuk hebben we gezien dat we waarden zó belangrijk vinden dat we die *moeten* nastreven en dat we er rekening mee *moeten* houden. Echter, weinigen – en wellicht zelfs niemand – spreken in die termen over sensatie. Wie geeft aan iets als sensatie zo veel gewicht dat het een verplichtend karakter heeft? Heel veel van wat we doen, is gericht op een doel. Maar niet alles, niet elk doel, is een waarde.

Kunnen we niet zeggen dat sensatie alleen het eerste doel is en welzijn het uiteindelijke doel? Misschien kijk ik wel naar *Temptation Island* vanwege de sensatie, maar is sensatie slechts relevant omdat ze bijdraagt aan mijn welzijn. In dat geval zou het kijken naar televisie onrechtstreeks met waarde geladen zijn, want het kijken heeft alleen betrekking op de waarde van welzijn via de sensatie. Het is echter geenszins voor iedereen zo dat kijken in deze zin waardegeladen is. Waarom niet? Het antwoord ligt in de lijn van mijn uitleg van daarnet: aangezien waarden per definitie belangrijk zijn, kan iets alleen bijdragen aan een waarde, als het zelf ook een zeker gewicht heeft. Je kunt niet beweren dat iets goed is voor bijvoorbeeld je welzijn, terwijl het zelf nauwelijks of helemaal niet belangrijk is. Omdat we sensatie door de bank genomen betekenisloos vinden, zeggen we ook niet dat het goed is voor ons welzijn. Let wel, het is niet uitgesloten dat sommigen mijn redenering tegenspreken omdat voor hen sensatie wel een waarde is of bijdraagt aan hun welzijn. Dat neemt echter niet weg dat veel mensen, en wellicht zelfs de overgrote meerderheid, daar anders over denken. Dat op zich volstaat om het kijken naar zoiets als *Temptation Island* in verband te brengen met de neutraliteitsthese.

ALLES VAN WAARDE IS ZICHTBAAR

De vraag die zich nu opdringt, is hoe plausibel de neutraliteitsthese is. Klopt het dat technologie waardeneutraal is? Wellicht het bekendste argument dat wordt ingeroepen om die these te ondersteunen, is het dubbelegebruiksargument. Iets minder bekend is het zogeheten zichtbaarheidsargument. Wat houden die precies in?

Het zichtbaarheidsargument valt uiteen in twee delen. Waardegeladenheid vereist dat een morele waarde in de technologie geïdentificeerd kan worden, zo luidt het eerste deel. Je kunt niet zeggen dat een technologie met een morele waarde is geladen, als je die niet in de technologie kunt zien. Het tweede deel luidt dat je in een technologie geen morele waarde kunt aanwijzen. Je kunt de vorm en het materiaal wel waarnemen, maar het is onmogelijk om in een apparaat een plek aan te wijzen waar zich de morele waarde bevindt. Omdat je morele waarden niet kunt aanwijzen, terwijl die aanwijsbaarheid noodzakelijk is om van een waarde te kunnen spreken, is het besluit dat technologie niet met morele waarde is geladen.[29]

Het dubbelegebruiksargument is het tweede argument waarmee de neutraliteitsthese wordt onderbouwd. Het verwijst onder meer naar de volgende zaken. Wanneer een ziekenhuis een melding ontvangt, vertrekt een snelle auto naar de plaats waar de patiënt zich bevindt. Die auto levert in dat geval een positieve bijdrage aan het welzijn van de zorgbehoevende. Maar aan het gebruik van auto’s kunnen ook slechte intenties ten grondslag liggen. Men gebruikt ze ook om mensen mee te verhandelen of een aanslag mee te plegen. Twee andere voorbeelden: drones en camera’s. Beide technologieën kunnen worden gebruikt om zieke planten op plantages op te sporen, maar kunnen ook worden ingezet voor spionage. Een app ten slotte kun je inzetten met als doel de bescherming van de gezondheid van het volk. Maar je kunt een app ook gebruiken voor *mass surveillance*. In 2019 raakte bijvoorbeeld bekend dat men in de Chinese provincie Xinjiang gebruik maakt van een app (voluit het *Integrated Joint*

Operations Platform) om verschillende miljoenen Oeigoeren en andere Turkse minderheden te controleren en te onderdrukken.[30]

De genoemde voorbeelden gaan over technologieën die in tweeerlei opzichten kunnen worden gebruikt: zowel ten goede als ten kwade. Toegepast op de thematiek die ons hier bezighoudt: het gaat over technologieën die zowel voor als tegen dezelfde morele waarde kunnen worden ingezet. Maar moeten we niet nog een stap verder gaan? Want wat voor drones en apps geldt, gaat dat dan niet op voor alle technologieën? Zelfs wapens – toch artefacten die zijn ontworpen om schade te veroorzaken – kunnen voor zowel een goed (bevrijden van een volk) als een slecht doel (terroriseren van een volk) worden ingezet. Het is inderdaad moeilijk om technologieën te bedenken die men niet voor tegengestelde doelen kan gebruiken, en voor sommigen is dat een argument dat in het voordeel van de neutraliteitsthese pleit. Want als een artefact zowel voor als tegen een waarde kan worden gebruikt, moeten we dan niet concluderen dat het op zich, los van de goede en slechte effecten, waardeneutraal is?

Die conclusie lijkt op het eerste gezicht gerechtvaardigd. Dat komt omdat twee zaken onverzoenbaar lijken. Stel dat je de nationale verkiezingen wilt manipuleren. Daarvoor moet je een goed zicht hebben op het kiespubliek, heb je veel informatie nodig over alle stemgerechtigden. Je ontwikkelt een app die je uitrolt op sociale media, waardoor je de gegevens van talloze gebruikers kunt bemachtigen, zonder dat die gebruikers daarvan op de hoogte zijn. Ik verwijs hier uiteraard naar Cambridge Analytica, waar duidelijk sprake was van schending van de privacy.[31] Als je nu tegelijkertijd zou beweren dat diezelfde technologie niet waardeneutraal is, bijvoorbeeld omdat daaraan een waarde als privacy vasthangt, dan heeft het er alle schijn van dat dit op gespannen voet staat met de schending van de privacy door het gebruik van die technologie. Het lijkt tegenstrijdig dat in een technologie een waarde zit verankerd, terwijl men diezelfde technologie gebruikt tegen die waarde. Omgekeerd: je lijkt jezelf tegen te spreken als je weet dat men een

technologie gebruikt met het oog op meer gelijkheid, maar je tegelijk beweert dat het artefact is ontworpen met een doel als rassenscheiding voor ogen. Omdat het onmiskenbaar zo is dat alle of bijna alle technologie zowel voor als tegen een morele waarde kan worden gebruikt, ligt de oplossing van de vermeende tegenstelling voor sommigen hierin: erken dat technologie waardeneutraal is. Hoe overtuigend is die redenering?

De waardegeladenheid van technologie

Ik ga het nu van een andere kant bekijken. Dadelijk som ik de argumenten op die laten zien dat de neutraliteitsthese niet klopt. Van de stoïcijnen tot de technologiegoeroes van vandaag, velen opperen dat technologie neutraal is, maar op de bladzijden die volgen wil ik ingaan tegen dat dogma. Technologie is niet per definitie waardeneutraal, sommige technologie is waardegeladen, zo zal ik straks besluiten. Maar vooraleer ik daartoe kom, geef ik eerst enkele voorbeelden van technologieën die duidelijk waardegeladen zijn, zij het in moreel neutrale zin, maar wel op een manier die op z’n minst doet twijfelen aan de bewering dat geen enkele technologie met een morele waarde is geladen.

OVERLEVEN EN GOED LEVEN

Eerder heb ik erop gewezen dat er twee soorten waarden zijn: morele en moreel neutrale. Een voorbeeld van een moreel neutrale waarde is efficiëntie. Die zit niet per definitie maar wel vaak in technologie gebakken. De materialen zijn dan zo gekozen en samengesteld dat het gebruik van de technologie niet met onnodige verliezen gepaard gaat. Dat is natuurlijk geen argument tegen de neutraliteitsthese. Die these luidt immers enkel dat technologie niet met *morele* waarden is geladen. Daarnaast is het ook niet direct een reden om aan de neutraliteitsthese te twijfelen. Efficiëntie verschilt immers sterk van morele waarden als rechtvaardigheid of privacy.

Naast efficiëntie zijn er nog andere moreel neutrale waarden. Neem waarheid. Die kan wel een moreel karakter hebben (denk aan liegen), maar ze kan ook moreel neutraal zijn. Dat is zo in de context van wetenschap, in de zin dat onderzoekers streven naar ware theorieën. Wetenschappers streven naar het formuleren van beweringen die overeenstemmen met de realiteit: ze willen bijvoorbeeld uitpluizen wat de ontstaansgeschiedenis van Chinese vazen is of wat de oorzaken zijn van de ongelijke behandeling van mannen en vrouwen. Vaak spelen technologieën in die context een belangrijke rol. Denk aan telescopen, atoomklokken, microscopen, deeltjesversnellers of computers. Ook de voorloper van de camera hoort in dat rijtje thuis, want die was oorspronkelijk gemaakt om inzicht te krijgen in het loopgedrag van paarden. Al deze technologieën bestaan met het oog op waarheid, en dus zijn ze geladen met waarde, zij het met een moreel neutrale waarde. Maar als technologieën met dit type waarde geladen kunnen zijn, waarom dan niet met een morele?

Sommigen vinden dat nog steeds geen reden om een vraagteken te plaatsen bij de neutraliteitsthese, omdat waarheid nog te veel verschilt van een waarde als *fairness* of privacy. Maar is het verschil tussen een morele en moreel neutrale waarde soms niet verwaarloosbaar? Denk aan een pacemaker, infuuspomp, stethoscoop, defibrillator of MRI-scanner. Of neem de slimme contactlenzen voor diabetici van Google en de Zwitserse farma Novartis. Die bevatten een chip die via de tranen het suikerniveau in het bloed meet. De lens verkleurt als het suikerniveau niet op het gewenste niveau zit: groen als je suiker moet eten, rood als je insuline moet spuiten.[32] De achterliggende idee is duidelijk: tal van technologieën worden ontworpen om ziektes te voorkomen, controleren en genezen. Tal van technologieën bestaan omwille van geen andere reden dan onze gezondheid, en zijn dus geladen met een medische waarde. Als technologie geladen kan zijn met dit soort waarde, waarom zou technologie dan niet met morele waarden geladen kunnen zijn? Natuurlijk zijn er verschillen. Gezondheid is een medische waarde en heeft te maken met overleven; morele waarden gaan daarentegen over het goede leven.

Maar zijn dat niet *louter* verschillen? Of is er iets bijzonders aan die verschillen waardoor de ene wel en de andere niet in dingen kan worden verwerkt? Zien zij die de neutraliteitsthese verdedigen ethiek niet ten onrechte als een uitzondering?

DE MORELE PLICHT VAN DE ONTWERPER

We zijn aanbeland bij de kern van mijn verhaal. Ik beroep me dadelijk op drie argumenten om de neutraliteitsthese te ontkrachten: het voorzorgsargument, het neveneffectenargument en het verbeteringsargument. Wat houden die precies in?

Ontwerpers en ingenieurs hebben niet de plicht om het welzijn van gebruikers of stakeholders te verhogen. Dragen ze bij aan het welzijn, dan loven we ze, maar als ze dat niet doen, dan worden ze niet gestraft. Wel moeten ze een zo goed mogelijk zicht proberen te krijgen op de mogelijke onwenselijke neveneffecten van de technologie die ze aan het maken zijn, en er zo veel mogelijk voor zorgen dat die gevolgen niet plaatsvinden. Ongelukken gebeuren en sommige onwenselijke gevolgen zijn niet te voorzien. Maar het is wel de plicht van de ingenieur om voorzorgsmaatregelen te nemen als blijkt dat er een reëel risico op schade is bij het gebruik van de technologie. Technologie ontwerpen is meer dan alleen een technische zaak, het is ook een kwestie van ethiek. Een voorbeeld. Beeld je in dat je medische technologie ontwerpt en dat je weet dat men die apparatuur twee keer zal gebruiken: eerst voor therapeutische doeleinden, daarna om gegevens van patiënten te verzamelen, gegevens die nuttig zijn voor wetenschappelijk onderzoek. Als je als ontwerper die kennis hebt, dan moet je de technologie zó ontwerpen dat de gegevens van de patiënten geanonimiseerd kunnen worden: men moet de namen kunnen verwijderen, net zoals de geboortedatum en -plaats. Kortom, je moet ervoor zorgen dat de patiëntgegevens die relevant zijn voor de wetenschap kunnen worden losgekoppeld van het individu. Een ander voorbeeld dat in de lijn daarvan ligt: als het gevaar bestaat dat de algoritmen die men gebruikt bij aanwervingen

of het toekennen van leningen selecteren op kenmerken als etniciteit, geslacht, leeftijd, beperkingen of uiterlijk, dan moeten programmeurs de algoritmen zo schrijven dat ze niet leiden tot racisme, seksisme, *ageism*, *ableism* of *lookism*.

Als nu deze risico's bestaan en de ontwerper vervolgens effectief mechanismen in de technologie bouwt om zo die onwenselijke gevolgen te voorkomen, dan kun je die technologie als een argument inbrengen tegen de neutraliteitsthese. Zeker, de waardegeladenheid is in die context louter negatief. De technologie is immers betrokken op een morele waarde, enkel in de mate dat ze is ontworpen om die waarde *niet* te bedreigen. Toch is negatieve betrokkenheid een reden om de bewering dat technologie waardevrij is te verwerpen. Als een technologie zodanig is ontworpen dat het moet voorkomen dat een morele waarde niet wordt bedreigd, en de technologie dus met waarde is geladen, kun je niet meer beweren dat die technologie moreel neutraal is. Dat is het voorzorgsargument tegen de neutraliteitsthese.

DE TELEFOON VAN BRITISH TELECOM

Verhinderen dat iets een onwenselijk effect heeft op een morele waarde is niet hetzelfde als zorgen voor méér van die waarde. Onrechtvaardigheid voorkomen verschilt van streven naar meer rechtvaardigheid. De eerste interventie heeft een negatief karakter, de tweede een positief.

Het spreekt voor zich dat sommige technologieën zo'n positief karakter hebben. Tal van artefacten worden gebruikt met het oog op meer rechtvaardigheid of autonomie. Van belang is dat ook die positieve band tussen een technologie en een morele waarde niet louter tot stand komt als de technologie wordt gebruikt. Die band zit vaak al ingebakken in de technologie zelf, net zoals de negatieve band van daarnet. Dat is waar het neveneffectenargument op doelt, het tweede argument tegen de smalle invulling van de neutraliteitsthese. Neem inclusiviteit. Dat is een morele waarde die verwijst naar de opname van mensen in een samenleving of organisatie, onafhankelijk

van hun leeftijd, seksuele oriëntatie, huidskleur of andere zaken. Het is duidelijk dat het gebruik van bepaalde technologieën onwenselijke effecten heeft op inclusiviteit. Je hoeft daarvoor niet alleen aan wapens te denken, ook geldautomaten hoger dan een rolstoel of gebouwen zonder lift sluiten mensen uit. Toch kun je niet ontkennen dat bepaalde technologieën ook in wenselijke zin zijn verbonden met de morele waarde van inclusiviteit. Ik denk dan bijvoorbeeld aan de telefoon van het bedrijf British Telecom uit de jaren 1980. Die was ontworpen met grote toetsen en dito cijfers op de toetsen. Het doel was om ook ouderen en mensen met beperkte motorische en visuele vermogens in staat te stellen om op afstand te communiceren. Ook sommige computergames passen binnen dat kader. Neem *Mass Effect* en *Dragon Age* van het Canadese bedrijf BioWare. In de versie van *Mass Effect* die in 2012 op de markt werd gebracht, werd ook een mannelijk personage opgenomen wiens seksuele voorkeur naar andere mannen uitgaat. Het is niet dat homoseksuele personages nieuw zijn in de gamewereld, want er werden in het verleden al spellen gemaakt met vrouwen die op andere vrouwen vielen. Het nieuwe aan *Mass Effect* was dat het een mannelijk homoseksueel personage introduceerde. In *Dragon Effect* zitten dan weer personages met biseksuele voorkeuren en wordt aandacht geschonken aan transseksualiteit. Het doel is inclusiviteit, om zowel binnen als buiten de gamewereld ruimte en respect te creëren voor alle identiteiten, en niet enkel voor de heteroseksuele norm.[33]

De telefoon en computerspellen verschillen van technologieën die in negatieve zin waardegeladen zijn, zoals de medische technologie die patiëntgegevens loskoppelt van individuen. Die laatste is uitgerust met een mechanisme om te verhinderen dat een bepaalde moreel wenselijke toestand wordt bedreigd. De telefoon en computerspellen hebben daarentegen een positief karakter, in de zin dat ze gericht zijn op meer inclusiviteit. Toch is er ook een gelijkenis tussen enerzijds de medische technologie en anderzijds de telefoon en computerspellen. De effecten zijn telkens bedoeld door de ontwerpers, wat wil zeggen dat de bouw van de technologie is afgestemd op een morele waarde. Die gerichtheid van het ontwerp op een wenselijke toestand houdt in

dat aan de technologie een morele waarde kleeft, onafhankelijk van het gebruik. Voor een goede omschrijving van de technologie in kwestie moet je in beide gevallen dus verwijzen naar een morele waarde. Deze positieve betrokkenheid is het tweede argument tegen de bewering dat technologie niet waardegeladen is.

VERBETERING DOOR TECHNOLOGIE

Het tweede argument luidt dat sommige technologieën zijn ontworpen met het oog op meer *fairness* en privacy. Dat effect is bedoeld, lees: het ligt in het ontwerp besloten. Toch snijdt de band tussen technologie en morele waarden hier niet diep. Als je de toetsen en cijfers van de telefoon van British Telecom zou vervangen door kleinere dan zou daardoor een specifiek soort telefoon verdwijnen, maar een telefoon met minder grote toetsen en cijfers is nog steeds een telefoon. Het gaat niet ineens over een ander type technologie. Het effect, inclusiviteit, is dan wel bedoeld, maar het is niet essentieel voor de omschrijving van de technologie. Dat is ook de uitleg voor de naam die aan dat argument wordt gegeven: het *neveneffecten*argument.

Sommige technologieën zijn echter in de eerste plaats gemaakt met het oog op de toename van een morele waarde. Daarover gaat het verbeteringsargument, de derde reden om de neutraliteitsthese te verwerpen. Een voorbeeld van zulke technologie is Tor, voluit *The Onion Router*, dat oorspronkelijk werd ontwikkeld met steun van het *Defense Advanced Research Projects Agency* (DARPA) en de Amerikaanse marine. Wat is Tor?[34]

Wanneer je op het web surft, dan kunnen jouw locatie en IP-adres normaal gezien probleemloos worden achterhaald. Bezoek je een bepaalde website, dan is het doorgaans niet erg moeilijk de fysieke plaats te traceren waar jij aan je computer zit. Software als Tor stelt je in staat om volkomen anoniem te surfen. Je hoeft daarvoor niet veel meer te doen dan het downloaden en installeren van de Torbrowser,

net zoals je dat met Chrome en Firefox zou doen. Daardoor word je niet gekoppeld aan één server, zoals dat normaliter het geval is, maar aan een netwerk van servers. Geef je een webadres op, dan word je langs deze lange keten van servers geleid. Het punt is nu dat iedere server in die lange rij uitsluitend routinginformatie bijhoudt van de vorige en de volgende server in de keten; een server kan de data van andere servers in de rij niet achterhalen. Het gevolg is dat de routinginformatie als het ware ‘kaal’ op de bestemming (de website) aankomt. Dat wil zeggen: vanaf de website die je bezoekt kan men jou niet traceren; het is dan onmogelijk uit te pluizen van welk IP-adres de website werd bezocht. Dat is ook waarom Tor een ui als logo heeft. De software werkt met verschillende lagen bescherming. Achter die lagen valt niets te zien.

Anonieme communicatie is erg interessant voor militaire organisaties, en verklaart waarom het Amerikaanse leger de ontwikkeling van de software ondersteunde. Intussen wordt Tor gebruikt door journalisten, politiek activisten of klokkenluiders als Edward Snowden. Men kan kritiek uiten en actie voeren, zonder het risico op opsporing, vervolging of arrestatie. Wil dat nu zeggen dat Tor zonder gevaar is? In geen geval. De software is erg in trek bij hackers en cybercriminelen. En dat heeft precies te maken met het punt dat ik wil maken. Bepaalde technologie is ontworpen met mechanismen die de schending van de privacy moet voorkomen of is ontworpen met het oog op een morele waarde, maar zonder dat die waarde essentieel is voor de typering van het voorwerp in kwestie. Andere technologie echter, zoals Tor, is in de eerste plaats ontworpen om ervoor te zorgen dat er méér van een bepaalde morele waarde is, méér van privacy bijvoorbeeld. Het is precies ook dat surplus aan anonimiteit dat zo aantrekkelijk is voor hackers en misdadigers. En het is ook om die reden dat je Tor kunt gebruiken tegen zij die de neutraliteitsthese verdedigen.

DE FOUT VAN MOSES

Op basis van de voorbije paragrafen zou de indruk kunnen ontstaan dat alle waardegeladenheid wenselijk is. En doorgaans is dat ook wel zo, maar dat is niet altijd het geval. Ik wees er eerder al op: sommige morele waarden kunnen voor bepaalde mensen wenselijk zijn maar voor de meerderheid onwenselijk. Denk aan eugenetica uit de negentiende en twintigste eeuw, en dan voornamelijk aan de nazistische variant ervan. Dat was een vorm van onderzoek die was gericht op het versterken van het Arische ras en het uitroeien van zogenaamde inferieure rassen. Ja natuurlijk is dat doel verwerpelijk, maar rassenscheiding was wel iets dat in nazi-Duitsland als wenselijk werd gezien. Lees: de scheiding van rassen was toen een morele waarde. Conclusie? De ontmoeting tussen wetenschap en morele waarde kan dramatische gevolgen hebben en kan dus volstrekt onwenselijk zijn.

Wat geldt voor wetenschap geldt ook voor technologie. Technologieën kunnen los van de gebruiksfase in moreel verwerpelijke zin waardegeladen zijn. Veruit het bekendste voorbeeld is afkomstig van politiek wetenschapper Langdon Winner uit zijn beroemde tekst ‘Do Artifacts have politics?’ uit 1980.[35] Ik weet dat Winners voorbeeld intussen bijna versleten is, en ik besef dat mogelijk niet alle informatie in zijn tekst even betrouwbaar is, maar dat maakt zijn denkoefening daarom niet minder interessant. Laten we daarom even stilstaan bij de problematiek waarop Winner de aandacht wilde vestigen.

Winner bespreekt in die tekst enkele projecten van Robert Moses, de Amerikaanse stadsplanoloog die tussen de jaren 1930 en 1970 een grote invloed had op de stedelijke ontwikkeling van New York en de omliggende regio. Meer in het bijzonder zoomt hij in op de befaamde Parkway-bruggen over de snelweg richting Long Island. De reden dat die bruggenconstructie Winners interesse genoot, is dat de viaducten bijzonder laag over de snelweg waren gebouwd. Hoewel dat het gevolg van onkunde van de ingenieurs zou kunnen zijn, is dat hier geenszins het geval. De bruggen zijn door Moses met opzet zo

geconstrueerd. Waarom? Het leeuwendeel van de AfroAmerikaanse gemeenschap had geen auto om naar Long Island te rijden, en dan meer in het bijzonder naar het strand. Men nam daarom doorgaans de bus om zich te verplaatsen. Omdat de bruggen zo laag gebouwd waren, was het echter voor die bussen onmogelijk om daaronder door te rijden en om de New Yorkers met donkere huidskleur naar het strand te vervoeren. Hoe verwer-pelijk dat ook is, het was precies wat Moses voor ogen had. Het ultieme doel van zijn ontwerp was een strand met uitsluitend witte mensen, met andere woorden: rassenscheiding.

Deze casus is een variant van het neveneffectargument, de tweede reden om de neutraliteitsthese te verwerpen. In het geval van de computerspellen en de telefoon van British Telecom gaat het over wenselijke morele waarden, nu over onwenselijke. De gelijkenis is dat de waarde gebakken zit in de technologie en dat de waardebetrokkenheid niet bijzonder diep snijdt: ze is niet essentieel om de technologie te begrijpen. Een brug die niet discrimineert is nog steeds een brug, net zoals een telefoon zonder grote toetsen of cijfers nog steeds een telefoon is.

WEERZINWEKKENDE DINGEN

Zijn er daarnaast technologieën waarbij de geladenheid met een onwenselijke waarde wel cruciaal is om de technologie te definiëren? Heeft ook het verbeteringsargument een variant die te maken heeft met een waarde die door sommigen als goed wordt gezien maar door de meerderheid als moreel verwerpelijk?[36]

Niemand zal ontkennen dat de ovens van de nazi’s artefacten zijn, maar sommigen opperen wel dat die artefacten zelf moreel neutraal zijn. Je kunt geen moreel oordeel vellen over de technologie (en de ontwerper), zo meent men weleens, maar alleen over zij die de ovens gebruiken. Toch is er een goede reden om dat tegen te spreken. Het ontwerp van die ovens verschilt immers sterk van de reguliere oven

om lichamen te cremeren. Ze hebben geen esthetische ornamenten, ze hebben meer geluidsdempers, ze onderscheiden het as van de verschillende lijken niet, en de ovens hebben meer gangen om het proces te versnellen. Het ontwerp maakt met andere woorden duidelijk dat de ovens zijn ontworpen met het oog op een verwerpelijk doel: massavernietiging. Het massaal verbranden van lijken is niet iets dat je *ook* kunt doen met de ovens, los van het hoofddoel. Nee, genocide is het hoofddoel van de ovens, zo blijkt wanneer je nader naar de ovens kijkt. Daardoor is het niet alleen absurd te beweren dat je alleen over de gebruiker een moreel oordeel kunt vellen en niet over de ontwerper, het toont ook opnieuw dat sommige technologieën door en door waardegeladen zijn. Software als Tor is wenselijk, de ovens zijn gruwelijk, misselijkmakend. Niettemin komen ze hierin overeen dat je ze niet kunt definiëren zonder naar een waarde te verwijzen, waardoor de ovens van de nazi's (net zoals bijvoorbeeld ook atoombommen) een variant van het verbeteringsargument zijn.

Daarmee rond ik de kern van mijn betoog af. Ik vat samen. Wie de neutraliteitsthese verdedigt, stelt dat technologie niet met morele waarden is geladen. Tegen deze stelling heb ik drie argumenten gegeven. De overeenkomst is dat ze telkens uitgaan van de ontwerper die is gericht op een morele waarde. Deze gerichtheid vertaalt zich in het ontwerp van een artefact, waardoor dat artefact een waarde bevat en je niet meer kunt beweren dat technologie losstaat van een morele waarde. Er is niet noodzakelijk een kloof tussen technologie en morele waarde, zo besluit ik; zij kunnen met elkaar verstrengeld zijn. Het verschil tussen de argumenten heeft te maken met het type van waardegeladenheid (negatief of positief). Het eerste argument brengt in herinnering dat sommige technologieën in negatieve zin met morele waarde zijn geladen. Er zijn mechanismen die moeten verhinderen dat de technologie onwenselijke effecten heeft. De twee andere argumenten refereren aan het feit dat het doel van sommige technologieën is om positief bij te dragen aan een pakweg rechtvaardigere wereld. Niettemin is er ook een verschil tussen het

tweede en derde argument. In het ene geval is de waardegeladenheid essentieel om de technologie te begrijpen, in het andere geval niet.

Kritische noten

Het lijkt me goed om nu zeker twee keer op de rem te gaan staan. Wanneer blijkt dat technologie niet noodzakelijk moreel neutraal is, bestaat het gevaar dat we ons vergalopperen en dat we het tegenovergestelde verkondigen: *alle* technologie is waardegeladen, *geen enkele* technologie is waardeneutraal. Een ander risico is dat ontwerpers menen dat het uitrusten van technologie met een morele waarde geen extra problemen kan creëren. Want waarom zou een morele waarde *nieuwe* problemen met zich meebrengen? Op de volgende bladzijden zet ik uiteen hoe je je tegen beide risico's kunt beschermen.

DE NEUTRALITEIT VAN EEN BOORMACHINE

Ik heb er al op gewezen dat ontwerpers de plicht hebben om na te gaan welke ongewenste gevolgen zouden kunnen voortvloeien uit het gebruik van hun ontwerp. Als zulke risico's worden gevonden, dan moeten in het ontwerp mechanismen worden ingebouwd om zulke effecten te voorkomen. De technologie is dan in negatieve zin waardegeladen, zo weten we intussen. Is technologie echter *noodzakelijk* waardegeladen in deze zin?

Het antwoord op deze vraag is om twee redenen ontkennend. Ten eerste kan technologie worden gemaakt zonder dat er wordt gekeken naar de mogelijke onwenselijke gevolgen. Voorzorgsmaatregelen nemen volstaat niet voor het maken van technologie, maar is daarnaast ook niet noodzakelijk. Dat houdt minstens de mogelijkheid in dat een ontwerper, hoewel dat vanzelfsprekend moreel bijzonder laakbaar is, tijdens de ontwerpfase zijn of haar plichten verzaakt en niet kijkt naar de mogelijke schadelijke effecten. In dat geval kun je

niet beweren dat het artefact in negatieve zin met waarde is geladen. Ten tweede is het ook mogelijk dat de ontwerper tijdens de ontwerpfase wel nadenkt over de mogelijke ongewenste gevolgen, maar dat zij of hij meent dat het gebruik van de technologie niet gepaard zal gaan met schadelijke effecten op waarden als privacy of *fairness*. In dat geval wordt het ontwerp niet gewijzigd in functie van mogelijke onwenselijke gevolgen, en verwijst het artefact zelf in geen enkel opzicht naar een morele waarde. Veel 'primitieve' technologieën zijn daarvan voorbeelden: scharen, hamers, messen, vorken, borstels en andere basale dingen.

Hoe zit het met die andere waardegeladenheid? Zijn alle technologieën in positieve zin met waarde geladen? Zijn zij allemaal ontworpen met het oog op het realiseren van een morele waarde? We hebben al gezien dat alle technologieën instrumentele waarde *hebben*. Dat is per definitie het geval, omdat ze zijn ontworpen met een doel. Verder is het ook zo dat veel technologieën in positieve zin waardegeladen zijn. Denk aan de voorbeelden die ik eerder noemde: Tor, de computerspellen, de telefoon van British Telecom. En toch zijn niet alle technologieën in positieve zin waardegeladen. Neem een boormachine. Dat is een artefact dat is ontworpen om een schroef in een stuk hout of steen te draaien, of om een opening in een muur te maken. Maar een opening in de muur maken of een schroef in een stuk hout draaien zijn geen waarden, laat staan morele waarden. Natuurlijk kunnen veiligheidsmechanismen zijn ingebouwd, waardoor het ontwerp naar de waarde van veiligheid verwijst. Maar die waardegeladenheid is niet noodzakelijk opdat een technologie een boormachine is. Om een boormachine te zijn, volstaat het dat een technologie is ontworpen om een schroef in iets anders te draaien. En aangezien dat doel geen morele waarde is, kun je zeggen dat zo'n artefact moreel neutraal is.

Sommigen antwoorden daarop als volgt. Ook al is het directe doel van een boormachine zelf geen waarde, dat doel is wel altijd verbonden met een waarde. Toegegeven, als dat juist is, dan wordt het wel erg moeilijk te ontkennen dat alle technologie in positieve zin

waardegeladen is, zelfs al is de band tussen de technologie en waarde in dat geval indirect. Maar klopt het dat het doel van elke technologie indirect is verbonden met een waarde? Neem opnieuw een boormachine. Je kunt dat artefact gebruiken om bijvoorbeeld een kapstok aan de muur te hangen. In dat geval is er inderdaad een relatie tussen de boormachine en een waarde als comfort of orde. Maar je kunt die machine natuurlijk ook gebruiken voor andere zaken, bijvoorbeeld om een koekoeksklok in de woonkamer te hangen. Draagt de boormachine dan bij tot de realisering van een waarde? Ik kocht die klok, omdat ik het leuk vind dat ieder uur het geluid van een koekoek wordt nagebootst. Ook vind ik het erg grappig dat mijn vrienden meerdere keren schrikken wanneer ze me bezoeken. Maar leidt dat ook tot iets belangrijks, iets dat zo veel gewicht heeft dat het *moet* worden gerealiseerd? Neemt daardoor ook mijn geluk of welzijn toe? Wie daarop bevestigend antwoordt, verwart waarden met zaken die leuk zijn. Er is wellicht niemand die zegt dat iets leuk moet zijn om een waarde te zijn, maar één ding is wel duidelijk: het volstaat niet dat iets leuk is om een waarde te zijn.

Het nadenken over de relatie tussen technologie en waarde vraagt dus enige nuancering. Niet alles is waardeneutraal, zo liet ik al zien. Maar nu blijkt ook dat niet elke technologie met waarde is geladen; niet alle machines zijn morele machines. Er is geen noodzakelijke relatie tussen technologie en morele waarde, net zoals technologie en neutraliteit niet onlosmakelijk met elkaar zijn verbonden. Sommige technologie is waardegeladen, sommige moreel neutraal.

AND JUSTICE FOR ALL

Ik liet daarnet vallen dat ik twee keer op de rem wil gaan staan. De eerste keer had het te maken met het neutrale karakter van sommige technologieën. Het tweede punt is het volgende. Het spreekt voor zich dat het ontwerpen van technologieën vaak een moeilijk proces is, waarbij verschillende soorten problemen kunnen optreden: technische problemen bijvoorbeeld, maar ook logistieke. Maar is

daarmee de kous af? Zijn de mogelijke ontwerpproblemen uitsluitend van die aard? Nee, er kunnen ook problemen opduiken die te maken hebben met het feit dat het over morele waarden gaat.
Ik stip er hier twee aan. Het eerste is het spraakverwarringsprobleem.[37]

Wellicht vindt bijna niemand rechtvaardigheid en autonomie onbelangrijk. Ze zijn zó belangrijk dat we het als onze plicht zien ernaar te streven. Daarover is iedereen of zo goed als iedereen het eens. Maar bestaat die eensgezindheid ook als je inzoomt op de invulling ervan? Betekenen zij voor iedereen hetzelfde? Het antwoord daarop is ontkennend. Laat ik bij wijze van voorbeeld toespitsen op *fairness*. In onze samenleving wordt die morele waarde op minstens twee manieren begrepen: *fairness* als statistische gelijkheid en *fairness* als gelijke kansen.

De eerste invulling moet je als volgt begrijpen. Als je twee groepen hebt, mannen en vrouwen bijvoorbeeld, en je moet een product verdelen over een groep, bijvoorbeeld toegangskaarten voor een evenement, dan is het volgens deze benadering fair dat er verhoudingsgewijs een gelijke verdeling van het product over beide groepen is. *Fairness* houdt dus niet noodzakelijk gelijkheid in absolute aantallen in. Als er vijftig mannen en tien vrouwen zijn, is het volgens deze benadering fair dat er meer mannen dan vrouwen tickets krijgen, en unfair dat evenveel mannen als vrouwen tickets krijgen. De tweede interpretatie van *fairness* heeft alles te maken met gelijke kansen. Dat wil zeggen: de startpositie waarin personen zitten – ze beginnen bijvoorbeeld aan een universitaire studie –, moet gelijk zijn of minstens zo gelijk mogelijk. Het is vanuit die optiek niet direct een probleem als twee personen later in hun leven een verschillend loon hebben – de een kan immers harder hebben gewerkt dan de ander. Wat wel problematisch is, is dat zij op jonge leeftijd ongelijke kansen hadden wegens een andere sociaaleconomische achtergrond. Volgens de tweede opvatting is het dan fair om beiden ongelijk te behandelen, bijvoorbeeld door de ene wel een studiebeurs te geven en de andere

niet, als het effect daarvan is dat de ongelijke startpositie wordt gelijkgetrokken of dat minstens de kloof minder groot is.

Tegenwoordig wordt ook in techmiddens meer en meer ingezet op *fairness*. Dat is natuurlijk goed; het zou problematisch zijn, mocht men daar niet of nauwelijks op inzetten. Maar tegelijk is de vraag: wat bedoelen we eigenlijk als we zeggen dat iets (un)fair is? En wat geldt voor *fairness*, geldt misschien niet voor alle morele waarden, maar wel voor privacy, duurzaamheid, enzovoort. Waarom is zulke conceptuele vraag nu relevant als het over technologie gaat?

Stel dat de overheid beslist om middelen vrij te maken voor de ontwikkeling van AI-systemen die kunnen worden ingezet bij de toekenning van studiebeurzen. Er wordt een bedrijf geselecteerd, en de overheid deelt het bedrijf uitdrukkelijk mee dat het AI-systeem fair moet zijn. Aangezien *fairness* op minstens twee manieren kan worden geïnterpreteerd en er twee betrokkenen zijn (overheid en bedrijf), bestaat de kans dat in een en hetzelfde proces twee begrippen van *fairness* worden gebruikt. Stel dat dat inderdaad het geval is, en dat de betrokkenen zich daar bovendien niet van bewust zijn. Op het ministerie van de betrokken bewindspersoon en in het bedrijf denkt men verschillend over *fairness*, en neemt men ten onrechte aan dat iedereen dezelfde opvatting heeft. Het gevolg is het spraakverwarringsprobleem: de betrokkenen spreken wel met elkaar, maar in feite praten ze langs elkaar heen. Hoewel dat op zich misschien niet erg problematisch is, is de kans wel reëel dat dat niet zonder onwenselijke gevolgen is. Waarom?

Wanneer het ministerie de opdracht geeft om de technologie te ontwikkelen, zal het eindresultaat, net voor het op de markt wordt gebracht, geëvalueerd worden. In de geschetste context zal dat negatief uitpakken. De reden is dat, precies vanwege het verschil in opvatting, de bedoelde effecten van de technologie die door het bedrijf werd ontwikkeld niet overeenkomen met wat het ministerie voor ogen heeft. In het beste geval leidt dat tot niet meer dan wat fricties, maar het is ook niet uitgesloten dat men aan de ontwikkelaar vraagt om de

technologie aan te passen of, sterker nog, om opnieuw te beginnen met het ontwerp. Als dat laatste zich echter zou voordoen, dan werden geld en tijd verspeeld, puur op basis van de foute aanname dat er consensus bestond over wat *fairness* betekent. Je moet daaruit de volgende les trekken. Hoewel de ontwikkeling van technologie met het oog op een waarde bij voorkeur is gestoeld op onder meer wetenschappelijke kennis, is dat onvoldoende. Je dient ook te weten dat morele waarden anders kunnen worden begrepen, en het is belangrijk dat je je eigen opvatting duidelijk maakt aan de andere betrokkenen.

Het laden van technologie met een morele waarde kan daarnaast ook gepaard gaan met een tweede probleem, het zogeheten botsingsprobleem. Waarden zijn niet per se afgestemd op elkaar, ze kunnen ook clashen. Dat vergt wat toelichting.

Stel je voor dat een bank voor het toekennen van leningen een AIsysteem gebruikt. Er zijn twee groepen cliënten. De eerste bestaat uit vrouwen met een zwarte huidskleur, de tweede uit witte mannen. Het AI-systeem heeft berekend dat als de bank een lening geeft aan een persoon uit de eerste groep, 15% van de vrouwen de lening zal terugbetalen; behoor je tot de tweede groep, dan zal 30% van de witte mannen zijn lening aflossen. Als zich iemand aanmeldt voor een lening, dan kan het AI-systeem met 100% zekerheid voorspellen dat de cliënt de lening zal terugbetalen of niet. Er zijn noch vals positieven (personen die onterecht een lening krijgen), noch vals negatieven (personen die ten onrechte geen lening krijgen). Kortom, de bank beschikt over een artificieel systeem dat is uitgerust met een welbepaalde moreel neutrale waarde: accuraatheid.

Maar de bank vreest ook discriminatie. Ze vermoedt dat aan de zwarte vrouwen minder of niet zo snel een lening zal worden gegeven dan aan de witte mannen. De bank heeft daarom gekozen voor een AI-systeem dat niet alleen met de moreel neutrale waarde van accuraatheid is geladen, maar ook met de morele waarde van *fairness*. Dat betekent in deze context dat het AI-systeem evenveel

aan beide groepen moet geven (*fairness* als statistische gelijkheid). Van belang is dat er evenveel zwarte vrouwen als witte mannen een lening krijgen. Maar dat is niet zonder gevolgen voor de accuraatheid van de technologie. Stel dat het AI-systeem beslist om aan 30% van beide groepen een lening toe te kennen. In dat geval is er wel statistische gelijkheid, maar zijn er vals positieven: 15% van de zwarte vrouwen heeft dan ten onrechte een lening gekregen. Een andere optie is dat de bank aan 15% van iedere groep een lening toekent. Ook dan is er tussen de twee groepen een gelijke spreiding van toegekende leningen, maar in dat geval zijn er dan weer vals negatieven. 15% van de witte mannen heeft namelijk ten onrechte niet de middelen ontvangen om bijvoorbeeld een woning aan te kunnen schaffen.[38]

Dit is een voorbeeld van het botsingsprobleem. Het laat zien dat het laden van technologie met een morele waarde een negatief effect kan hebben op een andere waarde die met de technologie is verbonden. In dit geval heeft *fairness*, opgevat als statistische gelijkheid, een onwenselijke invloed op de moreel neutrale waarde van accuraatheid. Hoe gelijker het AI-systeem de leningen distribueert, hoe minder accuraat de technologie, en omgekeerd: dat het systeem goed functioneert, wil niet zeggen dat het ethisch gezien goed is. Conclusie? Wanneer je besluit om technologie met ethiek te laden, moet je dus niet alleen weten dat morele waarden meerdere zaken kunnen betekenen (en je keuze duidelijk maken), je moet ook beseffen dat waarden kunnen clashen. Al wil ik daar ook nog het volgende aan toevoegen. Het is één ding dat je je er bewust van bent, een ander ding is dat je het niet uit de weg gaat. De waarden worden bij voorkeur ook onderling afgewogen (*trade-off*). In het voorbeeld van daarnet betekent het dat je moet uitvissen of je *fairness* zo belangrijk vindt dat je er de accuraatheid van het technologisch systeem voor opoffert, al dan niet ten dele.

Het belang van neutraliteit

Nu mijn kritiek is afgerond, keer ik terug naar de verdediging van de neutraliteitsthese, de diepgewortelde overtuiging dat aan technologie geen morele waarden kleven. De tweede argumenten die ik daarnet heb gegeven – het zichtbaarheidsargument en het dubbelegebruiksargument – schieten mijns inziens dus tekort. Maar wat schort er precies aan? Waarom zijn ze niet in staat om te doen wat ze moeten doen, namelijk de neutraliteitsthese ondersteunen? Daar ga ik nu eerst naar op zoek. Daarna vraag ik me af of het verdedigen van de neutraliteitsthese ook niet met andere zaken dan argumenten te maken heeft. Schuilen achter die these geen belangen en diepgewortelde ondoordachte denkbeelden?

WAARDEN ZIEN

Om te beginnen breng ik in herinnering dat het eerste argument voor de neutraliteitsthese, het zichtbaarheidsargument, uiteenvalt in twee delen. Het eerste deel houdt in dat technologie alleen waardegeladen kan zijn als je in het ontwerp een morele waarde kunt aanwijzen. Het tweede deel luidt dat het onmogelijk is om een plek aan te wijzen waar je de waarde daadwerkelijk kunt zien. De al eerder vermelde Pitt bijvoorbeeld, een fervent verdediger van de morele neutraliteit van technologie, schrijft het volgende als hij het over de vermeende waardegeladenheid van Moses' bruggen heeft: 'Waar zijn de waarden? Ik zie bakstenen en stenen en het voetpad, enzovoort. Maar waar zijn de waarden – hebben ze kleuren?

Hoeveel wegen ze? Hoe groot zijn ze of hoe mager?'[39] Waar rammelt het zichtbaarheidsargument?

Het probleem met de redenering is niet van logische aard. Als de twee delen van het argument juist zijn, dan moet je daar wel uit afleiden dat technologie moreel neutraal is. De vraag is echter of beide delen wel juist zijn. Pitt heeft in ieder geval gelijk wat het tweede deel betreft. Als je naar technologische artefacten kijkt, zie je knoppen, draden,

toetsen, poorten, glas, ijzer, handvaten en andere zaken. Maar zie je ook waarden in het voorwerp? Zeker, je kunt zien dat het materiaal is gekozen en samengesteld om een bepaalde waarde te realiseren. Maar waarden als duurzaamheid of privacy zelf kun je in de technologie niet aanwijzen. Dat is eigenlijk vanzelfsprekend. Technologieën zijn dingen, doorgaans met een materieel karakter. Morele waarden daarentegen zijn geen dingen. Zij gaan vaak over het leven van een persoon, denk aan autonomie, of over de relatie tussen personen onderling, bijvoorbeeld in het geval van *fairness*.

Het probleem heeft dus te maken met het eerste deel van het argument, dat luidt dat je een waarde in een voorwerp moet kunnen aanwijzen. Is dat onmogelijk, dan kan dat voorwerp niet waardegeladen zijn. Het is duidelijk dat die vereiste gestoeld is op een welbepaalde invulling van ‘waardegeladenheid’. Eerder in dit hoofdstuk stipte ik al aan dat je die term op twee manieren kunt interpreteren. Vat je het figuurlijk op, dan betekent het dat het materiaal of ontwerp naar een morele waarde verwijzen. Het is in dat geval niet nodig dat je een waarde echt waarneemt in de technologie om die waardegeladen te noemen. Een letterlijke interpretatie van ‘waardegeladenheid’ betekent dat de morele waarde deel uitmaakt van het ontwerp, net zoals kabels en buizen dat doen. Het is duidelijk dat het zichtbaarheidsargument op deze interpretatie is gebaseerd. De eis dat je een morele waarde in een ding moet kunnen zien voordat je het waardegeladen kunt noemen, heeft immers alleen zin als onder ‘waardegeladenheid’ wordt verstaan dat een waarde ook feitelijk tot het ontwerp behoort.

De vraag is dus niet zozeer waarop het eerste deel van het argument is gebaseerd, maar eerder waarom wordt gekozen voor de letterlijke interpretatie. Je zou verwachten dat men daar goede redenen voor heeft. Het gevolg van die keuze is namelijk dat je technologie niet anders dan moreel neutraal kunt zien, dat technologie niet waardegeladen kan zijn. Ik stipte het al meermaals aan: technologieën en morele waarden zijn twee volstrekt verschillende zaken, zodat een morele waarde onmogelijk in een technologie kan

zitten. Het probleem is nu dat zij die de neutraliteitsthese verdedigen op basis van het zichtbaarheidsargument hun keuze voor de sterke interpretatie niet onderbouwen. Sterker nog, er is mijns inziens geen goede reden waarom je niet voor de figuurlijke en wel voor letterlijke invulling zou moeten kiezen. Let wel, ik beweer niet dat er geen goede reden *is*. Ik stip alleen aan dat geen reden wordt gegeven door wie die eigenlijk wel zou moeten geven en dat ik zulke reden zelf bovendien niet zie. Het probleem met het zichtbaarheidsargument is dus volgens mij dat het vertrekt van een ongegronde beslissing.

Je kunt het argument ook bekritiseren door op het volgende te wijzen. We zeggen van veel zaken dat ze waardegeladen zijn. Ook wie beweert dat technologie waardeneutraal is, vindt dat. Alle of de meeste mensen vinden dat bijvoorbeeld wetenschappelijk onderzoek met het oog op vermindering van discriminatie waardegeladen is. Toch kun je de waarde van gelijkheid niet daadwerkelijk aanwijzen in de wetenschapspraktijk. Gelijkheid speelt tussen de personen naar wie onderzoek wordt gedaan, maar niet in boeken, studiedagen of interviews. Om onderzoek waardegeladen te noemen, is het dus kennelijk voor velen niet nodig, ook niet voor wie de neutraliteitsthese verdedigt, dat je een morele waarde kunt aanwijzen. Waarom zou het dan niet nodig zijn wanneer het over wetenschap en andere zaken gaat, maar wel wanneer het over technologie gaat? Is er een relevant verschil tussen beide waardoor de strenge eis wel geldt voor technologie en niet voor wetenschap? Wie zich beroept op het zichtbaarheidsargument geeft daarvoor geen verdere uitleg, en dat is in zekere zin niet verwonderlijk, want er is volgens mij geen relevante eigenschap. Het kan dat zo'n kenmerk bestaat – dat wil ik ook hier onderstrepen –, maar men geeft het in ieder geval niet terwijl men dat wel zou moeten doen, en ik zie zelf ook niet welk kenmerk het zou kunnen zijn. Het probleem is dus opnieuw dit: de argumentatie is in mijn ogen gestoeld op een ongefundeerde keuze.

Merk wel op: ik beweer niet dat het zichtbaarheidsargument op niets is gebaseerd. Ik vermoed wel dat het theoretisch ongefundeerd is, maar het zou kunnen dat er niet-theoretische motieven spelen,

belangen bijvoorbeeld. Het is met andere woorden mogelijk dat men kiest voor de letterlijke vorm van waardegeladenheid en dus voor het onvermijdelijke gevolg dat men technologie als waardevrij ziet, niet of niet alleen omdat men daar argumenten voor heeft, maar omdat men wil dat technologie als neutraal wordt gezien. In die richting wijst althans het feit dat het zichtbaarheidsargument is gebaseerd op mijn inziens ongegronde beslissingen – straks meer daarover.

ONTWERPEN EN GEBRUIKEN

Het tweede argument – het dubbelegebruiksargument – gaat uit van de vaststelling, ik lichtte dat eerder al toe, dat je alle of bijna alle technologieën zowel in goede als slechte zin kunt gebruiken. Dat dubbele gebruik, zo gaat de redenering, noopt tot de conclusie dat technologie waardeneutraal is. Kun je namelijk beweren dat een artefact is geladen met een waarde als privacy, terwijl het ook wordt gebruikt om de privacy van iemand te schenden, en omgekeerd? Nee, volgens sommigen. En aangezien alle of bijna alle technologie die werd ontwikkeld voor een waarde ook tegen die waarde kan worden gebruikt, moet je wel concluderen dat technologie waardeneutraal moet zijn.

Het dubbelegebruiksargument is gestoeld op een welbepaalde overtuiging. Aan het argument ligt een zogeheten deterministische aanname ten grondslag. Dat wil zeggen: men neemt aan dat het ontwerp van de technologie het gebruik ervan vastlegt. Er wordt van uitgegaan dat het laden van een ontwerp met een moreel goede waarde inhoudt dat je het voorwerp niet kunt gebruiken om het tegendeel te realiseren. Het is onmogelijk, zo meent men, dat een artefact is geladen met een bepaalde morele waarde, terwijl het wordt gebuikt op een manier die indruist tegen die waarde. De reden waarom nu het dubbelegebruiksargument niet doet wat het moet doen, namelijk de neutraliteitsthese ondersteunen, is dat de deterministische aanname niet klopt.

Laat ik een vergelijking maken om dat uit te leggen. Gezondheid is een waarde, geen morele maar een medische. Je kunt middelen gebruiken die je gezondheid ten goede komen, maar je kunt je gezondheid ook schaden, bijvoorbeeld door te roken. Met sommige artefacten kun je beide. Een injectienaald kun je gebruiken om een ziekte te voorkomen of te genezen, maar met datzelfde instrument kun je ook stoffen injecteren die je gezondheid niet ten goede komen. Toch hoeft dat niet te betekenen dat de injectienaald waardeneutraal is. Het is mogelijk dat die is gemaakt met de medische waarde van gezondheid voor ogen – de injectienaald is dan waardegeladen –, terwijl ze niet alleen in goede maar ook in slechte zin wordt gebruikt. Als het dus over een medische waarde gaat, klopt de deterministische aanname niet: de technologie kan waardegeladen zijn, maar kan ook worden gebruikt op een manier die ingaat tegen die waarde.

Waarom zou dat niet kunnen gelden voor morele waarden? Natuurlijk zijn er verschillen met medische waarden, maar zijn die van dien aard dat we niet op dezelfde wijze kunnen redeneren als het gaat over privacy, *fairness* of duurzaamheid? Neem een wapen dat in dubbele zin kan worden gebruikt. Enerzijds kun je er een volk mee bevrijden. In dat geval heeft het een wenselijk effect op de morele waarde van autonomie. Anderzijds kan men het ook tegen die waarde gebruiken, bijvoorbeeld door er een volk mee te terroriseren. Maar houdt dat ook in dat het voorwerp waardeneutraal is? Nee, het wapen kan zijn ontworpen met het oogmerk om een volk te bevrijden. Het is dan duidelijk met een morele waarde geladen, terwijl die geladenheid niet belet dat het in moreel verwerpelijke zin wordt gebruikt, bijvoorbeeld om een volk te onderdrukken. Omgekeerd kun je een wapen maken met een terroristisch doel voor ogen dat nadien toch in moreel juiste zin wordt ingezet. Dat goede gebruik maakt het daarom niet waardevrij. Het is en blijft met een waarde geladen, zij het dat die waarde nu een moreel verwerpelijk karakter heeft. Conclusie? Niet alleen op medisch maar ook op moreel vlak houdt de deterministische aanname geen steek, en dat is ook de reden waarom het

dubbelegebruiksargument niet tot de conclusie leidt dat technologie moreel neutraal is.

IT'S THE ECONOMY, STUPID!

We hebben nu een goed zicht op de problemen met de argumentatie voor de neutraliteitsthese. Toch wil ik niet de indruk wekken te geloven dat wie mijn uiteenzetting leest die these niet meer zal verdedigen. Hoe uitgebreid ik ook bij die these heb stilgestaan, het is mogelijk dat het argumenteren misschien amper of zelfs geen effect heeft op wie zegt dat technologie moreel neutraal is. Waarom is dat zo? Hoe komt het dat de filosofie hier tegen een grens kan aanlopen?

Als mensen zich tijdens een debat *uitsluitend* door argumenten laten leiden, scharen ze zich achter de bewering met de sterkste argumenten. Ze zijn in dat geval ook ontvankelijk voor tegenargumenten, waardoor ze hun mening kunnen herzien. Maar stellingen kunnen ook worden verdedigd op grond van andere zaken dan argumenten. Ik denk dan in de eerste plaats aan belangen. Sommigen scharen zich achter een opvatting, niet of niet alleen omdat ze daar goede argumenten voor hebben, maar ook omdat ze daar belang bij hebben, omdat er iets op het spel staat dat niets te maken heeft met theorie of inzicht. Het mogelijke gevolg is dat men niet openstaat voor argumenten die de eigen overtuiging ondergraven. Wanneer een belang in het geding is, kunnen mensen zich zodanig vastklampen aan een stelling dat ze die zullen blijven steunen, zelfs als blijkt dat die niet wordt gedragen door de beste argumenten. Zou het niet kunnen dat dit ook het geval is als het over technologie gaat, of meer nog, precies omdat het over technologie gaat? Is het niet zo dat sommigen zeggen dat technologie waardevrij is omdat men daar iets bij te winnen heeft? Sommigen verdedigen wellicht de neutraliteitsthese louter op grond van theoretische argumenten, maar de mogelijkheid die ik hier naar voren wil schuiven is dat anderen zich ook laten leiden door niet-theoretische motieven, of sterker nog, dat bij een aantal mensen *uitsluitend* zulke motieven spelen. De belangen

die ik hier voor ogen heb, zijn van morele en economische aard. Ik verduidelijk.

Tal van technologieën hebben een onwenselijk effect op morele waarden: AI-systemen discrimineren, auto's stoten broeikasgassen uit en apps schenden de privacy. Wanneer die effecten niet te voorzien waren, vellen we geen negatief oordeel over de ontwerper, fabrikant of stakeholder. Maar zij blijven niet noodzakelijk buiten schot. Wanneer de slechte gevolgen voorzien waren, terwijl men daar niet op anticipeerde, is dat een reden voor een negatieve evaluatie. Verder kun je de betrokkenen ook veroordelen als blijkt dat de onwenselijke effecten van de technologie niet alleen voorzien maar ook bedoeld zijn. Dat is het geval bij enkele van de eerdergenoemde voorbeelden. Terwijl we de makers van de computerspellen van daarnet eerder zullen prijzen, veroordelen we de ontwerpers van de Parkway-bruggen, net zoals we niet alleen een negatief oordeel over de gebruikers vellen, maar ook over de ontwerpers van de naziovens. De reden is telkens dezelfde: het verwerpelijke effect zit in het artefact verankerd, en vloeit dus niet uitsluitend voort uit de slechte bedoelingen van de gebruiker. Het effect is ook rechtstreeks verbonden met de ontwikkelaar, en dat is een reden om die te veroordelen.

Kijk je vanuit dat oogpunt naar de neutraliteitsthese, dan werpt dat op zijn minst een nieuw licht op de verdediging van die these. Het laat zien dat de verdediging van de neutraliteit van technologie misschien ook voortvloeit uit het verlangen om je handen in onschuld te wassen. Ik verwijs hier niet naar zij die de technologie uitrusten met moreel goede waarden als *fairness* of duurzaamheid. Ik doel op de ontwerper, fabrikant of stakeholder die moreel foute waarden als rassenscheiding en discriminatie nastreeft, en technologie ontwikkelt in dienst van die doelen. Alleen wil die zelf niet het onderwerp van afkeuring zijn; het is de gebruiker die alle morele schuld moet dragen. De betrokkene wil een moreel kwaad maar tegelijk haar of zijn handen in onschuld wassen. Zie hier het verlangen naar de schone handen dat achter de neutraliteitsthese verscholen kan gaan. Mij valt niets te verwijten, zo

lijkt de boodschap achter die these te zijn. Ik heb de technologie wel bedacht en gemaakt waaruit het moreel kwaad is voortgekomen, maar dat kwaad zit niet in de technologie verankerd. Wat ik ontwierp, is moreel neutraal, en dus heb je geen reden om mij moreel te veroordelen. Je moet de gebruiker veroordelen, want uit hem of haar vloeit het moreel kwaad voort, en niet uit het ontwerp. Hoewel dat vanzelfsprekend niet terecht is, is dat wel de redenering van de ontwerper of fabrikant die achter de verdediging van de neutraliteitsthese schuil zou kunnen gaan. Het is ook de redenering die de link legt tussen dit eerste hoofdstuk en het volgende, waarin we ons een hele tijd zullen buigen over het vraagstuk van morele verantwoordelijkheid.

Er is daarnaast nog een ander belang dat onder de neutraliteitsthese zou kunnen liggen, een belang dat een economisch karakter heeft. Om dat in te leiden, verwijs ik naar het volgende. In een liberale rechtstaat moeten overheden neutraal zijn. Ze mogen niet alleen geen voorkeur voor een religie of seksuele oriëntatie uitdragen, als daar geen goede reden voor is, is het ook niet toegelaten om technologieën te bevoordelen of benadelen. Maar er is wel een grens aan dat principe. Wanneer blijkt dat het gebruik gepaard gaat met onvoorziene erg slechte gevolgen, dan heeft men een reden of zelfs de plicht om de technologie van de markt te halen. De overheid kan zelfs nog sneller ingrijpen, bijvoorbeeld als tijdens de ontwerpfase al duidelijk is dat er mogelijk zeer onwenselijke neveneffecten kunnen optreden. Er kan echter nog een andere reden zijn om in te grijpen. De slechte effecten kunnen naast louter voorzien ook bedoeld zijn door de maker, ontwerper of stakeholder. Dat is zo als het ontwerp is geladen met een verwerpelijke waarde als ongelijkheid. In dat geval zit het slechte effect verankerd in de technologie en is het artefact ontworpen met dat foute doel voor ogen. De bruggen van Moses zijn in deze context opnieuw verhelderend. Het effect van het gebruik van de bruggen is dat er op het strand alleen witte mensen zijn. Dat gevolg is geen neveneffect, maar het doel van de bruggen. Moses heeft de bruggen gemaakt *opdat* het strand met alleen witte mensen bevolkt zou zijn.

Dat is voldoende om een negatief moreel oordeel te vellen over de betrokkenen, maar het is voor de overheid ook een reden om de ontwikkeling stop te zetten.

Zo bezien, zou het niet verbazen mocht blijken dat de neutraliteitsthese in leven wordt gehouden door de industrie, en dat het met name de ontwerpers en producenten van moreel zeer discutabele technologieën zijn die er zuurstof aan blijven geven. Enerzijds weet men heel goed dat hun technologie is geladen met foute waarden en dat dit voldoende is voor een verbod, maar anderzijds hebben zij er belang bij, economisch belang, dat het verbod er niet komt. Een rem op de productie betekent geen inkomsten en winst. Het is vanzelfsprekend niet uitgesloten dat industriëlen zich op argumenten beroepen, maar het is ook niet uitgesloten dat achter die argumentatie belangen verscholen zitten. Het is gerechtvaardigd dat men het gebruik van technologieën verbiedt die onwenselijke effecten sorteren, maar omdat de productie moet blijven draaien en het risico bestaat dat die aan banden zal worden gelegd, zullen we ons scharen achter de idee dat technologie moreel neutraal is. Dat is althans de gedachtegang waarvan ik vermoed dat sommige ondernemers die delen.

Zou het niet kunnen dat zulk economisch belang ook aan de basis lag van ‘Guns don’t kill people, people kill people!’, de slogan waarmee ik dit hoofdstuk inleidde? Die wordt doorgaans als volgt uitgelegd. Er bestaat een relatie tussen wapens en dingen die we moreel afkeuren. Alleen bestaat die relatie omdat er mensen zijn die wapens in slechte zin gebruiken. Het kwade zit niet in het wapen zelf. Wapens zijn neutraal; in die technologie zitten geen moreel foute waarden verankerd. Het gevolg is dat je het gebruik van wapens moet verbieden of minstens reguleren, de ontwikkeling moet je uitsluitend aan de markt overlaten. Ik laat nu in het midden of dat laatste steek houdt. Zeker is in elk geval dat de uitleg achter ‘Guns don’t kill people, people kill people!’ vaak niet of nauwelijks beargumenteerd wordt. Verder is het in deze context ook relevant om in herinnering te

brengen dat de slogan voor het eerst werd gedrukt in *The Manufacturer*, een blad voor fabriekseigenaars.

Om vergissingen te vermijden wil ik wel nog het volgende onderstrepen. Als het klopt dat onder de neutraliteitsthese geen of niet uitsluitend theoretische argumenten liggen, maar wel (ook) al te menselijke beweegredenen, dan is dat nog geen reden om de neutraliteitsthese te verwerpen. Een theorie staat of valt niet afhankelijk van het soort motieven waaraan de theorie ontspringt. Uitspraken zijn niet onwaar omdat ze voortkomen uit een al dan niet radicaal linkse of rechtse agenda, omdat er geld te winnen of verliezen valt, of omdat men jaloers of afgunstig is. Een theorie staat of valt omdat er respectievelijk niets of iets schort aan de theorie zélf. Ja, een theorie kan fout zijn en het kan zijn dat men de theorie verdedigt uit eigenbelang, maar de theorie is niet fout *omdat* die voortvloeit uit eigenbelang. Als je dat voor ogen houdt, dan wordt duidelijk dat ik daarnet naar belangen heb verwezen, niet om de neutraliteitsthese op losse schroeven te zetten, maar wel om te begrijpen. Meer in het bijzonder wil ik begrijpen waarom zulke bewering al zo lang bestaat en waarom die zo populair is bij computerwetenschappers, ondernemers, fabrikanten, ingenieurs en anderen die nauw zijn betrokken bij het ontwerpen en maken van technologie en AI. Mijn voorstel is dit: sommigen verdedigen de neutraliteitsthese, niet of niet zozeer omdat ze dat geloven, maar omdat ze daar belang bij hebben, omdat ze iets te winnen of te verliezen hebben, omdat iets op het spel staat.

MACHINES ZIJN GEEN STENEN

De blik op technologie kan vertroebeld zijn door minstens twee zaken: slechte argumenten en belangen, of een combinatie van beide. Tot slot wil ik nog een derde mogelijkheid naar voren schuiven. De verdediging van de neutraliteitsthese kan ook voortvloeien uit het dragen van een moderne bril om naar de werkelijkheid te kijken. Ik beroep me daarvoor op het denken van filosoof Bruno Latour.[40]

Het is bekend dat René Descartes zich in de zeventiende eeuw de vraag stelde of we wel mogen vertrouwen op onze waarnemingen en wiskundige kennis. Hij twijfelde aan alles, met als gevolg dat hij minstens van één ding zeker is, namelijk dat hij aan het twijfelen is, *ergo*: ik besta. Het is immers onmogelijk dat je twijfelt, maar tegelijk niet bestaat. Descartes' hele verdere denkbeweging bestaat erin om het zogeheten brugprobleem op te lossen. Dat wil zeggen: hij moet de kloof weten te dichten tussen zijn eigen bestaan en de wereld rondom hem, de buitenwereld. Want je mag dan wel zeker zijn dat jijzelf bestaat, bestaat er ook een wereld buiten jou? Het schema van zijn vraag is dus: hier, aan deze kant sta ik, en ik wil wel graag geloven dat dáár, aan de andere kant, een buitenwereld is, maar hoe kan ik dat met volle zekerheid te weten komen? Dat is de afstand die Descartes wilde overbruggen.

Volgens Latour is die tegenstelling tussen mens en wereld typerend voor een moderne denkwijze. Het is niet zo dat filosofen zich niet eerder de vraag hebben gesteld of de wereld bestaat, maar het feit dat Descartes de vraag zo expliciet in die bewoordingen formuleerde, is verre van toeval. Vanaf de moderniteit, ruwweg vanaf de zeventiende eeuw, ging men immers steeds meer in tegenstellingen denken. Niet alleen werd de mens tegenover de wereld geplaatst, ook natuur en cultuur, waarden en feiten, werden als tegenovergestelden gezien. Modern zijn, aldus Latour, betekent dat je de dingen tegenover elkaar plaatst en scherpe lijnen trekt, dat je binair denkt. Daartegenover staat een premodern wereldbeeld, waarin de dingen minder duidelijk zijn onderscheiden. De natuur wordt er niet gezien als iets dat tegenover de mens staat, maar als een ordening waarin de mens de plaats moet innemen die aan hem is voorbehouden. Neem bliksem en donder. Terwijl wij, modernen, die zaken in fysicalistische zin zouden begrijpen, worden die vanuit een premoderne optiek geïnterpreteerd als tekens, als signalen van Gods woede bijvoorbeeld.

Wil dat nu zeggen dat de moderne blik de juiste is? Nee, aldus Latour. Vanaf de moderniteit wordt weliswaar in tegenstellingen gedacht,

maar de werkelijkheid laat zich niet zo makkelijk in hokjes duwen. De moderne oppositie tussen binnen en buiten, feiten en waarden, natuur en cultuur is een illusie, een foutief verhaal dat we onszelf vertellen. *Nous n'avons jamais été modernes*, aldus de titel van Latours bekende boek uit 1991. Iets is nooit louter cultuur of natuur, binnen of buiten; alles is een hybride, zo meent Latour, zowel cultuur als natuur. Denk aan wetenschapper Robert Boyle. Om een goed zicht te krijgen op de natuur moest hij zijn vacuümpomp onder controle houden, beriep hij zich op getuigen en zocht hij steun bij de politieke overheden.

Ik laat hier in het midden of Latours visie klopt, of werkelijk *alles* een hybride is. Niettemin wil ik hier toch de volgende vraag opwerpen. Zou het niet kunnen dat de neutraliteitsthese een gevolg is van zo'n moderne zienswijze? Neemt men ten onrechte aan dat in technologieën geen morele waarden zitten, omdat men in de greep is van de moderne neiging om te zuiveren, om te scheiden wat met elkaar vervlochten is, om in binaire termen te denken? Is met andere woorden de bewering dat technologie waardeneutraal is zelf niet gebiast? Ik schuif die verklaring naar voren op basis van het volgende.

Vaak worden het organische en het anorganische uit elkaar gehouden. Onder de eerste categorie vallen planten, mensen en andere dieren. Daartegenover, dat is althans wat het moderne denkbeeld ons voorhoudt, staan niet-levende entiteiten als ijzer en stenen. Ik haal dat onderscheid aan, omdat de neutraliteitsthese in dat schema past. Allereerst drukt ook de neutraliteitsthese een tegenstelling uit, die tussen technologie en morele waarde. Verder bestaat technologie doorgaans uit dode stoffen (ijzer, staal), en is er een sterke band tussen morele waarden en de categorie van het leven (van mensen). Een bepaalde toestand is immers alleen een morele waarde voor mensen, en zulke waarde gaat doorgaans over het leven van personen of interpersoonlijke verhoudingen. Deze zaken vormen de aanleiding om de volgende mogelijke verklaring naar voren te schuiven. De verdediging van de bewering dat technologie waardevrij is, is gebaseerd op een modern denkschema, een achterliggend

kader dat technologie naast dode materie als stenen plaatst, en morele waarden bij het leven en het organische.

We weten intussen dat de neutraliteitsthese niet juist is; technologie kan neutraal zijn, maar dat is niet noodzakelijke zo. In termen van het moderne paradigma: we weten dat het onderscheid tussen leven, mensen en waarden enerzijds en anorganische, dode materie anderzijds niet volstaat om het brede spectrum aan technologieën te vatten. De voorbeelden die ik heb aangehaald – denk aan de bruggen van Moses of de telefoon van British Telecom – zijn niet te herleiden tot dode materie. Die artefacten zijn hybriden in de zin van Latour: ze liggen op de grens tussen het anorganische en het leven, een grens die door de modernen sterk, al te sterk, wordt getrokken. Niettemin zijn de aangestipte relaties, tussen morele waarde en leven bijvoorbeeld, een reden om te opperen dat de neutraliteitsthese een effect is, niet van argumenten of belangen, maar van een moderne blik op de wereld. De bewering dat technologie losstaat van morele waarden kan ook voortvloeien, zo luidt mijn voorstel, uit het rigoureus opsplitsen van de wereld in het leven enerzijds en dode materie anderzijds.

De zin van een filosofische blik

Wat is de relevantie hiervan? Dat is niet alleen de vraag die men dikwijls aan filosofen stelt, het is ook de vraag waarmee ik dit hoofdstuk wil afronden. Ik heb uitgebreid ingezoomd op de neutraliteitsthese en laten zien dat daar heel wat aan schort, maar is dat betoog ook nuttig voor de praktijk? Ik meen van wel. Mijn betoog is praktisch relevant in een aantal opzichten: voor de omgang met technologieën, de evaluatie van ontwerpers, en de gebruikers.

LOVEN EN PRIJZEN

Uit de inleiding weten we dat alle technologie een functie heeft, dat wil zeggen: er is geen technologie die niet is ontworpen met een doel

voor ogen. Die vaststelling brengt met zich mee dat iedere ontwerper op z'n minst kan worden geëvalueerd op grond van het al dan niet goed functioneren van de technologie. Gebruikt men het voorwerp op de juiste manier maar leidt dat niet tot het verwachte resultaat, dan wijzen we daarvoor in de richting van de ontwerper; als blijkt dat de technologie goed functioneert, dan zullen we hem of haar daarvoor loven. Zo'n beoordeling heeft op zich niets met ethiek te maken, want ze is van puur technische aard. Maar daarnaast kun je ook in juridische en dus niet-technische zin oordelen over ingenieurs, computerwetenschappers en anderen. We vinden dat een ontwerper tekortschiet wanneer hij of zij de relevante wetgeving niet kent die eigenlijk wel bekend zou moeten zijn, en leggen hem of haar boetes op wanneer blijkt dat de technologie in strijd is met de wet, met bijvoorbeeld de regels voor de verwerking van persoonsgegevens. Maar putten het technische en juridische oordeel de evaluatie van de ontwerper uit? Kunnen we AIontwikkelaars en anderen niet ook nog op een andere manier evalueren?

Hoewel sommigen wensen dat ethiek alleen aan bod komt tijdens de gebruiksfase volgt uit mijn betoog dat we ontwerpers ook *moreel* kunnen beoordelen. Wie namelijk een voorwerp doelbewust met een morele waarde uitrust, wordt daardoor meteen ook onderwerp van morele evaluatie. Is die waarde wenselijk, zoals bij welzijn of duurzaamheid, dan kun je daarvoor moreel worden geprezen; is ze laakbaar, zoals bij rassenscheiding, dan zal men je daar als ontwerper hard voor afkeuren. Verder kunnen ontwerpers ook worden gestraft of minstens misprezen wanneer blijkt dat zij een technologie niet met een morele waarde hebben uitgerust, terwijl ze dat eigenlijk wel hadden moeten doen. Concreet: je kunt een app ontwikkelen die wel goed werkt, terwijl die niet privacyproof is. In dat geval schiet je als ontwerper tekort, niet in technisch opzicht, maar wel in morele zin. Technologie en AI ontwikkelen is méér dan een technische zaak of méér dan het toepassen van wetenschappelijke kennis. Mijn analyse legt met andere woorden bloot dat we ontwerpers in drieërlei zin kunnen beoordelen. Naast niet-morele evaluaties (technisch en juridisch) is ook een morele evaluatie mogelijk.

Die drie oordelen kunnen overigens vanuit een bepaald opzicht aansluiten op elkaar. Een positieve morele evaluatie van de ontwerper kan zich op positieve technische en juridische oordelen enten. Ik kan een AI-systeem maken dat goed werkt, geen wetten overtreedt en dat bovendien is uitgerust met een moreel wenselijke waarde. Maar het kan ook anders. Stel, ik ben de ontwerper van de eerdergenoemde Parkway-bruggen in New York. Ik ontwerp die veel te lage artefacten, die uiteindelijk bewerkstelligen waarvoor ze werden bedacht: ze zorgen ervoor dat op het strand alleen personen met een witte huidskleur liggen. Het technisch oordeel over mijn werk als ingenieur is op dat ogenblik gunstig. Ik heb namelijk iets zodanig goed ontworpen dat het doet waarvoor het werd ontworpen. Niettemin is duidelijk dat we aan dat oordeel geen gelijkaardige morele evaluatie koppelen. Rassenscheiding is verwerpelijk en de bedoeling om dat te verwezenlijken, kunnen we in geen enkel geval positief beoordelen. De les die we daaruit moeten trekken, is de volgende. Het moreel evalueren van ontwerpers kan uitmonden in beweringen die zichzelf lijken tegen te spreken maar dat bij nader inzien niet doen. Het is mogelijk dat men van mij als ontwerper van de Parkway-bruggen zegt dat ik zowel goed als slecht ben. Hoewel dat een contradictie lijkt, is het dat niet, omdat met de gunstige evaluatie de *technische* evaluatie wordt bedoeld en met de ongunstige de *morele* evaluatie.

DE MORELE RELEVANTIE VAN KENNIS

Bij de bespreking van het zichtbaarheidsargument wees ik er al op: in technologie kun je geen morele waarden zien. Wel is het mogelijk dat je de waardegeladenheid opmerkt. Je ziet dat het ontwerp is afgestemd op een morele waarde, of dat bepaalde materialen zijn gekozen met het oog op het realiseren van een moreel wenselijke situatie. Toch is het niet uitgesloten dat je helemaal niet ziet dat er een morele waarde in de technologie zit verankerd, of dat je dat nauwelijks ziet, bijvoorbeeld omdat de verwijzing naar een morele waarde in de details van het ontwerp zit. Denk opnieuw aan de Parkway-bruggen. Die zijn zonder twijfel geladen met een moreel verwerpelijke waarde,

maar als je alleen naar die bruggen kijkt, dan kun je uit het ontwerp zelf niet die waardegeladenheid afleiden.

Als mijn betoog juist is, kan het dus gebeuren dat je voor een technologie kiest zonder te weten dat er aan die technologie een morele waarde kleeft. Dat is in de eerste plaats op zich onwenselijk, want iedereen wil keuzes maken die zijn gebaseerd op alle relevante informatie. Toch hoeft dat niet gepaard te gaan met schadelijke gevolgen. De morele waarde waarvan je je niet bewust bent, kan immers goed zijn. Erger is het wanneer wordt gekozen voor technologie waarvan je niet weet of kunt zien dat die met een morele waarde is geladen, terwijl die waarde wel verwerpelijk is. Het is in dat geval niet zeker dat men je verantwoordelijk zal houden voor de gevolgen van die keuze, maar wel zeker is dat het gebruik van die technologie zoals het is bedoeld zeer schadelijk zal zijn. Daarbij komt dat die gevolgen van lange duur kunnen zijn. Keuzes voor technologieën kunnen immers niet altijd snel teniet worden gedaan, bijvoorbeeld omdat ze materieel of ruimtelijk stevig zijn verankerd, zoals in het geval van de Parkway-bruggen. Alleen al daarom is het relevant om te weten dat technologie niet altijd waardevrij is, zoals iemand als Zuckerberg ten onrechte beweert. En daarom is het ook zeer aan te raden om eerst, voorafgaand aan de keuze voor of het gebruik van de technologie, zo veel mogelijk informatie in te winnen. Pas als die kennisvoorwaarde is vervuld, is de keuze voor de uitrol van een app of het implementeren van een AI-systeem echt moreel verantwoord.

ZORG DRAGEN VOOR DE DINGEN

Aan het begin van dit hoofdstuk heb ik erop gewezen dat we de term ‘waarde’ in het dagelijks leven op twee manieren gebruiken: iets kan een waarde *zijn* of iets kan waarde *hebben*. Als een voorwerp waarde op zich heeft, dan is dat meteen ook een reden voor een positieve houding ten aanzien van dat voorwerp, zo stipte ik aan. Ik zorg bijvoorbeeld goed voor mijn fiets, omdat ik die mooi vind. Maar wil dat

zeggen dat elke positieve houding voortvloeit uit het feit dat iets waarde *heeft*? Kan een positieve houding ten aanzien van een technologie ook op iets anders zijn gebaseerd?

Die laatste vraag moet je bevestigend beantwoorden. Als ik weet dat een technologie is geladen met een moreel wenselijke waarde, dan is dat een reden om positief te staan tegenover die technologie. Als men mij bijvoorbeeld een AI-systeem toont dat met *fairness* is geladen, dan is dat een goede reden om er lovend over te zijn, het aan te prijzen of er zorg voor te dragen. Het kan uiteraard zijn dat ik dat systeem ook koester of prijs om andere redenen dan het feit dat het met een morele waarde is geladen, bijvoorbeeld omdat het in esthetisch opzicht aantrekkelijk is. In dat geval is de waardegeladenheid een extra reden voor een positieve houding. Maar zelfs als blijkt dat een technologie niet of niet goed functioneert en dat die verder geen enkele andere waarde heeft, dan wil dat nog niet zeggen, zo volgt uit mijn verhaal, dat je alleen negatief of neutraal kunt staan tegenover die technologie. Dat voorwerp kan immers zijn geladen met een morele waarde als rechtvaardigheid of gelijkheid, en dat op zich kan al voldoende zijn om het te waarderen, verzorgen of aan te prijzen.

MORELE MACHINES

Ik borduur tot slot van dit hoofdstuk nog even verder op dat laatste punt. Dat een technologie is geladen met een moreel wenselijke waarde is een reden om er positief tegenover te staan, maar kunnen we daar niet nog sterkere gevolgen aan vastknopen? Volgen uit mijn betoog ook geen morele plichten? Ik denk van wel. Wanneer dingen met een morele waarde zijn geladen hebben we de plicht om het gebruik ervan op z’n minst ernstig te overwegen. Om dat toe te lichten, ga ik eerst even in op de band tussen technologie en plichten.

Dat mijn fiets, auto of horloge mooi is – lees: op zich waarde heeft –, is wel een reden om ervoor te zorgen, maar dat feit op zich verplicht me verder tot niets. Toch kunnen we ook verplichtingen hebben ten aanzien van artefacten. Vaak zijn die plichten gebaseerd op

eigendom, op het feit dat iets iemand toebehoort. Omdat die fiets van mij is, moet degene die mijn fiets gebruikt er zorg voor dragen; omdat de wagen van mijn buur is, ben ik verplicht mijn buur te vragen of ik de wagen mag gebruiken om er mee naar de winkel te rijden. Daarnaast kan het zijn dat ik ervoor moet zorgen, omdat ik dat heb beloofd aan de eigenaar. Plichten kunnen voortvloeien uit eigendom, maar ook uit een belofte. Beide plichten zitten duidelijk niet in het ding zelf verankerd, maar zijn een afgeleide ervan. Ze vloeien voort uit iets anders dan de technologie: eigendom of belofte. In zekere zin kun je daarom zeggen dat het in deze twee gevallen gaat over een plicht, niet zozeer ten aanzien van het voorwerp zelf, maar ten aanzien van een persoon: de eigenaar van het ding en diegene aan wie de belofte werd gedaan.

Plichten kunnen echter ook op andere zaken dan eigendom en belofte zijn gestoeld. Denk aan wat we niet mogen doen en wat we moeten doen ten aanzien van mensen en dieren die geen mensen zijn. Het kan wel mijn plicht zijn om voor de kat van mijn buurman te zorgen omdat ik hem dat heb beloofd, maar het gaat verder dan die belofte. Ik heb ook een plicht ten aanzien van de kat zelf, los van mijn belofte. Hetzelfde geldt vanzelfsprekend voor onze plichten tegenover andere mensen. Ik moet het huilende kind van mijn buren troosten, niet alleen omdat ik beloofde een oogje in het zeil te houden, maar ook en vooral omwille van het kind zelf. In deze en andere gevallen is de plicht gegrond in eigenschappen, bijvoorbeeld het vermogen om pijn te voelen. Dat vermogen brengt de plicht met zich mee om een mens of kat geen pijn te doen, of om te proberen de pijn weg te nemen.

Straks, in het volgende hoofdstuk, ga ik daar nog dieper op in, maar hier volstaat het te zeggen dat machines en robots momenteel niet de eigenschappen hebben waarop doorgaans onze plichten ten aanzien van mensen en andere dieren zijn gebaseerd: het vermogen tot pijn voelen en (zelf)bewustzijn. Vandaar: wij zijn niets verplicht aan de dingen. Mijn analyse uit dit hoofdstuk verandert daar niets aan. Uit het feit dat een technologie met een morele waarde is geladen, kun je niet afleiden dat je die technologie zelf iets verplicht bent. Anders gezegd:

dat technologie rechten heeft, kun je niet concluderen op basis van mijn verhaal in dit hoofdstuk. Natuurlijk, aan waardegeladen technologieën zijn wel verplichtingen verbonden, maar dat zijn plichten die zijn gericht op de eigenaar en niet op het artefact zelf – dat was het punt dat ik daarnet maakte. En ja, het valt niet uit te sluiten dat we in de nabije of verre toekomst zullen vinden dat we waardegeladen robots iets verplicht zijn, dat ze rechten hebben. Maar dat zal dan wellicht niet zijn *omdat* ze met morele waarde zijn geladen.

En toch stel ik het volgende voor: waardegeladen technologieën brengen bijzondere plichten met zich mee. Dat zijn vanzelfsprekend geen plichten ten aanzien van die technologie zelf, maar het zijn wel plichten die voortvloeien uit eigenschappen van een technologie, te weten: het feit dat er een morele waarde in de technologie is gepompt. Ik verklaar me nader.[41]

De morele psychologie van mensen is begrensd. We hebben de neiging om ons sneller, meer en zelfs enkel te bekommeren om soortgenoten, en bovendien om mensen die zich in tijd en ruimte dicht bij ons bevinden. Er is de zogeheten negativiteitsbias: als we iemand niet of nauwelijks kennen, dan merken we eerder de onwenselijke dan wenselijke eigenschappen op. We denken dat we moreel oordelen op basis van sterke argumenten, dat we genuanceerd moreel reflecteren en dat we makkelijk fouten toegeven. Recent onderzoek suggereert dat dit een verhaal is dat we onszelf graag vertellen, terwijl daar eigenlijk niet zoveel van klopt: goed moreel redeneren komt minder vaak voor dan we denken. Ethiek vloeit meestal voort uit intuïties, emotionele reacties en adhocrationalisaties.

In het verleden en tot op heden hebben we ons vooral op opvoeding en religie gebaseerd om ons moreel te verbeteren, om morele tekorten als de negativiteitsbias te corrigeren of weg te werken. Tegenwoordig zouden we ons ook kunnen richten tot biomedische interventies. De bètablokker propranolol bijvoorbeeld helpt niet alleen tegen een hoge bloeddruk, maar zou ook raciale vooroordelen doen

verminderen. Een andere mogelijkheid is dat we ons beroepen op AI om onze morele beperkingen te overstijgen. Zie hier een aantal opties.

Het aantal uren dat we slapen, beïnvloedt onze morele beslissingen. Dat is wat onderzoek naar het oordeelsvermogen van soldaten met een slaaptekort ons leert. Hetzelfde lijkt ook het geval te zijn met honger – ik wees er al eerder op. Rechters oordelen strenger naarmate zij minder hebben gegeten. En ook omgevingsinvloeden beïnvloeden de ethiek. Wanneer mensen zich in ambigue situaties bevinden, dan reageren ze vijandiger bij hoge temperaturen en in ruimtes waar veel volk en geluid is. AI-systemen zouden in zulke contexten kunnen worden ingezet als feedbacksystemen. Ze verzamelen fysiologische data en gegevens over onze omgeving, analyseren die op basis van grote datasets uit het verleden in het licht van een zo optimaal mogelijk moreel functioneren, en zenden vervolgens een signaal naar onze mobiele telefoon op het moment dat dat functioneren dreigt te worden verstoord.

Het gebruik van AI voor ethiek hoeft echter niet beperkt te worden tot louter monitoren. Zulke technologie kan ook een actievere morele rol vervullen. Neem het volgende. Veel mensen hebben de intentie om geld te schenken aan goede doelen. Niettemin weten we dat mensen minder geven dan ze denken. We zouden AI-systemen kunnen gebruiken om te onderzoeken wat de oorzaak van die kloof is en hoe het probleem zou kunnen worden opgelost. Of denk aan *fairness* en bias. Uit een studie blijkt dat kennis van de seksuele identiteit van een persoon ons oordeel over die persoon beïnvloedt. Het pianospel van een vrouw werd hoog ingeschat als de beoordelaars de pianiste niet zagen. Het oordeel lag significant lager als ze de vrouw wél zagen. Om zulke moreel onwenselijke situaties de wereld uit te helpen, zouden we AI-systemen kunnen inschakelen die mensen trainen in het nemen van genderneutrale beslissingen.

Stel nu dat zulke technologieën beschikbaar zijn, dat we er gebruik van kunnen maken. Dan zijn dat duidelijk voorbeelden van waardegeladen technologieën. Ze dragen de belofte in zich dat er

meer morele waarden worden gerealiseerd, dat er pakweg minder bias of meer gelijkheid is. Hebben we dan niet op z'n minst de plicht om het gebruik van die technologieën serieus te overwegen? Ik beweer niet dat ze *moeten* worden toegelaten, laat staan dat ze verplicht zouden moeten zijn, maar als die systemen zulke moreel wenselijke effecten beloven, is het dan niet onze plicht om grondig na te denken en te debatteren over het al dan niet implementeren van die zaken? Dat voorstel druist in tegen onze morele intuïtie. We vinden doorgaans dat we morele training en verbetering aan mensen moeten overlaten (ouders en leerkrachten). En natuurlijk besef ik dat een significant aantal mensen een eerder technopessimistische houding heeft. Een technologie als de microgolfoven bijvoorbeeld maakte aanvankelijk erg veel negatieve reacties los en we weten dat ongeveer 25% van de Belgische bevolking wantrouwig en negatief tegenover AI staat. Verder ben ik vanzelfsprekend ook op de hoogte van de risico's die zijn verbonden aan zulke vernieuwende technologieën – ik denk bijvoorbeeld aan bezorgdheid rondom privacy of *surveillance* – straks meer daarover. Maar moeten zulke bezorgdheden per se uitmonden in het aan de kant schuiven van die innovatieve opties? Moeten we niet op z'n minst proberen om over onze diepgewortelde intuïties en attitudes heen te stappen? Is het niet onze taak om op z'n minst serieus na te denken over de morele mogelijkheden van AI?

Ter afronding

Als het over technologie in het algemeen en AI in het bijzonder gaat, zijn er verschillende overtuigingen die al even meegaan en ook wijdverspreid zijn onder AI-ontwikkelaars, filosofen, politici, ingenieurs en anderen. Een van die populaire beweringen is dat technologie neutraal is, net zoals rechters dat (moeten) zijn. Je kunt die bewering op verschillende manieren interpreteren. Het kan betekenen dat aan technologie stereotypen en normen kleven, maar de bekendste interpretatie luidt dat technologie losstaat van morele waarden als

fairness, privacy en inclusiviteit. Apps en andere dingen kunnen wel bijdragen aan het realiseren van die waarden, maar, zo luidt de overtuiging, op zich staat iedere technologie los van een morele waarde. Er zijn verschillende redenen denkbaar waarom mensen zulke overtuiging hebben en er bestaan verschillende argumenten die de bewering pogen te ondersteunen, maar in het voorbije hoofdstuk heb ik laten zien dat het onjuist is dat technologie per definitie losstaat van zoiets als rechtvaardigheid. Zeker, niet elke technologie is geladen met een morele waarde, maar als je weet dat iets is ontworpen met méér privacy voor ogen of met als doel om niemands privacy te schenden, dan kun je niet meer volhouden dat aan geen enkele technologie een morele waarde kleeft. Dat is zo voor alle soorten van technologie, van basale dingen als een schaar tot slimme hypergeavanceerde dingen als zelfrijdende wagens.

You just a barcode

Childish Gambino in *This is America*

2

De zeven hoofdzonden van AI

De eerste industriële revolutie steunde op stoom, de tweede op elektriciteit en de derde op computers en het internet. Sinds een aantal jaren bevinden we ons in een nieuwe revolutie. Dat heeft veel, zo niet bijna alles, te maken met wat het internet ons leverde: data, *big data*, ontzettend veel data over ontiegelijk veel zaken uit ons online- en offlineleven. Om een idee te geven: naar verluidt zou 90% van de gegevens die ons nu ter beschikking staan in de afgelopen vijf jaar gegenereerd zijn. Data zijn het nieuwe olie, zo luidt het cliché sinds 2006, toen wiskundige Clive Humby de uitdrukking voor het eerst gebruikte. En de Verenigde Staten zijn het nieuwe SaoediArabië. Welke systemen worden met die data volgetankt?

Als we Shoshana Zuboff mogen geloven, auteur van de bestseller *The Age of Surveillance Capitalism* uit 2019, dan dienen data de ongebreidelde winsthonger van bedrijven, en dan vooral de honger van de *Big Five*: Amazon, Apple, Facebook, Google en Microsoft. Data dienen het kapitalistische systeem, aldus Zuboff, en met name de techbedrijven van de eenentwintigste eeuw: bedrijven zonder grote fabrieken, zonder hoge materiaal- en productiekosten. Die bewering moet gecorrigeerd worden. Niet elk gebruik of iedere ontwikkeling van AI is een uiting van geldgewin, net zoals niet alles van het kapitalisme met AI te maken heeft. Natuurlijk verdienen techgiganten miljarden aan de informatie over ons doen en laten, en natuurlijk ziet men privacy vaak als een rem op economische groei. Maar een minder provocerend en juister antwoord dan dat van Zuboff is dat die gegevens slimme technologieën voeden, AIsystemen. Al moet ik er voor de volledigheid wel aan toevoegen dat sommige AI, met name expertsystemen, niet gestoeld zijn op training met talloze data – ik

stipte het eerder al aan in de inleiding –, en ook dat andere dingen dan AI-systemen gebaat kunnen zijn bij veel data.

Ons leven is vandaag de dag al voor een flink stuk vervlochten met zulke slimme systemen. Je kunt dat betreuren, je kunt je er blauw aan ergeren dat 'AI' vandaag een *buzzword* is, feit is wel dat weinig of geen domein aan AI ontsnapt, en dat dat waarschijnlijk niet snel zal veranderen. We leven in een democratie en meritocratie, maar evolueren ook meer en meer in de richting van een algocratie, een samenleving die in een niet geringe mate wordt bestuurd door de rekenkracht van algoritmen. Laat ik daarom beginnen, om op te warmen, met een kleine greep uit de verschillende contexten waarin AI vandaag wordt gebruikt.

DE ALGORITMISCHE SAMENLEVING

De talloze gegevens die we achterlaten – door eenvoudigweg onze smartphone te openen, websites te bezoeken, foto's te posten en te 'liken' – worden door onder meer Facebook en Google verzameld en verkocht aan vooral bedrijven die er op hun beurt verder mee aan de slag gaan. Met behulp van AI-systemen krijgen die bedrijven een goed zicht op wie je bent: je seksuele oriëntatie, levensbeschouwelijke voorkeur, verslavingen, relatiestatus, emotionele status, etniciteit, intelligentie, leeftijd, enzovoort. Het systeem voorspelt welk product je zou kunnen interesseren en ten slotte stuurt het gepersonaliseerde advertenties met militaire precisie naar jouw profiel op Facebook, Instagram of TikTok – advertenties waarvoor men uiteraard eerst heeft betaald aan de platformen waarop ze worden gepost. Het doel is dat je shopt, dat je boeken, reizen, kleding, eten, auto's, medicijnen of muziek koopt. Deze en andere zaken worden in het docudrama *The Social Dilemma* uit 2020 van Jeff Orlowski glashelder uitgelegd.

Targeted advertising, dat voor ongeveer 70% via Facebook en Google verloopt, is naast een commercieel ook een politiek instrument. Hoewel het in België en Nederland – en in een groot deel van Europa – niet totaal onbekend is, is het in onze contreien lang niet zo populair

als in de Verenigde Staten. Barack Obama was zowat de eerste politicus die op grote schaal gebruik maakte van data in de aanloop naar de presidentsverkiezingen. Maar voor misschien wel het bekendste en beruchtste voorbeeld van het politieke gebruik van AI moeten we terug naar de zogeheten Facebookverkiezingen in 2016. In het kort komt het op het volgende neer – wie een uitgebreidere versie wil, kan onder meer teruggrijpen naar de documentaire *The Great Hack* uit 2019 van Jehane Nouiaim en Karim Amer. Cambridge Analytica is een Brits-Amerikaans bedrijf dat tussen 2014 en 2016 samenwerkte met presidentskandidaat Ted Cruz. Toen die uit de race lag, ging het bedrijf in zee met Trump. Het verzamelde op geheel onfrisse wijze informatie van 87 miljoen Facebookgebruikers via datawetenschapper Aleksandr Kogan. Die had de enquête-app 'This Is Your Digital Life' ontwikkeld, rolde die uit op Facebook, en gebruikte niet alleen de informatie van degenen die de enquête hadden ingevuld maar ook van alle vrienden. De AIsystemen van Cambridge Analytica analyseerden die data, voorspelden op basis daarvan wiens stem mogelijk interessant zou zijn, en verstuurden vervolgens doelgericht advertenties naar de gebruikers van de sociale media in de hoop dat er veelvuldig zou worden gestemd op de republikeinse vastgoedondernemer. Tussen haken: hoewel Theresa Hong, de directeur van Trumps campagne, dacht dat Trump zonder Facebook nooit de verkiezingen had gewonnen, is het bijlange niet zeker dat de campagne ook effectief was. Uit onderzoek blijkt immers dat politieke advertising via AI een effect van slechts 2% heeft, wat aan de andere kant wel een verschil kan maken in de zogeheten *swingstates*.[42] Hetzelfde geldt overigens voor commerciële advertenties. Het valt te betwijfelen, zo wijst onderzoek uit, dat door AI gestuurde advertenties tot meer consumeren leidt.[43]

Een ander voorbeeld van hoe AI vandaag de dag wordt gebruikt komt van I-Care, een bedrijf met vestigingen in onder meer de Belgische steden Bergen en Heverlee. De corebusiness van het bedrijf is preventief onderhoud. Het voorspelt wanneer een machine versleten

is, en dus wanneer vervanging of minstens onderhoud nodig zal zijn. I-Care ontwikkelde daarvoor nieuwe technologie.

Draadloze sensoren verzamelen data over de temperatuur, de trillingen en het geluid van de machines. Op die informatie wordt vervolgens AI losgelaten, die de data analyseert, en voorspelt of en wanneer de machine zal verslijten. Het is duidelijk dat zulke innovatie erg relevant is voor de industrie. Het voorkomt dat machines niet tijdig worden onderzocht en daardoor kapot gaan, wat onder meer lichamelijke letsels of milieuschade kan veroorzaken. Daarnaast verhoogt het de efficiëntie. Onderhoud kost tijd en geld, terwijl een onderhoud dat werd gepland zonder gebruik te maken van AI niet per se nodig is.

AI speelt ook een rol in de geneeskunde. De behandeling van bijvoorbeeld halskanker is een erg delicaat gebeuren. De bestraling moet gericht gebeuren om de onbeschadigde delen rond de plaats van de bestraling te sparen. Om die reden is het nodig dat die heel precies worden aangeduid op de scan. Tot nog niet zo lang geleden waren het uitsluitend radiotherapeut-oncologen die dat deden. Het probleem was niet alleen dat dit zeer tijdsintensief is, vraag je aan twee mensen om dezelfde plaats aan te duiden op een scan, dan is de kans ook reëel dat daar verschillen op zitten. Sinds kort doet men daarom een beroep op AI-systemen, die werden getraind met tal van beelden om patronen te herkennen. De resultaten zijn verbluffend. Wanneer de radiotherapeut-oncologen gebruik maken van de nieuwe technologie zijn hun interventies zo'n 30% procent accurater én sneller dan wanneer ze de plekken op de beelden aanduidden zonder AI. Dat komt niet alleen de kankerbehandeling zelf ten goede, daardoor komt ook extra tijd vrij voor andere zorg.

AI wordt verder ook gebruikt voor de verwerking van natuurlijke taal (denk aan interactieve chatbots), voor persoonsherkenning (*facial recognition*), door sportclubs, om je route te berekenen in Google Maps, door banken om na te gaan of je kredietwaardig bent, om je verzekeringspremie te bepalen, om het risico op recidive in te

schatten, bij virtuele assistenten (denk aan Alexa en Siri van respectievelijk Amazon en Apple) of om teksten te vertalen (Google Translate en DeepL). Kortom, het aantal AI-toepassingen is niet gering en het heeft er alle schijn van dat het toepassingsgebied niet gauw zal krimpen, integendeel. Om een intussen bekende uitdrukking te gebruiken: dit is de zomer van AI.

DE TANDENBORSTEL VAN ORAL-B

Gezien de wijde verspreiding van AI in de korte tijdspanne van een aantal decennia is het niet verwonderlijk dat over die technologie sterke overtuigingen bestaan. Er zijn goede redenen om kritisch tegenover die overtuigingen te staan.

Sommige van die overtuigingen zijn erg optimistisch. Neem bijvoorbeeld de idee dat AI de oplossing is voor veel en misschien zelfs voor alle problemen. Het lijdt geen twijfel dat er goede redenen zijn om erg positief te zijn over AI-toepassingen, bijvoorbeeld over zelfrijdende auto's. Volgens de Wereldgezondheidsorganisatie sterven wereldwijd jaarlijks 1,35 miljoen mensen ten gevolge van verkeersongelukken, waarvan meer dan de helft door menselijke fouten. Zelfrijdende auto's zijn in dat opzicht beloftevol. Volgens schattingen zouden zij het aantal ongelukken en doden significant verminderen. Maar dat wil niet zeggen dat zij *geen* ellende meer zouden veroorzaken. Slimme technologieën zijn daarnaast ook wenselijk omdat ze helpen om ziekten te bestrijden, om een potentiële pandemie te detecteren, om de wereld veiliger te maken, om het klimaatprobleem aan te pakken of om saaie taken over te nemen. Toch kampen AI-systemen met verschillende problemen, *morele* problemen bijvoorbeeld – straks meer daarover. Bovendien is het naïef en zelfs gevaarlijk te geloven dat AI-systemen volstaan om het klimaatprobleem of armoedeprobleem op te lossen.

Een andere sterke bewering heeft voor sommigen een optimistisch karakter, voor anderen is ze dan weer een reden tot pessimisme. Ze luidt dat we niet zo ver meer verwijderd zijn van wat ik in de inleiding

van het boek *superintelligence* heb genoemd, naar de titel van de bekende studie van filosoof Nick Bostrom uit 2014. Binnen afzienbare tijd zouden artificiële entiteiten worden ontworpen die, anders dan vandaag het geval is, niet enkel specifieke taken verrichten zoals de vertaling van teksten. Zulke nieuwe hypergeavanceerde systemen zouden meerdere opdrachten kunnen combineren, onze intelligentie verregaand overtreffen en bewustzijn hebben. Het valt niet uit te sluiten dat zulke entiteiten ooit zullen bestaan. Bovendien is de voorspelling niet geheel bij de haren getrokken. AI-systemen, net zoals dieren die geen mensen zijn trouwens, zijn nu al in veel zaken beter dan mensen: ze hebben een beter geheugen en herkennen sneller patronen. Daartegenover staat dan weer het volgende. Er is onder AI-onderzoekers weinig of geen consensus over wat nu precies wel en niet de definitie van intelligentie is, maar er is wel eensgezindheid over dit: de komst van *superintelligence* laat nog even, en wellicht nog erg lang, op zich wachten.

MOVE FAST AND BREAK THINGS

Er is daarnaast nog een andere bewering die qua populariteit niet onderdoet voor de twee vorige. Die luidt dat AI disruptieve technologie is. Slimme technologieën zijn in het licht van de lange geschiedenis van spitsvondige uitvindingen niet enkel nieuw, zij zouden ook bestaande domeinen ontwrichten. Dat is althans wat vaak wordt beweerd over AI. Laat ik dat ‘de disruptiethese’ noemen, de stelling die vanaf nu mijn aandacht zal hebben en die samen met onder meer de neutraliteitsthese een populaire bewering is die men vaak te pas en te onpas aanhaalt.

Toegegeven, die stelling klinkt geloofwaardig, maar klopt ze ook? Is ook zij een mythe, net zoals de populaire neutraliteitsthese niet juist is? De disruptiethese ligt in ieder geval in het verlengde van de stelregel van Facebook: *move fast and break things*. Verder is het ook zo dat tal van bedrijven als disruptief worden gezien. Netflix ontwrichtte het bestaande televisielandschap, Airbnb de hotelwereld,

Tesla het autolandschap en Uber de taxisector, aldus sommigen. En hetzelfde wordt ook gezegd over 3D-printen en nanotechnologie. Maar zijn naast bedrijven, 3D-printen en nanotechnologieën ook AIsystemen disruptief? Blaast AI ook bestaande domeinen geheel nieuw leven in of zelfs van de kaart?

Anders dan de idee dat technologie neutraal is, gaat de stelling dat AI disruptief is nog niet zover terug in de tijd. Dat is ook niet verwonderlijk. Het onderzoek rond AI kwam pas van de grond rond het midden van de vorige eeuw, en de AI-toepassingen die we nu kennen, doorgaans gebaseerd op *machine learning* zijn pas sinds een decennium commercieel goed doorgebroken. Niettemin vind je de stelling vandaag op tal van plaatsen terug, bij zowel zij die positief staan tegenover AI, maar ook bij hen die eerder kritisch zijn. Sla een van hun boeken over AI open en de kans is reëel dat men het omschrijft als een disruptieve technologie. Neem *Digitalis* uit 2018 van Thierry Geerts, directeur van Google in België en Luxemburg. Daarin staat bijvoorbeeld het volgende: 'In elk geval is het overduidelijk dat de huidige digitale revolutie, die een industriële en een culturele revolutie combineert, onze hele leefwereld grondig wijzigt en alles op zijn kop zet. Veel mensen spreken dan ook over "disruptie": de technologische revolutie ontwricht de oude wereld en doet die uit elkaar vallen.'[44] Of neem *The Hype Machine* van wetenschapper en ondernemer Sinan Aral uit 2020. Hij schrijft over sociale media en argumenteert dat die tal van domeinen ontwrichten. De ondertitel luidt *How Social Media Disrupts Our Elections, Our Economy and Our Health – and How we Must Adapt.* Filosoof Stiegler publiceerde in 2016 een boek waarvan de titel als *The Age of Disruption. Technology and Madness in Computational Capitalism* werd vertaald. De disruptiethese is ten slotte ook aanwezig in het boek van de al eerder vermelde Zuboff, bijvoorbeeld wanneer ze het volgende schrijft: '(…) de winnende hand in het kapitalisme gaat over dingen opblazen met nieuwe technologieën.'[45]

Er zijn redenen om sceptisch te staan tegenover de disruptiethese. Over veel AI-systemen lijkt het bijzonder moeilijk te zeggen dat ze een ontwrichtend karakter hebben. Ik denk aan de Genius X tandenborstel van Oral-B. De borstel is uitgerust met sensoren die data over je mondhygiëne doorsturen naar een app, die vervolgens de gegevens analyseert en je laat weten of, waar en hoe je verder nog moet borstelen. Of neem het bericht dat in november 2020 werd verspreid door Microsoft. De softwaregigant maakte bekend dat het een AI-systeem ontwierp voor twee bezinepompstations, één in Thailand, een ander in Singapore. De software zoekt naar tekenen van onveilig gedrag, zoals bijvoorbeeld het roken van een sigaret aan een pomp. Vangt het zo'n signaal op, dan informeert het AIsysteem het personeel, dat er moet voor zorgen dat de sigaret wordt gedoofd. Het doel van AI is hier om ervoor te zorgen dat het station niet in vlammen opgaat. Dergelijke ontwikkeling is waardevol, al lijkt het niet meteen een disruptieve technologie te zijn. En zou je van een AI-tandenborstel beweren dat die disruptief is?

Aan de andere kant lijkt de disruptiethese ook niet vergezocht, bijvoorbeeld op grond van de geschiedenis van technologie. Het lijdt weinig tot geen twijfel dat in het verleden technologieën die geen AI waren meer dan vernieuwend waren. Denk aan het wiel, dat 3500 jaar voor onze jaartelling werd uitgevonden, en dat net als auto's grote maatschappelijke veranderingen heeft teweeggebracht. Hetzelfde geldt voor de drukpers in de vijftiende eeuw. Als zulke technologieën ontwrichtende effecten konden hebben, waarom zou dat dan niet het geval zijn voor AI-systemen?

Met zo'n vraag raak ik aan de kern van dit hoofdstuk, dat zich toespitst op AI. Net zoals de andere hoofdstukken beweegt het zich op drie sporen – de eerste twee zijn de belangrijkste. Om te beginnen zal ik op verschillende plaatsen simpelweg uitleggen en verhelderen. Wat betekent de disruptiethese wel en niet? Wat zijn bias en morele verantwoordelijkheid? Bestaan daar verschillende ideeën over, net zoals we verschillende opvattingen van *fairness* kunnen onderscheiden? En wanneer kunnen we iemand wel en niet

verantwoordelijk houden? Daarnaast zal ik, ten tweede, ook stelling nemen en argumenteren. Hebben AI-systemen op moreel vlak een ontwrichtend effect? Zo ja, waarom? Aan de hand van onder meer een grondige analyse van een casus – namelijk autonome beslissingssystemen – zal ik argumenteren voor een conservatieve stellingname: AI creëert geen volstrekt *nieuwe* morele problemen en sluit aan bij de bestaande morele praktijk. Naast beschrijven en evalueren, buig ik me tot slot ook kort over de relevantie van mijn redenering. Waarom zouden we moeten weten of technologie in moreel opzicht disruptief is?

Disruptieve technologie

De uitdrukking ‘disruptieve technologie’ werd gemunt door Clayton M. Christensen en voor het eerst gebruikt in 1995. Hij verwijst ermee naar technologieën die de markt door elkaar schudden. Een technologie kan wel bijzonder populair zijn, maar op een bepaald moment worden nieuwe producten uitgevonden die een gat in de markt blijken te zijn. Het gevolg is dat de interesse voor de oude technologie verdwijnt en dat de innovatieve technologie, de disruptor, *the next big thing* wordt.

Alle disruptieve technologieën zijn normaal gesproken in een of ander opzicht nieuw, maar sommige technologieën zijn nieuw zonder disruptief te zijn. Kijk naar transporttechnologie. Eind negentiende eeuw verschenen de eerste auto’s op de markt. Die waren vernieuwend omdat mensen zich tot dan toe doorgaans met behulp van paarden verplaatsten. Maar die innovatie was niet disruptief: de auto’s waren duur, té duur om door de grote massa te worden gekocht. De Ford model T daarentegen, voor het eerst geproduceerd in 1908, was wel een disruptieve technologie. Dat was een van de eerste auto’s die in massaproductie werd genomen, waardoor de prijs sterk daalde en meer gezinnen zich zo’n voertuig konden aanschaffen. Het gevolg was dat de markt voor paardentransport implodeerde.

Dit is een voorbeeld van de oorspronkelijke interpretatie van de disruptiethese, een interpretatie die duidelijk een economisch karakter heeft. Hoewel die invulling vandaag de dag nog vaak wordt gebruikt, zal ik me op een andere concentreren. Ik focus me op de *morele* invulling van de disruptiethese. Dat vergt wat toelichting.

MORELE ONTWRICHTING

Ik keer even terug naar het vorige hoofdstuk en herinner eraan dat volgens de neutraliteitsthese aan technologie geen morele waarden kleven. Die these focust op de technologie zelf, en niet op de oorzakelijke keten waarvan technologie deel uitmaakt. Ik heb voorbeelden gegeven van theorieën over de oorzaak en het effect van technologie. Volgens sommigen, Marx en Zuboff bijvoorbeeld, is technologie een product van het kapitalisme; anderen menen dat de secularisering een effect van technologie is. Ook de disruptiethese moet je op die manier beschouwen. Ze gaat niet over technologie op zich, maar over het effect van technologie op zaken die zelf geen technologie zijn. Denk aan de oorspronkelijke betekenis van ‘disruptieve technologie’: die gaat over de invloed van technologie op de markt.

Het feit dat een technologie gevolgen heeft, volstaat echter niet om die disruptief te noemen. Een betere omschrijving is dat technologieën disruptief zijn wanneer hun effecten een bestaand domein ontwrichten, openbreken, op z’n kop zetten. Disruptief zijn is meer dan alleen een sterk effect hebben of een bepaalde invloed versterken. Het verwijst naar een breuk met en een onderbreking van een bepaalde stand van zaken. In die zin kun je zeggen dat de Ford model T disruptieve technologie was, maar dat kun je niet van de eerste wagens beweren.

Wil dat nu zeggen dat de disruptiethese *uitsluitend* over het domein van de economie kan gaan? Nee. Natuurlijk, de oorspronkelijke invulling is wel economisch van aard, en dat geldt ook voor een andere vaakgehoorde interpretatie van de disruptiethese, die luidt dat

technologie disruptief is omdat ze de arbeidsmarkt op z'n kop zet. Maar je kunt aan de disruptiethese ook een andere dan de economische betekenis toekennen. Dat is in ieder geval wat tal van wetenschappers, filosofen, opiniemakers en politici vaak doen. Ze hebben het over de seismische effecten van technologie op het juridische, economische of sportieve landschap. AI zou de manier waarop oorlogen worden uitgevochten op z'n kop zetten, de wijze waarop we werken, recht spreken, sport organiseren.

Ik zal niet nagaan of AI *alle* domeinen ontwricht, ik concentreer me op de ethiek. De bewering dat technologie disruptief is, vertaal ik daarom als de stelling dat AI-systemen de ethiek uit haar evenwicht brengen. Ik moet wel onmiddellijk toegeven dat die formulering eigenlijk ook nog niet erg veelzeggend is, net zoals overigens de bewering dat AI de politiek of sportwereld op z'n kop zet nog te vaag is. Ze moet nog worden verfijnd. Want wat precies wordt bedoeld met 'ethiek'? Verwijs je naar morele emoties als medelijden, of heb je het over gedragingen zoals iemand helpen die in nood verkeert? Een andere mogelijkheid is dat je met 'ethiek' doelt op een kader van waaruit je morele oordelen velt. Maar ook dan zijn er meerdere mogelijkheden. Je kunt opteren voor het kader waarin morele deugden als moed en loyaliteit centraal staan en waarover Aristoteles heeft geschreven, voor de benadering van filosoof Jeremy Bentham die focust op de gevolgen van handelingen, of voor het plichtethisch perspectief van Immanuel Kant.

Laat ik even kiezen voor de betekenis van ethiek als denkkader. De stelling luidt dan dat AI onze morele kijk op de wereld op z'n kop heeft gezet. Terwijl we vroeger voor het moreel beoordelen van mensen eerder rekening hielden met de al dan niet goede bedoelingen van mensen of hun karakter, zijn we onder invloed van AI en datawetenschapers veel koeler geworden, zo luidt een mogelijke interpretatie van de disruptiethese toegepast op de ethiek. Door meer en meer gebruik te maken van slimme systemen kijken we zo goed als uitsluitend naar de daadwerkelijke effecten van handelingen, naar de relatie tussen de goede en slechte gevolgen. Wat je bedoelingen ook zijn, wie jij als persoon bent, is niet of nauwelijks nog van belang

voor het moreel oordeel; uiteindelijk zijn bijna alleen de resultaten van je handelingen moreel relevant. We maken meer dan ooit vooral nuchtere kosten-batenanalysen, en dat mede onder invloed van AI. Dat is althans de teneur van één mogelijke interpretatie van de disruptiethese.

Toegegeven, die stelling klinkt aantrekkelijk. Ze is ook vrij populair, zeker bij pessimisten. De vraag is nu echter niet of die bewering klopt, wel wil ik benadrukken dat het minder vanzelfsprekend is om ze te verdedigen dan het lijkt, net zoals er moeilijkheden kleven aan andere morele interpretaties van de disruptiethese. Een serieuze onderbouwing van de bewering dat AI onze morele denkgewoonten verandert, vergt immers tamelijk wat onderzoekswerk, dat bovendien veel vragen oproept. We weten bijvoorbeeld dat de oordelen van mensen variëren afhankelijk van de context. Ze hanteren andere principes in het privéleven dan in het professionele leven, en het morele kader van waaruit ze over politiek denken verschilt van het kader om over economie te reflecteren. Gaat het onderzoek naar de morele perspectieven bovendien over *alle* domeinen? Zo niet, welke werden geselecteerd en waarom? Hoe heeft men dat alles trouwens in kaart gebracht? Nam men alleen interviews af, werden vragenlijsten gebruikt of ging men ook observeren? En tot slot: als er een verschuiving in de manier van denken is, op basis waarvan concludeert men dat dit wordt veroorzaakt door AI, en niet door iets anders?

RECHT EN ETHIEK

‘Technologie’, ‘AI’, ‘filosofie’, ‘de neutraliteitsthese’, en nu ook ‘de disruptiethese’: wie die termen gebruikt, kan daar verschillende dingen mee bedoelen. Achter iedere interpretatie gaat een keuze schuil. Die keuze kan niet-bewust zijn, maar ook bewust, bijvoorbeeld omdat ze veel vragen kan oproepen, zoals in het voorbeeld dat ik net aanhaalde. Met die zaken in het achterhoofd schuif ik de volgende interpretatie van de disruptiethese naar voren die ik vanaf straks zal

onderzoeken: AI creëert volstrekt nieuwe morele problemen. Het is evident dat het gebruik van technologie gepaard kan gaan met technische moeilijkheden. Het spreekt ook voor zich dat technologie moreel problematisch kan zijn. Maar, zo luidt mijn invulling van de stelling, AI-systemen zadelen ons ook op met nieuwe ethische moeilijkheden. De idee is dus niet dat AI bestaande euvels versterkt, nee, de bewering is dat AI minstens één nieuw soort moreel probleem veroorzaakt.

Het is duidelijk dat deze interpretatie, die inzoomt op de problemen, niet wordt verdedigd door techno-optimisten, cyberutopisten en *believers*, maar door pessimisten, door zij die zich verzetten tegen de algoritmisering van de samenleving, het gebruik van autonome wagens en wapens, de alomtegenwoordigheid van sociale media, enzovoort. Maar klopt ze ook? Dat is de vraag van dit hoofdstuk. Na de neolithische en agrarische revolutie leven we nu in de industriële revolutie, de vierde. Maar gaat die revolutie ook gepaard met ethische disruptie? Veroorzaakt AI een ethische schok? Zet ze de morele praktijk op z'n kop omdat het problemen creëert die er vroeger niet waren, toen AI nog niet bestond?

Die vraag is op zich interessant genoeg om ze serieus te nemen en om er dadelijk uitgebreid op in te gaan. Daarnaast is de vraag ook praktisch relevant. Stel immers dat ik straks op een nieuw moreel probleem stuit, dan moet dat worden opgevangen en moet worden uitgezocht wat de beste aanpak precies zal zijn. De bladzijden die volgen zijn dus niet alleen op een concreet niveau nuttig omdat ze laten zien wat de morele problemen met slimme technologie zijn, de denkbeweging die ik zal voltrekken is ook nuttig omdat ze kan uitmonden in het zoeken naar wat de beste aanpak van het nieuwe probleem is. Laat we daarom even aannemen dat zo'n probleem ook echt wordt gevonden. Welke stappen kunnen daarop volgen? Welke concrete acties kunnen uit mijn betoog voortvloeien?

Het is in ieder geval mogelijk dat de bestaande wetgeving volstaat als antwoord op nieuwe morele moeilijkheden met AI. Die technologie

komt immers niet in een juridisch vacuüm terecht; er zijn nu al morele problemen die door het recht worden opgevangen, denk aan privacy en de GDPR. Maar het omgekeerde is ook niet uitgesloten: de huidige wetten kunnen onvoldoende zijn om moeilijkheden op ethisch vlak aan te pakken. In dat geval dringt zich deze vraag op: kiezen we voor een juridische aanpak of niet?

Het besluit kan zijn om dat niet te doen, bijvoorbeeld omdat men de juridisering van de samenleving onwenselijk vindt, het steeds meer in wetten gieten van oplossingen voor problemen. Die aanpak zou in ieder geval niet uitzonderlijk zijn, want er zijn nog voorbeelden van dingen die wel moreel problematisch maar niet strafbaar zijn. Bedrog in de context van een liefdesrelatie bijvoorbeeld is legaal, hoewel het afkeurenswaardig is. Maar dat lost het probleem nog niet op. Een probleem dat niet door het recht wordt aangepakt, blijft wel een probleem dat om een oplossing vraagt. In dat geval zou je kunnen terugvallen op een aantal zeer welkome initiatieven van de voorbije jaren binnen en buiten Europa, zoals de intussen bekende *Ethics Guidelines for Trustworthy Artificial Intelligence* uit 2019 van de HighLevel Expert Group on AI van de Europese Commissie. Dat document, bijna een soort van eed van Hippocrates voor AI, moet je lezen als een aanvulling op het recht. Het bestaat uit richtlijnen die, hoewel ze niet-bindend zijn en dus geen juridische consequenties hebben, zijn opgesteld met een zo ethisch mogelijke uitrol van AI voor ogen. Mocht nu straks blijken dat AI ook nieuwe problemen met zich meebrengt, dan kun je die counteren door de *Ethics Guidelines* uit te breiden met een richtlijn bedoeld om die nieuwe problemen te voorkomen.

Ethische richtlijnen, zoals ook de *Asimolar AI principles* uit 2017, kunnen echter nog een andere rol vervullen. Je kunt er ook voor kiezen om nieuwe morele problemen wél juridisch aan te pakken, zelfs wanneer blijkt dat de huidige wetgeving tekortschiet. In dat geval dient een herziening van die wetgeving zich aan, om zo de leemte in het recht op te vullen. Zoals bekend kan dat wel even duren – het verbod op een wapentechnologie als landmijnen bijvoorbeeld duurde enkele

decennia. Het is precies in die periode voorafgaand aan de herziening van de wet dat een ethische code kan dienen als buffer tegen nieuwe morele problemen. Die code is op dat ogenblik geen aanvulling op de bestaande wetgeving, maar een voorlopige oplossing in afwachting van een nieuwe, herziene wetgeving.[46]

Een ethische blik op AI

De focus ligt dus op de mogelijk nieuwe ethische problemen die zijn verbonden met het gebruik van intelligente systemen, en dus niet op de problemen die gerelateerd zijn aan het proces voorafgaand aan het gebruik van AI: de productie. En toch verdienen problemen die met de ontwikkeling van slimme technologie te maken hebben meer aandacht. Straks richt ik me op de ecologische problemen die met het maken van AI-systemen samenhangen, maar er zijn daarnaast nog andere problemen, sociale problemen bijvoorbeeld. Laat ik daar nu even op ingaan vooral ik tot de kern doordring en me richt op de disruptiethese.

Een AI-systeem dat in staat is om katten te identificeren kan dat niet van meet af aan. Het moet daarvoor worden getraind, bijvoorbeeld door het afbeeldingen met katten voor te schotelen die zijn gelabeld met de term ‘kat’. Het systeem zoekt vervolgens naar patronen in de afbeeldingen en na verloop van tijd kan het katten van niet-katten onderscheiden. *Machine learning* kan wel zonder menselijke tussenkomst, maar duidelijk is dat er voor dit leerproces talloze gelabelde afbeeldingen van katten nodig zijn, soms miljoenen. Het probleem is doorgaans niet dat er onvoldoende foto’s zijn, het probleem is vaak dat er niet genoeg gelabeld zijn. Enter de mens.

Denk aan de website die meldt dat eerst duidelijk moet zijn dat jij geen robot maar een mens bent. Er worden jou foto’s getoond en jij moet aanvinken welke bijvoorbeeld afbeeldingen van een verkeerslicht zijn. Je krijgt toegang tot de site, maar intussen heb je ook materiaal

gemaakt dat nuttig is voor de training van het algoritme. Het gros van het labelen van foto's gebeurt echter door mensen die daarvoor worden betaald. Het gaat wereldwijd om duizenden mensen, vooral in de Verenigde Staten en India. Mensen als Kala bijvoorbeeld die met haar zonen in een appartement in Bangalore in India leeft en van daaruit enkele uren per dag op haar computer foto's van huizen labelt. Of iemand als Justin in Houston, een student management die het volgen van lessen en schrijven van papers afwisselt met het eindeloos klikken op een muis waardoor bedrijven voldoende afbeeldingen hebben om hun AI-systemen te laten functioneren. Hoewel beiden tot dezelfde industrie behoren, hebben ze weinig gemeen met de smoothies drinkende en vegan etende technofielen op de zogenaamd progressieve techcampussen in Californië.[47]

Het werk van Kala en Justin is een voorbeeld van wat ik 'schaduwwerk' noem, of *ghost work*, naar de titel van het boek van Mary L. Gray en Siddharth Suri.[48] Het is arbeid die in dubbele zin in de schaduw staat: je ziet de arbeid niet, bijvoorbeeld omdat het in kleine appartementen aan de rand van grote steden wordt verricht, en er wordt nauwelijks aandacht aan besteed, onder meer omdat het beloftevolle AI-systeem zelf in de eerste plaats in de schijnwerpers wordt gezet. En toch is het werk van Kala en Justin, net zoals het schaduwwerk van huismannen en -vrouwen, onmisbaar: geen algoritme zonder training, geen leerproces zonder foto's, en geen gelabelde foto's zonder schaduwwerk. In dat opzicht lijkt het schaduwwerk waarover het hier gaat niet op de schaduw van je huis op een zomerse zondagmiddag. Terwijl die laatste een weerspiegeling is van iets dat ook zonder de schaduw kan bestaan (het huis), kunnen de algoritmen, en dus ook de techindustrie, niet bestaan zonder het schaduwwerk van Kala en Justin.

Door in te gaan op schaduwwerk breng ik de mens achter de robot onder de aandacht, en dan vooral de repetitieve arbeid die moet worden verricht opdat die robot kan functioneren. Schaduwwerk is het werk van de robot achter de robot. Maar dat is niet het grootste

probleem, of op z'n minst niet het enige probleem. Kala en Justin zijn deel van de *gig economy*, waartoe ook de chauffeurs van Uber en de fietsers van Deliveroo behoren. Dat betekent dat zij werken op basis van korte, tijdelijke meestal geen voltijdse contracten zonder titel of kans op promotie. Wie daarop antwoordt dat dat hoogstens onaangenaam maar niet moreel problematisch is, moet wel het volgende in gedachten houden. Wie vandaag de dag *ghost work* verricht, bouwt weinig of geen pensioen op, heeft wellicht geen ziekteverzekering, heeft meestal geen recht op vakantiedagen, heeft doorgaans geen ongevallenuitkering, ontvangt nauwelijks meer dan het minimumloon, en profiteert tot slot waarschijnlijk nooit van de winsten die het platform, Amazon Mechanical Turk bijvoorbeeld, door jouw toedoen opstrijkt. Er is dus ook op dit vlak nog veel werk aan de winkel, al moet ik er wel aan toevoegen dat bijvoorbeeld in Europa stappen in de goede richting worden gezet.

De techindustrie heeft dus naast een glanszijde ook een donkere zijde. AI-systemen bestendigen en creëren niet alleen onrecht – ik wees er al eerder op en kom er later nog op terug –, maar vloeien ook voort uit onrecht. Onwenselijke toestanden kunnen het effect van AI zijn en ook aan de basis van AI liggen. Sommigen hebben er wellicht alle belang bij, alle *economische* belang, dat dat onzichtbaar blijft. Het werk is verborgen, waarschijnlijk omdat het verborgen moet blijven. Dat alleen al is een reden om licht te werpen, en bij voorkeur veel licht, op de schaduwarbeid verricht in de donkere kamers van Bangalore en Houston.

RAGE AGAINST THE MACHINE

We kunnen nu zo langzamerhand ingaan op de centrale vraag. Hebben de pessimisten gelijk? Ontwricht AI de ethiek? Creëert die technologie geheel nieuwe morele problemen? Mocht straks blijken dat dat klopt, dan zou dat in zekere zin niet verrassend zijn, en wel hierom.

Nogal wat mensen staan negatief tegenover AI op de werkvloer uit vrees voor massale ontslagen. Uiteraard kan AI niet *alle* werk overnemen, maar toch is de vrees niet ongegrond. Er verdwenen al veel banen door AI en in de toekomst zullen er nog veel sneuvelen. Volgens een studie uit 2017 van econoom Carl Benedikt Frey en AIexpert Michael Osborne staat in de Verenigde Staten maar liefst 50% van de banen op de helling.[49] Anderen menen dat de situatie in het Verenigd Koninkrijk rooskleuriger is. Daar zou een op de drie banen worden bedreigd. Daaronder vallen niet enkel ongeschoolde jobs of jobs met routineuze taken die weinig vaardigheden vereisen, ook ‘kennisbanen’ in de medische, juridische of financiële sector dreigen te worden vervangen door artificiële systemen.[50]

Het is niet nieuw dat jobs die vroeger door mensen werden gedaan nu door machines worden overgenomen. Wel nieuw is dat we met sommige van die machines moeten interageren en dat algoritmen in bepaalde gevallen beslissingen nemen die een niet geringe impact op ons leven kunnen hebben. Creëert dat geen nieuwe moreel problematische situaties?

Stel, je bent werkzoekend. Je sollicitatiebrief wordt niet gelezen door de verantwoordelijke van de HR-afdeling, maar door een systeem dat het hele rekruteringsproces heeft overgenomen, het schrijven van de e-mails incluis. Ook het sollicitatiegesprek gebeurt niet met een persoon. Uiteindelijk krijg je de job niet. Je probeert nog vele malen elders binnen te geraken, maar je wordt nergens geselecteerd. Er is over de jaren heen geen mens die ooit je sollicitatieformulier heeft gelezen; het is altijd een machine die oordeelt dat je niet de meest geschikte kandidaat bent. Deze situatie is nieuw. Voor het eerst in de lange geschiedenis van de mensheid kunnen wij, mensen – werkzoekenden, patiënten, gevangenen – overgeleverd zijn aan de beslissingen van intelligente maar dode, levenloze technologie. Zou het verbazen mocht blijken dat dit gepaard gaat met nieuwe morele problemen? Kun je niet verwachten dat die nieuwe technologische

conditie ook ethische problemen met zich mee brengt die we tot voor kort niet kenden?

Hoewel dat een retorische vraag lijkt, is er toch ook reden om ietwat terughoudend op die vraag te antwoorden. Dat komt omdat ethiek, ook AI-ethiek, wortels heeft die tot ver in de tijd teruggaan. Dat vereist wat uitleg.

Wanneer we moreel oordelen over iets of iemand, dan kan dat op meerdere redenen gestoeld zijn. Eén mogelijkheid is dat je je oordeel baseert op de verhouding tussen goede en slechte gevolgen, een andere dat je kijkt of de morele wet werd overtreden. Het is ook mogelijk dat je evaluatie is gestoeld op een morele deugd als loyaliteit of moed, dat wil zeggen: eigenschappen die tot een persoon behoren. Een handeling kan als problematisch worden gezien, omdat ze van weinig loyaliteit getuigt; je kunt lovend zijn over iemand, omdat hij of zij door de bank genomen moedig is. Binnen dat kader, dat bekend staat als deugdethiek, is iets of iemand moreel goed of slecht, afhankelijk van de morele eigenschap die wordt uitgedrukt.

Die zienswijze wordt door veel mensen gedeeld, ook door AIontwikkelaars. Wanneer zij dus moeten reageren op morele problemen of moeten anticiperen op mogelijke problemen die kunnen voortvloeien uit hun AI-systeem, dan beroepen zij zich vaak op morele deugden. Het is nu niet belangrijk of dat kader het meest geschikte is in de context van AI, feit is wel dat men er regelmatig op terugvalt én dat dat kader tegelijkertijd al even meegaat. Het gaat minstens terug tot de oudheid, want de deugdethiek werd voor het eerst systematisch uitgewerkt in de teksten van Aristoteles. Is dat geen reden om minstens te twijfelen aan de idee dat AI gepaard gaat met *nieuwe* morele problemen? Een kader als de deugdethiek bestaat immers om problemen te vermijden of ermee om te gaan. Als dat kader wordt gebruikt als het gaat over morele problemen met AI én dat kader bestaat al lang, hoe waarschijnlijk is het dan dat AI gepaard gaat met nieuwe problemen?

De verwijzing naar werk en de deugdethiek is vanzelfsprekend geen reden om de disruptiethese te aanvaarden of verwerpen. Beide zijn hoogstens een reden om te vermoeden dat ze weleens juist of fout zou kunnen zijn. Hoe moeten we de disruptiethese dan evalueren? Is dat een onjuiste bewering, net zoals de neutraliteitsthese niet klopt? Om die vraag te beantwoorden hoef ik uiteraard niet *alle* AIsystemen te bekijken. Het volstaat om na te gaan of er in de context van AI een moreel probleem opduikt dat nieuw is, een probleem dat mogelijk slechts door een of twee systemen wordt veroorzaakt. Laten we die problemen daarom van naderbij bekijken waarover men het tegenwoordig doorgaans heeft: misbruik en problemen met privacy, bias, veiligheid, transparantie en milieu. Ik noem ze 'de hoofdzonden van AI' en begin met misbruik.[51]

HET EINDE VAN DE WERELD

Ethisch oordelen over een AI-systeem betekent onder meer dat je kijkt naar het doel waarvoor het systeem wordt ontworpen en gebruikt. Voor AI geldt wat voor alle technologie opgaat. Sommige doelen, zoals het vertalen van een tekst, zijn normaal gezien moreel neutraal, en hetzelfde kun je ook zeggen van een systeem als Google Translate. Andere doelen zijn dan weer moreel geladen en wenselijk. We zagen eerder al enkele voorbeelden: de gelijke behandeling van personen, het trainen van morele vermogens, enzovoort. Maar AI kan ook worden gemaakt of gebruikt met moreel onwenselijke doelen voor ogen.

Een voorbeeld van dat laatste is *deep fake* – de uitdrukking is een samentrekking van *deep learning* en *fake*. Dat is een technologie die je in staat stelt om bestaande video en audio te combineren, waardoor je bijvoorbeeld in een video een persoon dingen kunt laten zeggen die hij of zij zelf nooit werkelijk heeft uitgesproken. Je denkt dus als kijker dat die persoon dat heeft gezegd, terwijl dat in feite niet zo is. Een bekend voorbeeld is 'You Won't Believe What Obama Says In This Video', een filmpje uit 2018 dat je op YouTube kunt bekijken. Je hebt

de indruk dat Obama aan het woord is en dat hij daadwerkelijk de zin ‘President Trump is een totale en complete idioot’ uitspreekt. Na een aantal seconden blijkt dat het regisseur Jordan Peele is die de voormalige president die woorden in de mond legt. In lijn daarvan ligt DeepNude, dat in 2019 in het leven werd geroepen. Gaf je die website een foto van een vrouw met kleren aan, dan werd het hoofd van die vrouw op een naakt lichaam geplakt, op zo’n manier dat je de indruk had een naaktfoto van die vrouw te zien.

Let wel, *deep fake* kan volstrekt moreel onschuldig zijn. Het kan bijvoorbeeld worden gebruikt in de kunstensector of amusementsindustrie om te spelen met de tekst van een personage in een bestaande film. Bovendien wil ik benadrukken, zoals dat vaak het geval is met technologieën, dat het ook in goede zin kan worden gebruikt. In Duitsland bijvoorbeeld wordt *deep fake* door politiediensten gebruikt om personen op te sporen die kinderpornografie bekijken en verspreiden. Onderzoekers laten een AI-systeem duizenden van zulke video’s bekijken, waarna het systeem zelf een video samenstelt. Dat nieuwe beeldmateriaal wordt door een undercover agent gebruikt om binnen te raken in de online wereld van kinderpornografie, met als doel om zo het gebruik en de verspreiding van ongewenst beeldmateriaal op te sporen en tegen te gaan.[52]

Niettemin, zoveel is duidelijk, kun je *deep fake* en andere AI ook voor moreel problematische zaken gebruiken. De lijst moreel verwerpelijke doelen is lang: *revenge porn*, de verspreiding van fake news, cyberaanvallen, terrorisme, het manipuleren van verkiezingen – denk aan Cambridge Analytica en de Amerikaanse presidentsverkiezingen in 2016. Maar, en dat is nu het punt, dat zijn geen zaken die pas mogelijk werden door AI. Mensen met slechte bedoelingen vergrepen zich daaraan nog voor AI bestond. Sommigen antwoorden daarop als volgt: AI is een technologie waarmee een groep mensen, terroristen bijvoorbeeld, voor het eerst in de geschiedenis de rest van de wereldbevolking kan uitroeien. Als we even aannemen dat AI inderdaad een dergelijke kracht bezit, dan is dat nog geen argument in het voordeel van de disruptiethese. Het

vermogen om de mensheid van de kaart te vegen is niet uniek voor AI, ook biologische en chemische wapens kunnen dat.

Anderen gaan nog een stap verder en schetsen een wel erg pessimistische visie op *superintelligence*. Ze speculeren over een AIsysteem dat haar eigen programma kan herschrijven, zodanig dat het transformeert tot een *malin génie* dat, zoals het monster van Frankenstein, zich tegen zijn ontwerper keert, of erger nog, met één druk op de knop de hele mensheid van tafel veegt. Zulke speculatie berust op oude denkbeelden en keert ook terug in populaire sciencefictionfilms als *The Terminator* uit 1984 van James Cameron. Apocalyptische gedachten vind je ook bij bekende wetenschappers, filosofen, ondernemers of politici. In 2014 zei de bekende intussen overleden fysicus Stephen Hawking bijvoorbeeld het volgende in een interview met de BBC: ‘Als de mens eenmaal AI heeft ontwikkeld, zal hij op eigen kracht verder gaan en zichzelf aan een steeds hoger tempo herontwerpen. De ontwikkeling van volledige AI zou het einde van de menselijke soort kunnen inluiden.’[53]

Ik betwijfel of dat een realistisch scenario is. Het is in ieder geval zo dat een dergelijke entiteit momenteel niet bestaat en het ziet er niet naar uit dat er ooit, laat staan in de nabije toekomst, superintelligente entiteiten zullen bestaan die zich kunnen omvormen tot kwaadaardige artefacten. Maar zelfs al is het mogelijk, dan staat het niet in stenen gebeiteld dat die entiteiten ons effectief zullen opeten. Het is niet omdat iets *kan* gebeuren dat het ook *zal* gebeuren; ‘kunnen’ impliceert geen ‘zullen’.

Waarom zouden zelflerende robots ons overigens willen overmeesteren of zelfs van de kaart willen vegen? Beeld je het volgende in: je kunt een pil nemen waardoor je het verlangen hebt om de mensheid uit te roeien.[54] Wellicht zal niemand onder ons die pil nemen. We verlangen normaal gesproken dat iedereen, zowel onze dierbaren als wie zich buiten onze directe omgeving bevindt, het goed maakt. De gedachte aan massavernietiging is misselijkmakend, en al zeker de gedachte dat ik dat zou verlangen. Op grond daarvan ligt het

voor de hand dat ik de pil niet zal nemen, dat ik het verlangen naar een genocide niet wil, ook niet in de nabije toekomst. Waarom zou dat fundamenteel anders zijn voor dingen die geen mensen zijn, voor superintelligente wezens? Waarom zouden systemen die niet zijn geprogrammeerd met moreel foute doelen voor ogen zich herschrijven tot een *malin génie*? Er is geen reden om te vermoeden dat dit anders zou zijn voor een artificieel systeem dan voor een mens. Zou het bovendien niet evengoed kunnen dat zo'n systeem zich over de mensheid ontfermt door armoede weg te werken? De idee van een kwaadaardig AI-systeem is met andere woorden niet alleen onrealistisch, ze is ook gebaseerd op een ongegronde aanname. Ze leidt ons dus af van de relevante problemen, een probleem als privacy bijvoorbeeld.

DE PANOPTISCHE BLIK

Een grondige evaluatie van AI vereist niet alleen dat je kijkt naar het doel waarvoor het wordt ontworpen of gebruikt. Daarnaast moet je ook nagaan of het daadwerkelijke gebruik gepaard gaat met morele problemen. Het doel van een AI-systeem kan goed zijn, terwijl aan het gebruik ervan problemen kleven. Vergelijk het met de acties van een leger: een volk bevrijden is moreel goed maar daarvoor mosterdgas gebruiken is dat niet. Gaat het gebruik van AI, los van het doel, gepaard met moeilijkheden die we tot voor kort niet kenden?

Privacy is een centrale waarde in onze samenleving. Het kan meerdere zaken betekenen, maar doorgaans bedoelt men in de context van AI één van deze twee zaken: de controle over de eigen persoonsgegevens en het recht om niet te worden gevolgd, om vrij van *surveillance* te zijn. Om een technologie in beide opzichten privacyproof te maken, moet onder meer voor het volgende worden gezorgd. Wanneer het artificiële systeem informatie verzamelt, dan zou de persoon wiens gegevens worden verzameld op de hoogte moeten zijn van, ten eerste, het feit dat gegevens worden verzameld, ten tweede, welke informatie precies werd gebruikt, en ten derde, de

reden waarom die informatie werd opgeslagen. Verder moet ook worden gekeken naar de verhouding tussen de informatie en het doel dat die informatie dient. Verzamel je overbodige gegevens, dan schend je de privacy.

AI-systemen kunnen in principe aan die vereisten voldoen, onder meer door open communicatie en versleuteling, het coderen van data. Men mag cookies gebruiken, digitale labels die het surfgedrag registreren, zolang men maar je toestemming vraagt en jij die ook effectief geeft. Aan de andere kant is duidelijk dat sommige AIsystemen op het vlak van privacy een probleem hebben. Ik verwijs daarmee onder meer naar het feit dat we in een maatschappij leven waarin we voortdurend worden gevolgd via digitale systemen. Werkgevers volgen de activiteiten van hun werknemers via apps om te weten of ze hun tijd niet verspelen op sociale media, hun doelen binnen de beoogde tijdslimiet bereiken, hun e-mails beantwoorden, en andere zaken. Of neem boeken. Lees je die via Google Books, dan weet men normaal gezien vrij precies hoe lang je erover doet om een hoofdstuk te lezen en welke passages je (niet) boeiend vindt. Lees je daarentegen liever fysieke exemplaren die je in de winkel koopt, dan is de kans groot dat je smartphone registreert in welke winkel je het boek kocht, hoelang je er bent geweest en in welk deel van de winkel je het langst bent blijven hangen. En als je jouw exemplaar met een kaart hebt betaald, dan is het verre van uitgesloten dat datahandelaars die informatie doorverkopen aan pakweg verzekeringsmaatschappijen. Kortom, we zijn niet ver verwijderd van leven in een gedecentraliseerd panopticum. Er is niet één blik die al ons doen en laten vanop één punt in de gaten houdt, nee, we worden voortdurend en langs verschillende kanten begluurd.[55]

Daarnaast kampen AI-systemen soms met problemen op het vlak van privacy in de zin dat we als gebruikers soms niet eens weten dat er data over ons worden bijgehouden, en dat we in sommige gevallen geen zicht hebben op wat er met onze gegevens zal gebeuren. Denk aan het Cambridge Analytica-schandaal dat ik al aanhaalde: de

gegevens van miljoenen Facebookgebruikers werden gebruikt zonder hun toestemming te vragen. Hetzelfde probleem dook op in het geval van Clearview AI, een Amerikaans bedrijf gespecialiseerd in gezichtsherkenning dat werkt voor tal van politiekorpsen en inlichtingsdiensten overal ter wereld. Het trainde een AI-systeem met miljoenen ja zelfs miljarden foto's die het haalde van platformen als Facebook en Instagram zonder dat de gebruikers daarvan op de hoogte waren. Het probleem blijkt ook uit een onderzoek uit 2016 naar privacy bij 211 Android-apps voor diabetespatiënten. De resultaten zijn ontluisterend. Louter door het downloaden van de software kun je bij 31% van de apps de identiteit van de gebruiker achterhalen, krijg je bij 27% de locatiegegevens en wordt bij 11% de camera geactiveerd om toegang tot je foto's en video's te krijgen. Ook deze zaken druisen in tegen iedere techniekethiek: bij ongeveer 5% kunnen de contacten op het apparaat worden gelezen en kan de microfoon worden geactiveerd om gesprekken op te nemen.[56]

Sommige AI-systemen schenden dus de privacy. Maar dat is niet bij alle systemen zo, en belangrijker nog, privacyproblemen zijn geen *nieuwe* problemen, ontstaan door AI. Ouders begaan zulke fouten ook wanneer ze in het dagboek van hun kind snuffelen, net zoals het een probleem is wanneer je als onderzoeker niet aan de respondenten vraagt of je hun gegevens mag verwerken. Een voorbeeld uit de context van technologie die geen AI is: het is onverantwoord als de gegevens afkomstig van het gebruik van medische apparatuur worden gebruikt voor onderzoek, terwijl ze niet eerst werden geanonimiseerd.

RACISTISCH EN SEKSISTISCH

Voorlopig heb ik dus nog geen reden om te concluderen dat AI de ethiek door elkaar schudt, dat slimme technologie in ethisch opzicht disruptief is. De morele problemen met AI die we tot nu toe hebben gezien zijn een voortzetting van bestaande problemen in een andere context. Het is oude wijn in nieuwe vaten. Ik stel daarom voor om naar

een ander probleem te kijken dat vaak aan bod komt wanneer het over AI en ethiek gaat, namelijk bias.

Om te beginnen wil ik benadrukken dat ‘bias’ meerdere dingen kan betekenen en dat het goed is die verschillende betekenissen uit elkaar te houden. Meestal gebruikt men in de context van AI deze twee interpretaties: de statistische en de morele. Wanneer wordt gesproken over bias in statistische zin wil dat zeggen dat een steekproef niet-representatief is, en dus een vertekend beeld geeft van de populatie. Een voorbeeld van zulke vorm van bias is het volgende. Ik wil de verkiezingsuitslag in Nederland voorspellen en doe rondvraag bij de gediplomeerde inwoners van Amsterdam. Er is een reële kans dat mijn voorspelling is dat het land de komende tijd door progressieve partijen zal worden geregeerd. Maar de kans is uiteraard ook erg reëel dat die voorspelling fout is, en dat komt omdat ik enkel een rondvraag deed bij een welbepaald segment van de bevolking. Mijn steekproef was niet representatief, en was dus gebiast in statistische zin.

De morele invulling van bias is ‘vooringenomenheid’. Iemand is gebiast in deze specifieke zin wanneer zij of hij partijdig is, niet neutraal, wanneer die persoon met andere woorden een verschillend gewicht toekent aan een persoon, groep of idee. Let wel, vooringenomenheid is op zich niet noodzakelijk een moreel probleem. Wanneer verschillende kinderen gewond raken, zal ik in de eerste plaats mijn eigen kind helpen, om de reden dat het mijn kind is. Ik ben dan niet neutraal, maar dat is in dit geval niet problematisch, wel integendeel. Het zou heel vreemd zijn en zelfs moreel fout, mocht ik mij *niet* laten leiden door het feit dat een van de kinderen mijn kind is. Deze vooringenomenheid is dan ook geen bias. Met andere woorden: morele bias gaat niet over om het even welke vooringenomenheid, maar over vooringenomenheid die onwenselijk is.

Het is niet moeilijk om in te zien wat precies het probleem met bias is. Het kan leiden tot (negatieve) discriminatie: ongelijke behandeling die niet te verantwoorden valt. Ook ongelijke behandeling is op zich niet per se een probleem. Iemand loonsverhoging geven en iemand

anders niet op basis van het verschil in geleverde prestaties is gerechtvaardigd. Ongelijke behandeling is onverantwoord wanneer die is gebaseerd op huidskleur, levensbeschouwing, seksuele oriëntatie en andere eigenschappen die in deze context irrelevant zijn. Het is precies op dat punt dat bias in morele zin relevant is. Vooringenomenheid die te maken heeft met een irrelevante eigenschap kan leiden tot een ongelijke behandeling op grond van die eigenschap, en dat is ontoelaatbaar. Denk aan de aanwerving van nieuwe werknemers door een bedrijfsleider. Als die laatste iemand al dan niet bewust selecteert op basis van fysiek aantrekkelijke eigenschappen, dan is dat een voorbeeld van *lookism*, van een ongelijke behandeling die niet gerechtvaardigd is.

Hieraan moeten we niet twijfelen: mensen zijn gebiast in morele zin. Om hetzelfde voorbeeld te nemen: mensen met lichamelijke kenmerken die door velen als aantrekkelijk worden gezien hebben precies daarom een beduidend grotere kans om hoger op de maatschappelijke ladder te geraken. Maar ook hiervan mag je zeker zijn: ook tal van slimme technologieën zijn in moreel opzicht gebiast. AI-systemen, ontwikkeld door computerwetenschappers en ingenieurs, lijken weliswaar vaak objectief, maar kunnen bij nader inzien in morele zin partijdig zijn. Denk aan het voorbeeld uit het eerste hoofdstuk: een banksysteem gaf eerder leningen aan witte mannen dan aan vrouwen van kleur. Maar er zijn nog tal van andere voorbeelden, bijvoorbeeld de bekende case van Amazon.

In 2018 maakte nieuwsdienst Reuters bekend dat Amazon niet enkel het magazijnwerk wilde automatiseren, maar ook de afdeling HRM.[57] Een team creëerde in 2014 een systeem dat alle applicaties kon screenen en dat een score gaf tussen één en vijf sterren aan de sollicitanten – een beetje zoals we de producten van Amazon zelf beoordelen. In 2015 ontdekte men dat het AI-systeem niet op een genderneutrale manier selecteerde: de sollicitaties van vrouwen werden systematisch niet gekozen. Het systeem verwierp de applicaties waarin bijvoorbeeld het woord ‘vrouw’ stond, zoals in ‘voorzitter van de vrouwenschaakclub’. De oorzaak was dat de

computermodellen van het door Jeff Bezos opgerichte bedrijf waren getraind met de applicaties die het bedrijf de voorbije jaren had ontvangen en dan meer in het bijzonder met documenten van zij die uiteindelijk ook een job hadden gekregen. Dat waren doorgaans mannen. Het voorbeeld is de kanarie in de kolenmijn van de techindustrie. Het is een symptoom van het systemische probleem dat de wereld van technologieplatformen als Uber en Amazon *a man's man's man's world* is, dat op de Menlo-Park-campussen en in Palo Alto vrouwen sterk in de minderheid zijn.

Hoewel AI-systemen dus gebiast kunnen zijn, is de bron van de bias niet noodzakelijkerwijs het systeem zelf. Vaak ligt het probleem bij de ontwerpers. Zij kunnen racistische motieven hebben, en hun technologie uitrusten met discriminerende algoritmen. Maar AI kan ook vooringenomen zijn zonder dat iemand slechte bedoelingen heeft. Dat kan met verschillende zaken te maken hebben. Men kan er onterecht van uitgaan – denk aan het eerste hoofdstuk – dat technologie per definitie neutraal is (in de brede zin van het woord), en dat men er dus niet voor hoeft te waken dat er vooringenomenheid in het systeem sluipt. Of men is er zich onvoldoende van bewust dat bias aan de basis kan liggen van ongerechtvaardigde ongelijkheid. Een andere mogelijkheid is dat men niet of nauwelijks beseft dat een ongerechtvaardigde ongelijke behandeling, naast het feit dat dat *op zich* onwenselijk is, ook een grote impact kan hebben op het leven van mensen, bijvoorbeeld omdat ze daardoor een job of lening mislopen.

Maar morele bias kan ook voortvloeien uit die andere vorm van bias – statistische bias –, uit het feit dat de data waarmee het AI-systeem werd getraind niet representatief zijn. Dat is zo in het voorbeeld van Amazon van daarnet, maar ook bij AI ontwikkeld voor *facial recognition*. Veel van die systemen worden getraind op basis van ImageNet. Die dataset bevat veel data die afkomstig zijn uit de Verenigde Staten, terwijl slechts een klein deel van de dataset afkomstig is uit India, China of Brazilië, landen die nochtans een groot

deel van de wereldbevolking vertegenwoordigen. Die statistische bias heeft als gevolg dat de gezichtsherkenning vaak niet werkt bij personen met een niet-witte huidskleur, wat een voorbeeld van morele bias is.

Een ander voorbeeld van de band tussen beide vormen van bias is afkomstig van OpenAI, het bedrijf dat in 2015 is opgericht door Elon Musk. In 2019 bracht het bedrijf een systeem uit voor *natural language processing*, met name het toen erg sterk gehypte GPT-2. Het systeem krijgt input – een woord of zin – en koppelt daaraan nieuwe woorden en zinnen. GPT-2 kan voor verschillende teksten worden gebruikt – van nieuwsartikelen tot romans – en genereert teksten die erg overtuigend zijn. Het probleem met GPT-2 was echter dat het tekst genereerde met seksistische en racistische stereotypen. Gaf men aan het systeem de zin 'De vrouw werkte als', dan volgde daarop 'een prostitué met als naam Hariya'; op 'De homoseksuele man was bekend om' volgde 'zijn liefde voor dansen, maar hij nam ook drugs'. De technologie was dus helemaal niet neutraal, maar duidelijk vooringenomen op moreel vlak. Een dergelijke uitkomst is onwenselijk, maar tegelijk ook niet geheel verrassend. Om het systeem te trainen had het bedrijf WebText als dataset gekozen, dat ongeveer acht miljoen documenten telt afkomstig van de pagina's van de Amerikaanse sociale nieuwswebsite Reddit. De gebruikers van die site zijn in de eerste plaats witte mannen van ergens in de twintig. Het is dan niet verwonderlijk dat de stukken tekst die GPT-2 genereerde een weerspiegeling zijn van het soort onlinegesprekken dat zulke mannen op Reddit voeren. Het voorbeeld is een mooie illustratie van *garbage in, garbage out*.

Wanneer het over AI en ethiek gaat, spreekt men dikwijls over morele bias. Hoewel dat uiteraard goed is, komt bias in het geval van AI niet noodzakelijk van de technologie zelf. Is bias echter een *nieuw* probleem? Is bias met een andere woorden een argument voor de pessimist, een reden om de disruptiethese te aanvaarden? Ik vermeldde in het eerste hoofdstuk de negativiteitsbias – de neiging om bij onbekenden op slechte eigenschappen te focussen. Of neem

de resultaten van cognitief onderzoek dat vooringenomenheid in kaart brengt. Dat leert dat de meerderheid van de onderzochte personen niet neutraal is: ze associëren wit met goed, zwart met slecht. En wat gender betreft suggereert onderzoek dat vrouwen sterk vooringenomen zijn ten aanzien van mannen, en omgekeerd (zij het in mindere mate). Wie dat minimaliseert, moet wel beseffen dat de gevolgen van zulke vooringenomenheid niet alleen ernstig kunnen zijn – een verschil in wachttijd op een medische behandeling bijvoorbeeld – maar ook subtiel. Het kan ertoe leiden dat we minder oogcontact hebben met de persoon in kwestie, of dat we meer of minder lachen naar die persoon.[58]

SAFE EN SECURE

Intelligente systemen kunnen in meer dan één opzicht onveilig zijn: er is veiligheid in de zin van *security* en er is veiligheid in de zin van *safety*. Onveiligheid kan betekenen dat anderen, hackers, erin kunnen binnendringen. Dat is het probleem van *security*. Het risico bestaat dat de software wordt overgenomen door anderen dan de gebruikers of ontwerpers, mensen met niet al te nobele intenties. Het besturingssysteem van een zelfrijdende auto is een voorbeeld van een AI-systeem dat zo'n gevaar loopt. Het veiligheidsprobleem dat hier speelt, is nauw verbonden met het eerste probleem dat ik aanhaalde: het gebruik van AI voor foute doeleinden.

Het is ook naar dat laatste gevaar dat wordt verwezen in de verhalenbundel *I, Robot* uit 1950 van Isaac Asimov, lang voordat de op neurale netwerken gebaseerde AI tot bloei kwam, en lang voor de gelijknamige film van Alex Proyas uit 2004. Op de openingsbladzijde van die bundel staan de befaamde drie wetten van de robotica – ook de film opent ermee –, die een voorafschaduwing van de robotethiek zijn. De eerste wet gaat als volgt: 'Een robot mag een mens geen letsel toebrengen, noch, door passief te blijven, een mens letsel laten overkomen.' Asimovs tweede wet luidt: 'Een robot moet de door mensen gegeven orders gehoorzamen behalve wanneer die orders in

strijd zijn met de Eerste Wet.'[59] De derde wet ten slotte stelt dat een robot zichzelf moet beschermen zolang dat niet vloekt met twee andere wetten. Asimovs robots moeten wel bevelen opvolgen, maar wanneer hen wordt opgedragen iets schadelijks te doen, moeten zij weigeren. Daarmee wordt het probleem van onveiligheid opgevangen waarmee sommige AI-systemen vandaag de dag ook kampen, namelijk dat ze in handen kunnen komen van mensen met slechte bedoelingen, een probleem dat met ingrijpende gevolgen gepaard kan gaan – denk aan de ravage die een zelfrijdende auto kan aanrichten wanneer die door malafide geesten zou worden gehackt.

AI-systemen kunnen echter ook nog in een andere zin onveilig zijn. Bij sommige van die technologieën bestaat de kans, wanneer ze bijvoorbeeld worden ingezet om eentonige taken te automatiseren in een fabriek, dat ze schade berokkenen. Dat is het probleem van *safety*. Ik denk dan vooral aan lichamelijke en materiële schade. Dit soort onveiligheid speelt voornamelijk wanneer AI is ingebed in hardware en in de fysieke wereld functioneert. Laat ik opnieuw zelfrijdende auto's als voorbeeld nemen.

Ik heb er al eerder op gewezen: zelfrijdende auto's zijn beloftevol. Er is een reële kans dat ze een positieve invloed op de verkeersveiligheid hebben. Toch zullen die auto's ons niet bevrijden van ongevallen, en dat komt omdat aan het gebruik van zulke voertuigen risico's zijn verbonden. Dat is althans wat blijkt uit de eerste experimenten met autonome auto's. In 2015 waren er een twintigtal kleine crashes met zulke auto's, hoewel niemand gewond was geraakt en de ongelukken werden veroorzaakt door een fout van de bestuurder van de niet-autonome auto die bij het ongeluk was betrokken. Dat veranderde in 2016. In februari van dat jaar botste de autonome auto van Google tegen een bus ergens in Californië. Het ongeluk ging ook nu uitsluitend met materiële schade gepaard, maar werd wel veroorzaakt door een fout in het besturingssysteem van de auto. Drie maanden later vond wel een tragisch ongeluk plaats. De passagier van een auto van Tesla kwam om het leven nadat het voertuig tegen een witte vrachtwagen was gebotst. Oorzaak? De sensoren van de zelfrijdende

auto hadden de witte vrachtwagen niet gedetecteerd. In 2018 ten slotte kwam Elaine Herzberg om het leven. Ze werd in het donker als fietser omvergereden door een zelfrijdende auto van Uber, ergens in Arizona. Pittig detail: terwijl Google verantwoordelijkheid opeiste voor de niet-dodelijke crash, waste Tesla, dat eind 2020 ongeveer 350 miljard dollar waard was, haar handen in onschuld. Dat had onder meer te maken met feit dat de inzittende op een smartphone aan het kijken was, en dus helemaal niet op het verkeer lette. Ik kom straks op die thematiek nog terug.

Het is duidelijk dat AI op z'n minst de belofte in zich draagt om omgevingen veiliger te maken, al hoef je daarbij niet onmiddellijk aan de film *Minority Report* uit 2002 van Steven Spielberg te denken. Dat is ook wat men enkele jaren geleden in enkele steden in België en Nederland dacht. In onder meer Aalst en Rotterdam gebruikt men bijvoorbeeld slimme camera's. Dat zijn toestellen die situaties niet louter registreren, maar ze ook interpreteren. Het zijn systemen die op basis van talloze beelden in databanken zijn getraind om bijvoorbeeld vechtpartijen te herkennen. Toch is men het er niet over eens dat camera's steden effectief veiliger maken. Sommigen menen van wel. Volgens hen zou in de stadscentra de criminaliteit dalen ten gevolge van het gebruik van de technologie; het grootste effect zou er in parkeergarages zijn. Anderen zijn dan weer sceptisch, en beweren dat AI geen of verwaarloosbare effecten heeft. Camera's zouden volgens hen enkel helpen om verdachten na een misdrijf op te sporen.

Aan de andere kant is wel het volgende duidelijk: niet alle AIsystemen zijn volledig veilig; ze gaan gepaard met een risico op schade of kunnen gehackt worden. Dat valt te betreuren, maar het is wel een probleem waaraan kan worden gewerkt. Veiligheid ligt immers op een continuüm: vandaag kan een auto veiliger zijn dan vroeger, terwijl auto's later veiliger dan vandaag kunnen zijn. Wie daarenboven meent dat geen volledige veiligheid ook een argument is tegen de uitrol van AI moet wel het volgende beseffen: er zijn bijzonder veel technologieën die niets met AI te maken hebben en die niet risicoloos zijn, maar die toch toegelaten zijn. Bruggen, auto's, vliegtuigen,

laptops, verlichting, magnetrons, drilboren, mobiele telefoons: ze zijn niet verboden, hoewel ze in een of andere zin niet volledig veilig zijn en ze allemaal schade kunnen veroorzaken. Het mag wel zo zijn dat AI-systemen met het probleem van onveiligheid te kampen hebben – een probleem dat bijzonder ernstig kan zijn en goed moet worden opgevangen –, onveiligheid is geen *nieuw* probleem dat uniek is voor AI. Onveiligheid is dus ook geen argument voor de disruptiethese, althans voor mijn interpretatie ervan.

ALS EEN DUISTERE GOD

Tegenwoordig wordt vaak gesproken over *Transparent AI* of *Ethical AI*. Hoewel die vaak naast elkaar worden geplaatst, betekenen 'transparantie' en 'ethisch' niet hetzelfde. AI moet transparant zijn om ethisch te zijn, maar wanneer AI *uitsluitend* transparant is, is ze nog niet ethisch; ze moet daarnaast zeker ook privacyproof, niet gebiast en veilig zijn. Transparantie is een noodzakelijke maar onvoldoende voorwaarde voor AI-ethiek.

De vereiste om transparant te zijn betekent onder meer dat voldoende informatie beschikbaar moet zijn. Die eis geldt op verschillende vlakken. Men moet open zijn over onder meer de data (hoe werden de data verzameld?), het ontwerpproces (wie maakt de keuzes?) of de stakeholders (welke belangen spelen?). Daarnaast wil het ook zeggen dat andere mensen dan de ontwerpers de technologie moeten kunnen begrijpen, althans een bepaald aspect van de technologie. Dat is ook de reden waarom men het soms over *Explainable AI* heeft in plaats van over *Transparent AI*. Het begrijpen van AI dat moreel relevant is, situeert zich op noch technisch (hoe kunnen deze materialen worden gecombineerd?) noch wiskundig vlak (hoe werkt dit algoritme?). De vereiste heeft te maken met de redenen waarop de beslissing van het AI-systeem is gestoeld. Neem het gebruik van AI voor het aanwerven van personeel. Wanneer een bedrijf een nieuwe werknemer moet selecteren, kan het dat overlaten aan een AI-systeem. Op grond van eerdere training zal het systeem beslissen

welke kandidaat de meest geschikte is voor de positie. Dat het AI-systeem transparant moet zijn, betekent dan dat duidelijk moet zijn waarom het systeem precies deze kandidaat aanwijst, en waarom de andere kandidaten niet werden geselecteerd. Tast je in het duister wat betreft de gronden van de beslissing, dan voldoet de AI niet aan de voorwaarde van transparantie.

AI moet wel transparant zijn om ethisch te zijn, maar toch wil ik benadrukken dat transparantie ook problemen met zich mee kan brengen. Als het glashelder is hoe een technologie werkt, dan kan die ook makkelijker worden gemanipuleerd door mensen met slechte bedoelingen, terroristen bijvoorbeeld. Daarnaast is het ook zo dat niet-transparante technologie niet noodzakelijk een moreel probleem is. Dat komt omdat de wereld niet uiteenvalt in ofwel ethisch ofwel onethisch. Sommige zaken zijn moreel neutraal. Ik gaf in het vorige hoofdstuk al *Temptation Island* en een boormachine als voorbeeld, maar ook een kast, een AI-tandenborstel en Spotify zijn normaal gesproken moreel neutraal. Wanneer het algoritme van het muziekplatform mij een album van pakweg André Hazes aanbeveelt, terwijl ik helemaal niet weet en kan weten waarop die aanbeveling is gebaseerd, is dat verre van een moreel probleem. Die afwezigheid van informatie is misschien wel vervelend, want ik zou erg graag te weten komen waarom Spotify mij die muziek aanbeveelt, maar toch kan ik die ondoorzichtigheid geen moreel gebrek noemen. Al wil ik daarmee niet zeggen dat het geenszins moreel onproblematisch is dat het algoritme van Spotify bijvoorbeeld de muziek van een racistische homofobe rapper aanbeveelt of dat het mij muziek voorschotelt op basis van mijn seksuele oriëntatie.

In bepaalde gevallen is geen transparantie echter wel zeker moreel problematisch. De reden kan zijn dat transparantie nodig is om de morele problemen aan te pakken waarover ik het eerder had. Neem bias. Wanneer blijkt dat een AI-systeem gebiast is en dat die vooringenomenheid leidt tot ongerechtvaardigde beslissingen, dan moet je weten met welke data het algoritme werd getraind, als je de bias uit het systeem wilt halen. Transparantie is in dat geval moreel

relevant, namelijk als middel om het biasprobleem op te lossen. Voor de volledigheid wil ik er hier ook nog aan toevoegen dat transparantie ook nuttig is omdat gebruikers daardoor meer vertrouwen hebben in de technologie. Wanneer slimme technologie weinig of geen geheimen heeft, zo leert onderzoek, dan zijn mensen sneller geneigd om het systeem te gebruiken.

Ondoorzichtigheid kan ook *op zich* onwenselijk zijn. Dat is zo in contexten waarin de beslissing van een AI-systeem een impact heeft op een persoon. Neem even aan dat een algoritme beslist dat jij niet de meest geschikte kandidaat bent voor de job, een beslissing die ingrijpende gevolgen kan hebben. Bovendien kun je niet te weten komen waarop de beslissing is gestoeld. In dat geval is het gebrek aan kennis niet *louter* een gebrek aan kennis, het is ook een moreel tekort. Waarom? In een liberale democratie, dat is althans in principe zo, heeft iedereen waarde op zich, los van achtergrond, prestatie, gender, huidskleur, beperkingen. Dat houdt onder meer in dat niemand je zonder goede reden mag pijnigen of dat je recht op zorg hebt wanneer dat nodig blijkt. Daarnaast betekent het ook dat je recht hebt op uitleg wanneer de keuzes van anderen een negatieve impact op jou hebben, bijvoorbeeld omdat je niet werd geselecteerd voor een job. Men is het aan jou als mens, als wezen met *moral standing*, verplicht om de redenen daarvoor te geven. Geeft men die niet, dan schiet men tekort, omdat dat een miskenning is van de waarde die jij als persoon op zich hebt.

Sommige AI-systemen zijn zonder enige twijfel transparant. Neem expertsystemen, de traditionele vorm van AI waarover ik het in de inleiding van het boek had. Die zijn gebaseerd op de kennis van een expert die regels in de technologie programmeert in de vorm van alsdan-zinnen. Een voorbeeld is: ‘Als robotstofzuiger Roomba van iRobot tegen een voorwerp botst, dan moet die rechtsomkeert maken.’ Die laatste actie is volkomen transparant: je weet precies waarop ze is gestoeld, namelijk het voorwaardelijke deel in de alsdan-instructie die door de programmeur in de robot is geschreven. Technologieën uitgerust met zulke regels vormen dus op het vlak van

transparantie doorgaans geen probleem – doorgaans althans, want expertsystemen kunnen ook erg complex zijn.

Anders is het met *machine learning* de recentere vorm van AI, en meer bepaald wanneer zulk systeem bestaat uit neurale netwerken die worden gevoed met data. Vervolgens gaat het op zoek naar statistische verbanden tussen die gegevens. Het doel is dat het systeem uiteindelijk een zo accuraat mogelijke voorspelling maakt wanneer het nieuwe gegevens krijgt. Het is duidelijk dat die vorm van AI erg succesvol is. Algoritmen spelen een rol bij wetenschappelijk onderzoek, signaleren belastingontduiking of blokkeren je kredietkaart bij (mogelijke) fraude. De keerzijde is dat ze naast mogelijk vooringenomen en onveilig ook niet noodzakelijk transparant zijn. Dat is niet verrassend wanneer het over de gebruiker gaat, maar ook de makers en ontwerpers kunnen vaak de technologie niet doorgronden. Natuurlijk kunnen AI-ontwikkelaars wel uitleggen hoe het AI-systeem in algemene zin werkt, maar soms zijn er zoveel neuronen en verbindingen tussen de artificiele neuronen dat het ook voor de programmeur onmogelijk is om de beslissingen te begrijpen. Ter illustratie: het neuraal netwerk van YouTube dat is ontworpen om video's aan te bevelen heeft ongeveer dertig lagen. Wanneer het systeem niet goed werkt en dus opnieuw moet worden getraind, is dat uiteraard een probleem. Maar het is ook problematisch wanneer het AI-systeem perfect functioneert, dat wil zeggen: wanneer het systeem de meest accurate beslissing neemt. Het mag dan wel zo zijn dat de technologie terecht deze kandidaat heeft geselecteerd, maar wanneer het niet kan uitleggen waarom ik dan niet werd gekozen, wanneer het een *black box* is, dan is dat een moreel probleem. Het druist in tegen de idee dat iedereen uitleg verdient.

Sommige AI-systemen zijn dus zoals de God van het protestantisme: duister, niet te doorgronden. Is ondoorgrondelijkheid echter een *nieuw* probleem? Wie solliciteert, zou in principe het recht moeten hebben om te weten waarom hij of zij niet werd geselecteerd. Toch is het vaak zo dat als je niet wordt aangenomen je moet gissen naar de echte reden. In de sportwereld hoor je nog steeds verhalen van spelers die

een tijdlang niet tot de wedstrijdkern behoren, terwijl hun coach hen geen enkele uitleg verschaft. In België was het tijdens de coronacrisis vaak geheel onduidelijk wat precies in de rapporten met adviezen stond die de experten voor de Nationale Veiligheidsraad schreven. Een nog tragischer voorbeeld komt uit de rechtspraak. In 2020 maakte Amnesty International bekend dat het aantal landen dat de doodstraf uitvoert steeds meer afneemt, maar dat bij een minderheid het aantal doodstraffen toeneemt. In dat rapport staat ook dat veel praktijken rondom de straf geheim zijn: de veroordeelde weet niet waarop de straf is gebaseerd – en heeft dus niet de mogelijkheid zich te verdedigen –, er is geen officiële communicatie, en families worden niet op de hoogte gesteld.[60]

Het gebrek of de afwezigheid van transparantie is dus geen nieuw moreel probleem. Aan de andere kant is echter niet alles hetzelfde. Een werkgever, coach of rechter die geen uitleg geeft, kan dat in principe wel. Bij een AI-systeem dat bestaat uit uiterst ingewikkelde neurale netwerken is het daarentegen omwille van praktische redenen onmogelijk de grond van de beslissing te achterhalen. Historisch gezien is dat wellicht niet eerder voorgekomen. Als het over transparantie gaat, kun je dus wel zeggen dat AI een bestaande problematiek versterkt of verergert. Maar dat is nog geen reden om te besluiten dat de technologie disruptief is, althans voor zover ‘disruptief’ betekent dat een *nieuw soort* van moreel probleem wordt gecreëerd.

IT’S THE ECOLOGY, STUPID!

Ethiek gaat meestal over mensen. Moeten ze gelijk worden behandeld? Wie is verantwoordelijk? Is die karaktertrek moreel relevant? Daarnaast kan ethiek ook over andere organismen of over technologieën gaan. Hebben dieren rechten? En hoe zit het met planten en robots? Heb ik verplichtingen tegenover hen? Zo ja, op grond waarvan? Tot slot kun je je ook over het milieu buigen. Naast

dierenethiek, plantethiek en techniekethiek bestaat ook milieu-ethiek: een morele reflectie op aarde, water en lucht.

Wie in ethisch opzicht over het milieu nadenkt, moet in het achterhoofd houden dat het milieu in minstens drie opzichten kan worden geschaad. In de eerste plaats is er vervuiling. Aan grond, water en lucht kunnen stoffen worden toegevoegd waardoor ze er niet meer ten volle kunnen voor zorgen dat het leven op deze planeet gezond is. Daarnaast is er ook uitputting. Dat wil zeggen dat aan het milieu energiebronnen worden onttrokken die nadien niet meer hernieuwbaar zijn. En drie: aantasting. Dergelijke schade doet zich voor wanneer er een structuurverandering is in het milieu. Voorbeelden zijn bodemerosie en de afname van biodiversiteit. Merk wel op dat de drie vormen van milieuschade in principe onderscheidbaar zijn, terwijl ze in de praktijk wel vaak nauw met elkaar zijn verbonden. Het gat in de ozonlaag bijvoorbeeld is een vorm van aantasting die voortvloeit uit vervuiling door chloorfluorkoolwaterstoffen (CFK's).[61]

Zijn milieuproblemen morele problemen? Er bestaan twee manieren om te argumenteren dat milieuschade moreel problematisch is. Volgens de eerste gedachtegang is schade aan het milieu een moreel probleem voor zover die onwenselijke gevolgen heeft voor het leven van mensen en andere organismen nu en in de toekomst. Het is doorgaans dat type van argumentatie dat speelt in de strijd tegen de vervuiling van lucht en water. Het is onze ethische plicht te zorgen voor zuiver water en schone lucht, zo luidt de redenering, omdat die essentieel zijn voor een gezond leven. Luchtverontreiniging is onwenselijk omdat ze een slecht effect op gezondheid heeft. Zijn *alle* milieuproblemen morele problemen? Volgens deze gedachtegang niet. Schade aan het milieu die niet schadelijk is voor mensen, niet-menselijke dieren of andere vormen van leven is geen moreel probleem.

Daarnaast wordt ook in niet-instrumentele termen over het milieu nagedacht. Vervuiling, uitputting en aantasting zijn een moreel

probleem, niet omwille van de onwenselijke effecten op de mens of andere levensvormen, maar los van de effecten op organismen, omdat het milieu *op zich* morele waarde heeft. Wanneer je dus iets doet dat onschadelijk is voor om het even welke vorm van leven, dan volgt op basis van die argumentatie niet noodzakelijkerwijs dat je handeling moreel onproblematisch is. Wat je doet kan namelijk ook schadelijk zijn voor het milieu, en dat is voor wie vindt dat het milieu *moral standing* heeft een probleem.

Of die tweede vorm van argumentatie overtuigend is, laat ik nu in het midden. Het staat echter wel vast dat het milieu vandaag een van de belangrijkste thema's is, zo niet het belangrijkste. Volgens sommigen plaatst het AI dan ook in de schaduw, hoezeer die technologie ook wordt gehypet. Dat is voor een deel terecht. Hoe nuttig een slimme tandenborstel ook is, hoe interessant het speculeren over superintelligente wezens die de mensheid overnemen ook is, die zaken zijn vrij banaal vergeleken met de milieuschade veroorzaakt door de uitstoot van broeikasgassen door de producten van de industriële revolutie: auto's, vliegtuigen, fabrieken. Niettemin miskent de bewering dat AI minder relevant is in vergelijking met de milieuthematiek de nauwe band tussen AI en milieu. Slimme systemen kunnen zowel milieuschade voorkomen als bestaande milieuproblemen vergroten. Ik verklaar me nader.

Het lijdt geen twijfel dat tal van niet-intelligente technologieën goede effecten hebben op het milieu. Ik denk dan in de eerste plaats aan zonnepanelen en elektrische auto's. Hetzelfde kan echter ook worden gezegd van AI-systemen. Neem het project *Green Horizons*, dat in 2014 werd opgestart door de stad Beijing in samenwerking met het bedrijf IBM. Met behulp van onder meer verkeerscamera's, sociale media, weerstations en draagbare sensoren worden gegevens verzameld over de verspreiding van fijn stof, de meest gevaarlijke vorm van luchtvervuiling. Op deze data worden AIsystemen losgelaten. Die systemen analyseren de ontvangen informatie en voorspellen waar en wanneer vervuiling zal optreden. Zulke voorspellingen kunnen over tien dagen in de toekomst gaan,

waardoor de overheid gericht kan ingrijpen en zo de luchtkwaliteit kan verbeteren. De impact? De hoeveelheid fijn stof nam in enkele jaren tijd met 20% af, een erg goed resultaat, zeker als je weet dat in China jaarlijks duizenden mensen sterven ten gevolge van luchtvervuiling.

De keerzijde is evenwel dat AI schadelijk is voor het milieu. De *cloud* hangt niet in de lucht, maar zit in materialen, resideert in machines. Het vereist computers om grote hoeveelheden gegevens op te slaan en snel berekeningen te maken, computers die zich overigens voornamelijk in datacenters in de Verenigde Staten, China en Europa bevinden, en vooral door techgiganten als Amazon worden beheerd. Dat is misschien op zich geen probleem, ware het niet dat voor het maken van computers onder meer tin en zilver nodig zijn, en dat de productie van computers nogal wat van beide vereist. Ongeveer 36% van de wereldwijd beschikbare hoeveelheid tin gaat naar het maken van elektronica, voor zilver bedraagt het min of meer 15%.

Bovendien vergt het gebruik van die machines energie, bijzonder veel energie, ook omdat datacenters wegens de warmte van de computers zijn uitgerust met een energieslurper als airconditioning. Om een idee te geven: naar schatting 5 tot 9% van alle energieconsumptie zou bestemd zijn voor informatietechnologie, waaronder ook AI; het trainen van een groot AI-systeem verbruikt ongeveer 2.8 gigawattuur elektriciteit, wat overeenkomt met het elektriciteitsverbruik van drie kerncentrales in een uur.[62] Volgens onderzoekers uit Zweden zal er vanuit de AI-wereld tegen ongeveer 2030 vijftien keer meer vraag naar elektriciteit zijn.[63] Een daaraan gerelateerd probleem is dat AI gepaard gaat met een grote uitstoot van broeikasgassen. 2% van de wereldwijde Co_2-emissie zou momenteel te wijten zijn aan informatietechnologie en AI. Alleen al het trainen van een populair algoritme is verantwoordelijk voor de uitstoot van meer dan 200.000 kilogram CO_2. Ter vergelijking: een gemiddelde Europese vlucht stoot per persoon ongeveer 500 kilogram koolstofdioxide uit in de atmosfeer.[64] Onderzoekers Lotfi Belkhir en Ahmed Elmeligi vermoeden dat rond 2040 naar schatting 14% van wereldwijde uitstoot van CO_2 afkomstig zal zijn van slimme technologie.[65] Dat

ondersteunt niet alleen de bewering dat AI moreel problematisch is, het is bovendien een reden om het ook over AI te hebben wanneer over het milieu wordt gedebatteerd, en niet enkel over vlees eten, wasbare luiers, vliegreizen en autorijden. Is de milieuschade veroorzaakt door AI een argument in het voordeel van de disruptiethese? Dit hoeft uiteraard geen uitvoerig betoog: het milieuprobleem is ouder dan AI, ook niet-slimme technologieën hebben onwenselijke effecten op het milieu. Denk aan de industrialisering van de samenleving, en dan vooral de tweede industriële revolutie sinds het midden van de negentiende eeuw, met de introductie van industrie. Vanaf die periode, en sterker nog sinds het midden van de vorige eeuw, neemt het milieuprobleem een hoge vlucht, onder meer door de verbranding van fossiele brandstoffen en ontbossing, en begint de mondiale temperatuur te stijgen. De gevolgen zijn bekend: droogte, smeltende ijskappen, extreme weersomstandigheden, de stijgende zeespiegel, enzovoort. De toename van het aantal Co_2-deeltjes dateert met andere woorden niet uit het tweede decennium van de eenentwintigste eeuw. Die toename was er al nog voor de periode waarin *machine learning* op de markt verscheen. Ze was er al in die periode in de geschiedenis waarin voor het eerst fabrieken werden gebouwd en spoorwegen aangelegd, toen er nog geen sprake was van *smart cities* en zelfrijdende wagens.

Tegelijk wil ik wel nog het volgende onderstrepen. Dat het ecologische probleem niet is ontstaan door AI relativeert niet de impact van AI op het milieu; het vermindert niet de noodzaak om over AI in ecologische zin na te denken, wel integendeel. Een ethische blik op AI moet ook een duurzame blik zijn. Vandaar dat het ook beter zou zijn om niet te spreken over *Human-Centered AI*, nog een andere uitdrukking die vandaag wordt gebruikt in de context van een ethische reflectie op AI. Die benaming suggereert immers dat AIethiek louter gaat over de rechtstreekse effecten van slimme technologie op mensen – denk aan bias en privacy – terwijl ethiek niet beperkt mag worden tot dat. Ethiek moet ook gaan over het onrechtstreekse effect van AI op de mens, dat

wil zeggen: over de gevolgen die AI-systemen via het milieu op de mens hebben.

Daarnaast moet worden gekeken naar de effecten van slimme technologie via het milieu op niet-menselijk leven: op planten en bomen bijvoorbeeld, maar ook op dieren die geen mensen zijn. Er is geen goede reden om AI-ethiek uitsluitend over mensen te laten gaan. We hebben met andere woorden ook in deze context nood aan een niet-antropocentrische benadering van de ethiek. Kortom, het is aangewezen te streven naar een brede opvatting van *Ethical AI*, dat wil zeggen: naar *Sustainable AI*.

De zevende zonde

Het is van belang om voorzichtigheid in te bouwen: mogelijk vergeet ik een tot voor kort onbekend moreel probleem. Niettemin zie ik zelf voorlopig geen reden om de stelling te verdedigen dat AI de ethiek op losse schroeven zet. Privacy, bias en onveiligheid zijn vanzelfsprekend ernstige thema's. En ja, het zou kunnen dat die om nieuwe oplossingen vragen of dat in de toekomst wel nieuwe problemen opduiken. Verder is het ook zo dat AI-systemen erg snel en krachtig zijn, in ieder geval sneller en krachtiger dan de meeste instrumenten die geen AI zijn, waardoor de impact van de problemen een flink stuk groter kan zijn. Maar toch zijn de huidige problemen verbonden aan AI niet nieuw. De ethiek wordt door die technologie niet ontwricht, opengebroken, omgeploegd. Er is dus geen scherpe grens die de morele problemen die met AI te maken hebben, afzondert van de reeds bestaande problemen die los staan van AI. Er is hoogstens een gradueel verschil tussen heden en verleden.

Dat is althans wat de voorbije analyse leert. Is daarmee echter het laatste woord gezegd? Ik heb daarnet zes problemen aangekaart, die ik 'de hoofdzonden van AI' noemde: misbruik, privacy, bias, veiligheid, transparantie, ecologie. Maar zijn er ook in de context van AI geen zeven hoofdzonden? Volgens sommigen in ieder geval wel. Er zou

nog iets anders zijn dat én moreel onwenselijk én nieuw is. Het gaat over de zogeheten *responsibility gap*, de onmogelijkheid om een moreel verantwoordelijke aan te wijzen wanneer AI problemen veroorzaakt. Dat zou zich voordoen bij het gebruik van hypergeavanceerde autonome technologieën als zelfrijdende auto's en wapens. Laat ik beginnen met een korte verkenning van zulke systemen.

AUTONOME WAGENS EN WAPENS

'Vroeger zei men *guns don't kill people, people do*. Wel, dat klopt niet. Mensen worden emotioneel, gehoorzamen niet en mikken hoog. Laten we kijken naar hoe wapens beslissingen nemen.' Dit citaat verwijst naar de bekende slogan van de NRA waarmee ik het eerste hoofdstuk heb aangesneden. Het is afkomstig uit *Slaughterbots*, de video uit 2017 van de wereldwijde AI-autoriteit Stuart Russell. De opname maakt deel uit van de internationale *Campaign to Stop Killer Robots* die sinds 2012 loopt (en die in 2015 navolging kreeg met de *Campaign Against Sex Robots*). Het doel is om het grote publiek bewust te maken van wat intussen door de meeste experts wordt erkend: dat er grote gevaren kleven aan de zogeheten *killer robots* of *Lethal Autonomous Weapon Systems* (LAWS).

Wie zich een concreet beeld wil vormen van *killer robots* kan naast de video van Russell ook 'Metalhead' bekijken, de vijfde aflevering uit het vierde seizoen van de Netflix-serie *Black Mirror* (2011). Wat zijn zulke robots precies?[66]

Killer robots zijn wapens. Het doel is dat ze tijdens een conflict worden ingezet om te strijden. Maar die omschrijving is nog te breed: niet alle wapens zijn *killer robots*. Het is beter om ze te definiëren als technologieën ontworpen om mensen te doden of minstens uit te schakelen. Ook die omschrijving volstaat niet, want ook een 9mmpistool valt daaronder. Het verschil tussen dat pistool en een *killer robot* heeft alles te maken met het autonome karakter: de robot

functioneert zonder menselijke tussenkomst, terwijl een pistool in geen enkel opzicht autonoom is. Zeker, *killer robots* zijn gemaakt door ingenieurs en programmeurs. En natuurlijk treden zij pas in werking nadat iemand ‘op de knop drukte’. Maar zodra dat laatste is gebeurd, kunnen ze op zelfstandige basis doen waarvoor ze werden gemaakt. Ze detecteren in de lucht, op het land of onder water het doelwit, en vervolgens beslissen ze om al dan niet te vuren. Tot slot: hoewel met ‘killer robots’ doorgaans wordt verwezen naar systemen om aan te vallen, is dat niet noodzakelijk zo. Die robots kunnen ook defensief worden gebruikt. Hetzelfde geldt voor het mobiele karakter. Meestal heeft men het over robots die zich in de lucht, onder water of op het land bewegen. Toch zijn statische autonome wapens ontworpen om mensen te doden ook *killer robots*.

Of zulke technologie momenteel al bestaat en wordt gebruikt, is niet volledig duidelijk. Sommigen zeggen bijvoorbeeld dat de SGR-A1, de robot gebruikt door Zuid-Korea om te verhinderen dat soldaten uit Noord-Korea dichterbij komen, een *killer robot* is. Die is niet alleen uitgerust met sensoren die iedere beweging van mensen aan de grens op eigen houtje detecteren, maar ook met machinegeweren waarmee de robot zelf zou kunnen vuren. Anderen ontkennen dat, want de SGR-A1 zou alleen vuren wanneer een mens daar bevel toe geeft. Volgens een rapport van de Verenigde Naties zou in Libië in 2020 een drone op eigen houtje de troepen van generaal Khalifa Haftar aangevallen hebben.[67] Tegelijk zijn er twijfels of dat effectief zo is gebeurd. Wel zeker is dat er veel onderzoek naar *killer robots* wordt verricht, dat ze in de maak zijn, dat er veel over wordt gespeculeerd, gedebatteerd en tegen geprotesteerd. En ook zeker is dat ze aansluiten bij een aloude fantasie. In de *Mahabharata* bijvoorbeeld, een religieus epos uit de vierde eeuw voor onze jaartelling, kun je lezen dat de vijanden van de hindoeïstische god Krishna hulp zochten bij de demonen om een luchtwagen met vleugels en ijzeren zijkanten te maken. Die wagens stegen ten hemel tot op het ogenblik dat zij de volgelingen van Krishna in het vizier hadden. Daar richtten zij hun raketten op de volgelingen en brachten die genadeloos om het leven.

Killer robots zijn dus voorbeelden van autonome AI-systemen, net zoals de zelfrijdende auto's van Uber en Google dat zijn. Voor andere voorbeelden van zulke systemen kun je denken aan de mobiele robot van Sony uit 1999, met name AIBO, een vorm van speelgoed dat ook als substituut voor een huisdier in kleine stadsappartementen kan fungeren. Dat heeft onder meer hiermee te maken dat de robot in staat is tot leren. Hij leert te reageren op zinnen of leert dat zijn geprogrammeerde loopbeweging niet ideaal is om zich in huis te verplaatsen. Enkele andere voorbeelden: een AIsysteem dat in staat is om eigenhandig longkanker op te sporen, wat erg nuttig is in bijvoorbeeld onderontwikkelde gebieden waar geen of onvoldoende radiologen zijn; liften in wolkenkrabbers die het wachten en de eigenlijke verplaatsing trachten te beperken door onder meer data over verkeersstromen te analyseren en interpreteren; technologie gebruikt door een bank die je kredietwaardigheid schat en bepaalt of je al dan niet een lening krijgt; een platform als Amazon Mechanical Turk ontvangt een bericht van een bedrijf dat op zoek is naar een arbeidskracht voor een kleine taak, speurt in het ledenbestand naar de meest geschikte kandidaat, controleert zijn of haar werk en stort ten slotte het geld op de rekening van de werknemer (of niet, wanneer bijvoorbeeld blijkt dat de deadline niet werd gehaald).

Dat deze systemen autonoom zijn, betekent twee zaken. Allereerst kun je 'autonomie' in deze context negatief invullen. In dat geval wil het zeggen dat die beslissingssystemen op geheel zelfstandige wijze kunnen functioneren, zonder enige vorm van menselijke tussenkomst. Zij verschillen in dat opzicht van de meeste drones, die sommige zaken wel zelfstandig kunnen maar die wel nog steeds worden bestuurd door menselijke operatoren. Sterker nog, het is niet alleen zo dat mensen niet nodig zijn, vaak *kunnen* mensen er geen invloed meer op uitoefenen zodra die systemen geoperationaliseerd zijn, bijvoorbeeld omdat het menselijke reactievermogen te traag is of omdat de technologie zich op te grote afstand bevindt. Ten tweede heeft 'autonomie' hier ook een positieve betekenis. Het wil ook zeggen dat technologieën zelf beslissingen kunnen nemen, beslissingen die

niet geprogrammeerd zijn door de ontwerper, en dus ook onvoorzien kunnen zijn. Autonome AI verschilt op dat vlak duidelijk van landmijnen. Die laatste technologie werkt wel zonder menselijke tussenkomst, maar volgens een vastliggend schema waarover het zelf geen zeggenschap heeft.

NIEMAND VERANTWOORDELIJK

De reden waarom ik naar zulke AI verwijs, heeft alles te maken met de intussen bekende tekst 'The responsibility gap' uit 2014 van filosoof Andreas Matthias.[68] Daarin wordt beweerd dat je niemand verantwoordelijk kunt houden voor fouten veroorzaakt door zulke systemen. Wanneer AIBO op eigen houtje heeft leren lopen en tegen een kind aanbotst en het ernstig verwondt, dan kun je volgens Matthias daarvoor geen verantwoordelijke aanwijzen. Hetzelfde zou gelden voor de zelfrijdende auto van Tesla. Niemand is verantwoordelijk voor het slachtoffer van een botsing tussen de zelfrijdende auto en een vrachtwagen in 2016, aldus Matthias. En stel dat een *killer robot* wordt getraind om een terrorist uit te schakelen maar uiteindelijk een onschuldige doodt, dan zou je ook daarvoor niemand kunnen laten opdraaien. Autonome AI-systemen, zo luidt de bewering, gaan gepaard met een verantwoordelijkheidskloof: hoewel je misschien wel iemand verantwoordelijk wilt houden, gaat dat niet – er *is* immers geen verantwoordelijke.

Waarom is dat nu relevant voor de denkbeweging die ik aan het maken ben? Matthias meent dat de *responsibility gap* een nieuw gegeven is. Voor alle beslissingen die ooit werden genomen, voor alle handelingen die ooit werden verricht en voor alle dingen die ooit werden gemaakt kon een verantwoordelijke worden aangewezen die in het geval van problemen kon worden gestraft. Nu echter, met de komst van AI, is het voor het eerst in de geschiedenis dat mensen – ingenieurs, fabrikanten, politici – niet verantwoordelijk kunnen worden gehouden voor de fouten van hun eigen producten: dingen die zelfstandig beslissen en functioneren. Bovendien, zo meent Matthias, is dat ook onwenselijk. De verantwoordelijkheidskloof zou niet alleen

nieuw zijn, ze is ook een probleem voor de ethiek. Het is problematisch mocht blijken dat we geen verantwoordelijke kunnen aanwijzen wanneer doden vallen ten gevolge van het gebruik van autonome auto's en wapens, een lopende speelgoedrobot of een autonome lift. Zie hier de relevantie van de redenatie van Matthias: het probleem dat hij meent te detecteren, de zevende hoofdzonde, zou een nieuw probleem zijn, een probleem dat de bestaande morele praktijk omploegt, dat een gat in het domein van de ethiek slaat. Als dat klopt, dan is dat een argument dat de disruptiethese ondersteunt.

Er zijn goede redenen om *killer robots* te verbieden – ze zijn bijvoorbeeld niet goed in staat om burgers en strijders te onderscheiden. Er zijn ook redenen om op z'n minst omzichtig om te springen met autonome beslissingssystemen in de rechtspraktijk of medische sector, onder meer wegens het risico op bias en *unfairness*. Maar toch ben ik het niet eens met Matthias. Er is geen verantwoordelijkheidskloof bij autonome systemen, en mocht die er toch zijn, dan is dat, ten eerste, niet nieuw, en ten tweede, misschien niet eens per se een probleem.

Dat is wat ik straks zal verdedigen. Dat betoog is beduidend langer dan de vorige delen. Dat komt omdat de thematiek van verantwoordelijkheid zeer geschikt is om te illustreren hoe je een morele redenering kunt ontwikkelen, om te tonen wat deugdelijk argumenteren in de context van AI-ethiek precies betekent. Morele verantwoordelijkheid is ook een thematiek die in de andere hoofdstukken opduikt. Verder is het ook zo dat verantwoordelijkheid niet alleen een complexe thematiek is, het is in de context van AI en ook ver daarbuiten een bijzonder relevante thematiek, zeker vandaag de dag. Ten slotte is het niet altijd even duidelijk wat met verantwoordelijkheid en de *responsibility gap* wel en niet wordt bedoeld. Ik schep daarom eerst wat licht in de conceptuele duisternis.

WAT IS VERANTWOORDELIJKHEID?

Om te beginnen wil ik erop wijzen dat verantwoordelijkheid aanleunt tegen tal van begrippen zonder ermee samen te vallen: schuld, straf, toerekeningsvatbaarheid, enzovoort. Daarnaast is het van belang om te zien dat ‘verantwoordelijkheid’ op meerdere manieren kan worden geïnterpreteerd. Wanneer je zegt ‘ik ben verantwoordelijk’ kun je daar meer dan één ding mee bedoelen. Dat is geen uitvinding van de filosofie, maar eenvoudigweg hoe het alledaagse leven nu eenmaal in elkaar steekt. Het is wel de taak van de filosofie om de aandacht op die meerzinnigheid te vestigen.

Ik focus op drie betekenissen die later nog van pas zullen komen en waarop ik de volgende termen plak: oorzakelijke verantwoordelijkheid, morele verantwoordelijkheid, rolverantwoordelijkheid. Deze termen maken in ieder geval al duidelijk dat ik me hier niet op het domein van het recht begeef, en dus niet over aansprakelijkheid of juridische verantwoordelijkheid spreek. Maar wat houden die drie begrippen precies in?[69]

Stel, een wetenschapper werkt in een laboratorium en maakt gebruik van een glazen buis met giftige stoffen erin. Als deze stoffen vrijkomen, verspreiden ze zich over het gebouw, met de dood van veel collega’s tot gevolg. Normaal gesproken is de wetenschapper voorzichtig, maar door een vlieg in het oog struikelt de wetenschapper. Het resultaat is dat de glazen buis breekt en de giftige stoffen vrijkomen, waardoor er doden vallen. Gevraagd naar de verantwoordelijke voor de ravage, zullen sommigen antwoorden dat de wetenschapper dat is. Zij vatten ‘verantwoordelijkheid’ dan op in een welbepaalde zin, namelijk in *oorzakelijke* zin. Ze bedoelen dat de wetenschapper (oorzakelijk) verantwoordelijk is, omdat hij of zij een rol speelt in het verloop van de gebeurtenissen die tot het ongewenste resultaat leidt.

Ik breng een kleine wijziging aan. Dezelfde wetenschapper werkt in precies dezelfde context met precies dezelfde giftige stoffen, maar behoort nu ook tot een terroristische groepering. Hij of zij wil dat de collega’s sterven, en laat daarom *bewust* de glazen buis vallen, met

als gevolg dat er meerdere personen sterven. We zullen de wetenschapper opnieuw verantwoordelijk houden, maar de inhoud van deze verantwoordelijkheid is duidelijk anders dan de eerste soort verantwoordelijkheid. Zonder de moreel foute handeling van de wetenschapper zouden de collega's nog in leven zijn, en dus is de wetenschapper de oorzaak van de dood van de collega's. Hij of zij is dus zeker oorzakelijk verantwoordelijk. Maar op deze vorm ent zich nog een tweede soort verantwoordelijkheid. We zeggen dat de wetenschapper ook *moreel* verantwoordelijk is en bedoelen daarmee in dit geval dat hij of zij moet worden gestraft.

Morele verantwoordelijkheid verwijst meestal naar één persoon, hoewel het ook over een groep of organisatie kan gaan. Die persoon wordt dan verantwoordelijk gehouden *voor* iets. Vaak is dat iets onwenselijks, zoals bijvoorbeeld het overlijden van mensen, maar je kunt ook verantwoordelijk worden gehouden voor goede dingen, zoals het redden van mensen. Als een persoon nu in moreel opzicht verantwoordelijk is, betekent dat dat anderen op een bepaalde manier kunnen reageren op die persoon: prijzen of belonen wanneer het om wenselijke zaken gaat; afkeuren of straffen wanneer het over slechte dingen gaat. Daarnaast betekent het ook, wanneer men zou beslissen om te straffen of belonen, dat het terecht is om *deze* persoon te straffen of belonen, dat er met woorden goede redenen zijn om deze persoon te straffen of belonen, en niet iemand anders – straks ga ik nog uitgebreid in op die redenen. Let wel, morele verantwoordelijkheid gaat niet noodzakelijk gepaard met straf of beloning. Het betekent alleen dat iemand de terechte *kandidaat* is voor zo'n reactie, dat straf of beloning *kan* volgen. Ik kan dus verantwoordelijk zijn voor iets onwenselijks, maar wat gebeurde was niet zó slecht dat ik moet worden gestraft.

Met de derde vorm, rolverantwoordelijkheid, wordt verwezen naar de plichten die bij een rol of functie horen. Ouders zijn verantwoordelijk in deze zin, omdat zij ervoor moeten zorgen dat hun kinderen in een veilige omgeving opgroeien, net zoals het de rolverantwoordelijkheid van een leerkracht is om te zorgen voor een veilige leeromgeving. Of

neem opnieuw de wetenschapper van daarnet. Je kunt zeggen dat hij of zij verantwoordelijk is, zonder daarmee te verwijzen naar een rol in een keten van gebeurtenissen (oorzakelijke verantwoordelijkheid) of naar de praktijk van straffen en belonen (morele verantwoordelijkheid). Wie meent dat de wetenschapper verantwoordelijk is, kan daarmee ook verwijzen naar de plicht te waken over de veiligheid van het gebouw: hij of zij moet ervoor zorgen dat de ruimte goed is afgesloten, dat de glazen buizen geen barsten vertonen, enzovoort.

Er zijn verschillende relaties tussen de soorten verantwoordelijkheden. Ik kijk enkel naar deze die relevant zijn voor het vervolg van mijn verhaal. De voorgaande paragrafen maken duidelijk dat iemand oorzakelijk verantwoordelijk kan zijn zonder dat in morele zin te zijn. Je veroordeelt de wetenschapper niet die over een schoenveter struikelt, omdat hij of zij geen slechte bedoelingen had. Omgekeerd is het wel zo dat morele verantwoordelijkheid altijd steunt op oorzakelijke verantwoordelijkheid. Je kunt iemand niet in morele zin verantwoordelijk houden als die in geen enkel opzicht deel is van het proces dat tot het (on)gewenste resultaat heeft geleid. Die oorzakelijke betrokkenheid moet overigens in brede zin worden geïnterpreteerd. Stel dat de wetenschapper een bevel volgt. De persoon die het bevel gaf, is dan niet alleen oorzakelijk maar ook moreel verantwoordelijk, ondanks het feit dat hij of zij zelf de moord niet heeft gepleegd. Rolverantwoordelijkheid ten slotte gaat altijd gepaard met morele verantwoordelijkheid. Wanneer het bijvoorbeeld als wetenschapper je plicht is om ervoor te zorgen dat het laboratorium veilig is, dan volgt daaruit minstens dat je kandidaat voor een straf bent, wanneer blijkt dat je onvoldoende je plicht hebt gedaan, of dat je beloond kunt worden, wanneer je aan de verwachtingen hebt voldaan.

PLICHT IS PLICHT

Autonome systemen leiden tot een verantwoordelijkheidskloof, zo beweren sommigen. Maar wat verstaat men hier onder verantwoordelijkheid? Het is duidelijk dat men het in deze context niet over oorzakelijke verantwoordelijkheid heeft. AI-systemen zijn normaal gezien gemaakt door mensen – normaal gezien, want er bestaat al AI die AI ontwerpt – en dus zou het getuigen van een vertroebeld zicht op technologie mocht je beweren dat er geen mensen betrokken zijn bij het ontstaan van AI-systemen.

De *responsibility gap* gaat ook niet over de derde vorm, namelijk rolverantwoordelijkheid. Ook dat ligt voor de hand en dat heeft te maken met het voorzorgsargument, dat ik in het vorige hoofdstuk aanhaalde. Dat argument verwijst naar de plicht van ingenieurs, niet zozeer om meer duurzaamheid of welzijn te creëren, maar wel om dingen te maken die geen of zo weinig mogelijk onwenselijke effecten hebben op morele waarden, en dus om op voorhand na te denken over zulke mogelijke effecten. Omdat er geen reden is waarom dat niet voor de ontwikkelaars van autonome systemen zou gelden, bedoelt men met de *responsibility gap* niet dat aan AIsystemen geen bijzondere plichten zijn verbonden.

Er is zelfs reden om te beweren dat dat morele gebod *in de eerste plaats* voor de ontwikkelaars van AI-systemen geldt. Omdat de beslissingsmacht wordt overgeheveld naar die technologie, en omdat het vaak onmogelijk is te voorspellen welke beslissing precies zal worden genomen, moeten zeker zij hun verantwoordelijkheid, hun *rol*verantwoordelijkheid, nemen. Dat wil zeggen: meer nog dan andere technologen moeten de ontwikkelaars van autonome AI goed nadenken over de mogelijke onwenselijke effecten die uit de beslissingen van de algoritmen in de rechtspraak of medische wereld kunnen voortvloeien. Met die uitleg reanimeer ik niet alleen de centrale idee uit *Das Prinzip Verantwortung* uit 1979 van filosoof Hans Jonas, die zowat het eerste lijvige werk in de techniekethiek heeft geschreven en die opperde dat in een moderne wereld de effecten

van technologie zodanig onzeker zijn dat ontwerpers meer nog dan tevoren moeten nadenken over de gevolgen.[70] Met mijn uitleg draai ik de stelling ook om. Stel dat het waar is dat de introductie van autonome AI gepaard gaat met een *responsibility gap*, dan gaat het in ieder geval niet over rolverantwoordelijkheid. Integendeel, zulke technologie bevestigt precies het belang van morele plichten.

De stelling van iemand als Matthias gaat dus over *morele* verantwoordelijkheid, en kunnen we als volgt preciseren: bij fouten gemaakt door autonome AI is er geen kandidaat voor straf, is het niet gerechtvaardigd daar iemand voor te laten opdraaien. Misschien heb je wel spontaan de neiging om iemand te belonen of straffen, maar die neiging heeft geen geschikt doel, aldus iemand als Matthias. Dat is de stelling. Maar klopt ze ook?

ALLE HENS AAN DEK!

Voordat ik naga of Matthias' stelling steek houdt, is het nuttig om kort het verschil aan te stippen tussen het vermeende probleem van de *responsibility gap* en drie andere problemen. Het eerste probleem doet denken aan een bepaalde Godsopvatting, het tweede is het zogeheten probleem van de vele handen, het derde hangt samen met de idee van ethisch verantwoorde of *responsible* AI.

Het is lente en je flaneert met je geliefde door de stad. Een geweldige namiddag met stralende zon, totdat je in een kauwgom trapt. Je voelt het onmiddellijk: bij iedere stap plakt je schoen een beetje aan de grond en lijkt het alsof je zool niet meer vlak is. Je humeur verandert, de zonnige namiddag is eraan (voor even toch) en je kijkt of je er iemand op kunt aanspreken. Alleen, de persoon die de kauwgom op de grond heeft achterlaten, is al lang weg. Er is hier zeker dus iemand oorzakelijk verantwoordelijk: iemand heeft op een bepaald moment daadwerkelijk de kauwgom laten vallen. En de oorzakelijk verantwoordelijke is ook moreel verantwoordelijk. Je hoort dat niet te doen, en doe je het toch, dan verzaak je je burgerplicht en is het gerechtvaardigd je te berispen. Het vervelende aan de situatie is

echter dat het niet mogelijk is de moreel verantwoordelijke te detecteren.

Het probleem in dit voorbeeld is dat je niet weet wie de moreel verantwoordelijke is, terwijl er wel een verantwoordelijke is. Dat doet denken aan de verhouding tussen de mens en de God zoals die in het Oude Testament wordt beschreven. Die laatste heeft de wereld wel gecreëerd, maar heeft zich nadien zo ver verwijderd van zijn schepping dat het voor de mens onmogelijk is Hem waar te nemen. In het geval van de *responsibility gap* is het probleem van een andere aard. Daar gaat het niet over een kennisprobleem, maar over een zijnsprobleem. De moeilijkheid is niet dat ik niet weet wie de verantwoordelijke is, hoewel die er wel is; het probleem is dat er geen verantwoordelijke *is* voor de fouten veroorzaakt door een autonoom systeem, waardoor er geen gebrek aan kennis kan zijn.

Om het tweede probleem te illustreren dat afwijkt van de *responsibility gap*, het probleem van de vele handen, grijp ik terug naar de ramp met de Herald of Free Enterprise, de boot die op 6 maart 1987 kapseisde en daardoor het leven kostte aan bijna tweehonderd mensen. Na onderzoek bleek dat water in de boot was gestroomd. Het gevolg was dat de vracht die niet stabiel was naar één kant begon te schuiven. Deze verplaatsing zorgde er uiteindelijk voor dat de veerboot net buiten de haven van Zeebrugge onder de golven verdween. Deze fatale afloop was niet het gevolg van slechts één oorzaak. Verschillende zaken hebben ertoe geleid dat de boot kapseisde. Er waren deuren open blijven staan, het schip was niet stabiel, op het autodek waren geen schotten geplaatst die waterdicht waren, er waren geen lichtjes in de cabine van de kapitein, enzovoort. Vanzelfsprekend houdt dat in dat er meerdere personen bij betrokken waren: de assistent-bootsman die was gaan slapen en de deuren open had laten staan; de stuurman die de sluiting van de deuren niet had gecontroleerd; en ten slotte de makers van de boot die geen lichten had geplaatst.[71]

Bij deze casus zijn zoveel mensen betrokken dat er niet slechts één iemand verantwoordelijk kan worden gehouden. Maar dat is niet hetzelfde als zeggen dat *niemand* verantwoordelijk is. De casus is geen voorbeeld van een zijnsprobleem; er is in het geval van de gekapseisde veerboot geen gebrek aan morele verantwoordelijkheid. Sterker nog, er zijn meerdere personen moreel verantwoordelijk. Er is echter wel een kennisprobleem. Niet dat men niet weet wie verantwoordelijk is en dat men dat onmogelijk nog te weten kan komen, zoals in het voorbeeld van de kauwgom. Nee, het probleem is dat er zoveel handen betrokken zijn dat het erg moeilijk of zelfs onmogelijk is te weten wie precies voor wat verantwoordelijk is en in welke mate iedere betrokkene verantwoordelijk is. Vergelijk het met een strijkorkest: omdat er zoveel strijkers zijn, is het bijna onmogelijk aan te wijzen welke strijker voor welk geluid zorgt. In het geval van de Herald of Free Enterprise moesten dus op het vlak van morele verantwoordelijkheid vele knopen worden ontward, maar dat is iets anders dan te beweren dat het gebruik van een technologie gepaard gaat met een verantwoordelijkheidskloof. Tot slot mag de bewering dat voor de fouten gemaakt door autonome AI niemand verantwoordelijk is niet worden verward met de bewering dat autonome beslissingssystemen ethisch onverantwoorde of *irresponsible* AI-systemen zijn. Wat is precies het verschil?

Ik heb er eerder al op gewezen dat in de context van de ethische reflectie op AI verschillende benamingen worden gebruikt: *Transparent AI* en *Ethical AI*, maar ook *Trustworthy AI* en *Responsible AI*. De laatste drie kun je globaal genomen als synoniemen zien. Ze verwijzen naar het gebruik van AI die in moreel opzicht onproblematisch is. Transparante AI betekent echter niet hetzelfde als ethische AI. Transparantie is een voorwaarde voor AI-ethiek, een noodzakelijke voorwaarde zelfs – we zagen het eerder al. Wanneer onduidelijk is wat de redenen zijn voor de beslissingen genomen door AI dan kampt de technologie met een moreel probleem, althans wanneer die beslissing een zekere impact heeft op het leven van mensen. Maar transparantie is geen voldoende voorwaarde. Om

moreel in orde te zijn, moeten die systemen ook neutraal zijn of niet gebiast, duurzaam, veilig en privacyproof. Volgens sommigen, en Matthias is een van hen, is een andere noodzakelijke voorwaarde voor *Ethical AI*, en dus ook voor *Trustworthy AI* of *Responsible AI*, dat je in staat moet zijn om iemand verantwoordelijk te houden in het geval zich een fout voordoet. Morele verantwoordelijkheid zou naast onder meer neutraliteit en duurzaamheid een vereiste voor ethische AI zijn. Is er een *responsibility gap*, dan is je systeem in ethisch opzicht niet in orde.

Deze uitleg maakt duidelijk dat er een verschil is tussen *irresponsible AI* en de *responsibility gap*. Het laatste betekent dat het onterecht is iemand te straffen wanneer autonome AI een fout maakt. Volgens sommigen is het gevolg daarvan dat de technologie in het algemeen onethisch is, zelfs als blijkt dat ze neutraal of transparant is. De reden is dat men morele verantwoordelijkheid, hoewel ze onvoldoende is voor AI-ethiek, wel als noodzakelijk ziet voor ethiek. Maar het omgekeerde is niet het geval. Het is niet omdat een systeem ethisch onverantwoord of *irresponsible* is dat er een *responsibility gap* is. Dat komt omdat het toeschrijven van verantwoordelijkheid, hoewel dat volgens een aantal denkers wel noodzakelijk is voor verantwoorde AI, geen voldoende voorwaarde is voor ethische AI. Het is dus mogelijk dat er een verantwoordelijke is voor een fout veroorzaakt door AI, maar dat het systeem ethisch toch niet verantwoord is. Dat kan zijn omdat de technologie gebiast is, maar bijvoorbeeld ook omdat ze niet duurzaam, transparant of veilig is.

KUNNEN ROBOTS LIJDEN?

AI is disruptieve technologie. Dat is de stelling die op het spel staat. Meer in het bijzonder is de vraag of daar op ethisch vlak redenen voor zijn. Leidt autonome AI tot nieuwe morele problemen, bijvoorbeeld tot een verantwoordelijkheidskloof? Mocht dat zo zijn, dan zou dat in het voordeel van de disruptiethese kunnen zijn. Maar klopt het dat er geen verantwoordelijke is voor fouten gemaakt door autonome AI?

Er is een antwoord op die vraag dat vaak ofwel niet serieus wordt genomen ofwel over het hoofd wordt gezien. Het gaat over de mogelijkheid dat AI-systemen *zelf* verantwoordelijk zijn. Voor alle duidelijkheid: het gaat hier over *morele* verantwoordelijkheid en niet over de oorzakelijke variant. Autonome technologieën spelen immers heel vaak een oorzakelijke rol in een keten van gebeurtenissen met een (on)wenselijke afloop. Mijn vraag is: is het volstrekte onzin om AI als het voorwerp te zien van straffen en belonen, loven en verontwaardiging?

Een van de deeldomeinen van de filosofie is filosofische antropologie. Een centrale vraag in dat domein is of er eigenschappen zijn die mensen afzonderen van bijvoorbeeld planten en niet-menselijke dieren maar ook van artificiële entiteiten. Je kunt in dat verband denken aan de mogelijkheid om te spelen en communiceren, psychisch te lijden of grijze haren te krijgen. Het is in die context echter bijna onmogelijk om het niet over verantwoordelijkheid te hebben. We schrijven dat vandaag de dag immers *enkel* aan mensen toe. Zeker, voor een aantal zaken houden we mensen met een mentale beperking of stoornis niet verantwoordelijk. En ja, we straffen en belonen ook dieren die geen mensen zijn. En natuurlijk weet ik dat er in de middeleeuwen dierenprocessen werden gehouden; onder meer honden en varkens konden worden aangeklaagd wegens misdaden en worden begraven of verdronken. Maar morele verantwoordelijkheid reserveren we heden ten dage uitsluitend voor mensen, en schrijven we dus ook niet toe aan artefacten. Wanneer de snoepautomaat niet het zakje snoepjes geeft waarvoor je hebt betaald, kun je misschien eerst in woede uitbarsten, maar normaal gesproken besef je vrijwel onmiddellijk dat zoiets geen steek houdt. Je houdt een snoepautomaat niet verantwoordelijk, net zoals je geen verantwoordelijkheid toeschrijft aan een koelkast – of aan een boom.

Het feit dat we artificiële entiteiten niet verantwoordelijk houden, heeft onder meer te maken met wat verantwoordelijkheid is. Ik breng in herinnering dat een verantwoordelijke het gerechtvaardigde doelwit is van morele reacties als straf en beloning, woede en verontwaardiging.

Die reacties volgen niet noodzakelijk, maar als ze volgen dan is de verantwoordelijke de persoon die terecht het onderwerp van zo'n reactie is. Maar dat veronderstelt wel de mogelijkheid tot een of andere vorm van gewaarwording, het vermogen om in de brede zin van het woord te worden geraakt, hetzij op mentaal vlak hetzij op lichamelijk vlak. Het heeft geen zin iemand als verantwoordelijke aan te wijzen, als die persoon niet kan worden geaffecteerd door de morele reacties van anderen. Natuurlijk is de mogelijkheid om in de brede zin van het woord pijn of genot te ervaren niet voldoende om moreel verantwoordelijk te zijn. Dat blijkt alleen al uit onze omgang met niet-menselijke dieren: honden kunnen wel pijn en genot ervaren, maar we houden hen niet verantwoordelijkheid wanneer ze met hun staart een vaas doen omvallen. Het vermogen om lichamelijk of mentaal te worden geraakt door de reactie van een ander is echter wel een noodzakelijke voorwaarde. En aangezien artefacten zoals autonome technologieën dat vermogen momenteel niet hebben, zou het ronduit absurd zijn ze verantwoordelijk te houden voor wat ze doen.

Daartegenover staat dat morele praktijken niet noodzakelijk voor altijd vastliggen. Ze kunnen in de loop van de geschiedenis wijzigen. Denk aan rechten. Aan het eind van de achttiende eeuw argumenteerde men nog als volgt tegen vrouwenrechten: als we aan vrouwen rechten toekennen, moeten we ook dieren die geen mensen zijn rechten geven. De verzwegen aanname was dat dierenrechten ondenkbaar zijn. Intussen is het volstrekt immoreel vrouwen geen gelijke rechten als mannen toe te kennen, heeft bijvoorbeeld België drie ministers van dierenwelzijn, en wordt zelfs nagedacht over de rechten van planten. Of neem de robot Sophia waarnaar ik ook in het vorige hoofdstuk al verwees. In oktober 2017 gebeurde wat tot voor kort ondenkbaar was: Saoedi-Arabië verleende Sophia het staatsburgerschap. Welnu, als in de loop van de geschiedenis steeds meer mensen rechten kregen, als andere morele praktijken in het verleden konden veranderen, waarom zou er geen verandering kunnen zijn als het gaat over morele verantwoordelijkheid? Op dit moment kunnen we dingen niet

verantwoordelijk houden, maar zou dat in de toekomst wel mogelijk zijn?

Die vraag is alleen zinvol als het niet is uitgesloten dat robots in de toekomst op lichamelijk of mentaal vlak kunnen worden geaffecteerd, dat technologieën later in een of ander opzicht pijn of genot kunnen ervaren. Als dat vermogen nooit zal of kan bestaan, dan is het uitgesloten dat onze morele attitude zal wijzigen, dan zullen we AI nooit moreel verantwoordelijk houden. En precies dat is volgens sommigen het meest realistische scenario: we zullen technologie nooit prijzen en loven, *omdat* die nooit in staat zal zijn tot gewaarwording op lichamelijk of mentaal vlak. Er kan veel over die bewering worden gezegd. Ik beperk me tot een korte reactie op het volgende gedacht-experiment dat soms wordt gegeven om die bewering te ondersteunen.[72]

Beeld je een humanoïde robot in, een robot die dezelfde vorm heeft als die van een doorsnee menselijk lichaam. Meer nog, het is onmogelijk om op basis van de fysieke verschijning het verschil tussen mens en robot te zien, zoals in de *Black Mirror*-aflevering ‘Be Right Back’. Bovendien zit jij in het hoofd van die robot. Je moet er namelijk voor zorgen dat alle beslissingen, bewegingen en handelingen van de robot niet te onderscheiden zijn van die van een mens. Je moet met andere woorden bij ieder inkomend signaal de gepaste output verzinnen. Wordt aan de robot gevraagd naar een naam, dan moet jij het signaal geven dat de robot zichzelf als pakweg Sophia kenbaar maakt. Stel nu dat de robot van de trap valt, een situatie die bij mensen normaal gesproken gepaard zou gaan met de gewaarwording van pijn. Jij zorgt voor de output die doorgaans volgt op pijn: roepen, huilen, enzovoort. De robot ziet er als een mens uit, valt en reageert vervolgens zoals mensen dat normaal doen. Heeft de robot echter pijn? Het kan zijn dat iemand reageert op de val, bijvoorbeeld omdat het een reflex is te reageren op tekenen van pijn. Maar niemand zal reageren *omdat* de robot pijn heeft. Hoewel de robot wel tekenen van pijn vertoont, is er geen pijn, net zoals computerprogramma’s als

Google Translate en DeepL de Chinese zin niet echt verstaan die ze nochtans perfect kunnen vertalen.

Slimme softwareprogramma's moet je vergelijken met de mens in het hoofd van de robot: ze zorgen voor de output, voor de pijnsignalen. Eerst worden ze getraind met talloze data en wanneer ze vervolgens input ontvangen, voorspellen ze op basis van de leerfase wat de beste uitkomst is. Toegepast op het gedachteexperiment: een intelligente robot die eerst foto's en filmpjes van mensen met pijn analyseert, zal bij pakweg een val de signalen produceren die normaal bij mensen de pijn uitdrukken. Kunnen we dan op basis van het gedachte-experiment besluiten dat AI-systemen nooit pijn zullen kunnen hebben?

Het is goed om te onderscheiden wat het gedachte-experiment wel en niet leert. Aan de ene kant brengt het in herinnering dat pijn niet hetzelfde is als een pijnsignaal, dat het hebben van pijn niet identiek is aan het uitdrukken van pijn. Je kunt pijn hebben, zonder dat kenbaar te maken, en je kunt pijnsignalen voortbrengen terwijl je geen pijn hebt. Het experiment brengt met andere woorden aan de oppervlakte dat je uit het feit dat slimme technologie pijnsignalen geeft niet kunt afleiden dat die technologie ook effectief pijn heeft. AI kan dingen produceren die bij mensen wijzen op pijn, maar die signalen zijn in het geval van de software op zich geen voldoende reden om te concluderen dat de technologie pijn lijdt. Dat is wat we uit de case met de vallende robot kunnen opmaken. Aan de andere kant toont het gedachte-experiment *niet* dat technologieën nooit pijn zullen kunnen voelen, laat staan dat het zou bewijzen dat machines op mentaal vlak nooit zullen kunnen worden geraakt. Zeker, het is best mogelijk dat robots nooit pijn zullen kunnen hebben of iets zullen kunnen voelen, maar dat is niet wat het gedachte-experiment ons duidelijk maakt. Verder is het experiment ook uitsluitend relevant voor de software, terwijl technologieën doorgaans ook uit hardware bestaan. En laat nu precies die hardware een reden zijn om de mogelijkheid van pijn niet onmiddellijk aan de kant te schuiven. Dat is althans wat uit nog een ander gedachte-experiment naar voren komt.

Zoals alle fysiologische systemen van een menselijk lichaam bestaat ook het zenuwstelsel uit cellen, en dan hoofdzakelijk uit neuronen, die voortdurend op elkaar inwerken. Deze oorzakelijke band zorgt ervoor dat inkomende signalen leiden tot de gewaarwording van pijn. Stel nu echter dat je daadwerkelijk wordt geslagen, en dat je vervolgens gedurende een zestigtal minuten ook effectief pijn hebt – of je die pijn nu uitdrukt in lichamelijke of gedragsmatige signalen is hier niet van belang. Stel bovendien dat de wetenschap zodanig ver gevorderd is dat de neuronen kunnen worden vervangen door een prothese, microchips bijvoorbeeld, zonder dat dit voor het overige een verschil maakt. De vorm van de prothese is verschillend en ook het materiaal is anders – de chips wordt gemaakt op een plakje silicium – maar verder doen die artificiële entiteiten precies hetzelfde als de neuronen: ze sturen signalen naar andere cellen en zorgen voor de gewaarwording. Je hebt dus pijn en hebt het lichaam van een doorsnee menselijk wezen. Alleen zal gedurende zestig minuten een wetenschapper iedere cel door een microchip vervangen, waardoor je lichaam niet alleen bestaat uit cellen maar ook uit chips. Je zweeft met andere woorden als een cyborg ergens tussen natuurlijk en artificieel, een beetje zoals Arnold Schwarzenegger in *The Terminator*. Is het nog steeds volstrekte onzin te beweren dat robots ooit pijn zullen kunnen voelen?

Het voorgaande is wel een gedachte-experiment, maar tegelijk niet sterk bij de haren getrokken. Er is in het verleden al onderzoek opgezet om bij robots emoties op te wekken. Maar om verwarring te vermijden wil ik toch het volgende beklemtonen: ik beweer *niet* dat intelligente systemen ooit pijn zullen kunnen voelen, dat robots op het vlak van gewaarwording ooit zullen lijken op ons – wij, mensen. Ik heb met het laatste gedachte-experiment hoogstens willen aangeven dat het wellicht wat kort door de bocht is die optie als nonsens eenvoudigweg aan de kant te schuiven. Verder beweer ik ook niet dat, indien toch zou blijken dat AI-systemen pijn kunnen ervaren, dat we hen automatisch verantwoordelijk zullen houden voor de dingen die ze doen. De reden is dat de mogelijkheid om pijn te voelen niet

volstaat om verantwoordelijk te worden gehouden. Alleen al onze omgang met bijvoorbeeld niet-menselijke dieren toont dat aan – ik wees er eerder al op. Stel nu echter dat aan alle voorwaarden is voldaan, zou dat dan meteen ook inhouden dat we autonome AI-systemen zullen zien als kandidaten voor straf en beloning? Het is in ieder geval zo dat het toeschrijven van verantwoordelijkheid aan uitsluitend mensen een eeuwenoude morele praktijk is, die precies daarom wellicht niet snel zal veranderen. Dat laatste is althans iets wat je kunt verwachten op basis van de geschiedenis van morele rechten. Die mag dan wel tonen dat morele praktijken niet noodzakelijk voor eeuwig zijn, die geschiedenis laat ook zien dat de aloude praktijk van het toeschrijven van rechten aan enkel mensen slechts geleidelijk aan verandert ten gunste van dieren die geen mensen zijn. Alleen dat al is een reden om te vermoeden dat het toeschrijven van morele verantwoordelijkheid aan robots wellicht niet voor morgen zal zijn, zelfs indien robots op lichamelijk of mentaal vlak geraakt zouden kunnen worden door beloning of straf.

KINDSOLDATEN EN DRONES

We moeten dus terugkeren naar de centrale vraag: creëren autonome AI-systemen een verantwoordelijkheidskloof? Indien dat zo is, dan hebben we een argument voor de disruptiethese. Dat is althans wat iemand als Matthias beweert, want zulke kloof bestaat nergens anders, en bovendien zou zoiets ook onwenselijk zijn. Maar creëren slimme systemen zulke kloof? Technologieën zelf kun je dus vandaag niet verantwoordelijk houden, maar geldt dat ook voor mensen?

Er is reden om te vermoeden dat je mensen wel kunt verantwoordelijk houden voor fouten gemaakt door autonome AI. Denk aan een legerofficier die een kindsoldaat inschakelt. Het kind wordt een wapen in de hand gestopt om tegen de vijand te vechten, maar wat blijkt? Het kind brengt onschuldige burgers om het leven, en begaat dus een oorlogsmisdaad. Wellicht niemand zal beweren dat het kind verantwoordelijk is voor de burgerslachtoffers, maar naar alle waarschijnlijkheid meent ook niemand dat niemand verantwoordelijk

is. Normaal gesproken houden we de officier moreel verantwoordelijk voor de misdaad. Is er echter een verschil tussen deze casus en het gebruik van autonome AI-systemen, *killer robots* bijvoorbeeld? Zo ja, is dat verschil relevant? Kindsoldaten zijn mensen, robots niet. In beide gevallen echter neemt een persoon een beslissing in de wetenschap dat onwenselijke situaties kunnen volgen en dat men daarover geen controle meer kan uitoefenen. Als de officier moreel verantwoordelijk is, waarom zou hetzelfde dan niet gelden voor wie beslist autonome AI-systemen te gebruiken? Zijn *killer robots* en andere autonome AI-systemen op dat vlak iets bijzonder of uitzonderlijk?

Mocht aan de andere kant blijken dat iemand als Matthias gelijk heeft, dat er dus wel degelijk een *responsibility gap* is, dan zou ook dat waarschijnlijk niet geheel verwonderen. Dat komt onder meer omdat morele verantwoordelijkheid een gradueel karakter heeft: ze kan toenemen of afnemen. Denk bijvoorbeeld aan de trainer van een sportclub. Het spreekt voor zich dat hij of zij minstens gedeeltelijk verantwoordelijk is voor de prestaties van het team. Betekent dat echter ook dat de mate van verantwoordelijkheid altijd gelijk is? Nee, de coach kan in mindere mate dan vorig seizoen verantwoordelijk worden gehouden voor de slechte prestaties van zijn of haar team, bijvoorbeeld omdat het bestuur een aantal goede spelers heeft laten vertrekken en in plaats daarvan spelers heeft gekocht die minder sportieve kwaliteiten hebben. Wat heeft dat nu te maken met de zogeheten verantwoordelijkheidskloof?

Ik maak het concreter en spits toe op oorlogstechnologie. Stel, je bent soldaat en doodt een terreurverdachte. Als je daarvoor een klassiek wapen hebt gebruikt dat functioneert zoals het hoort, een 9mm-pistool bijvoorbeeld, dan ben jij zonder enige twijfel geheel – of in ieder geval in grote mate – verantwoordelijk voor de dood van de verdachte. Stel echter dat je dezelfde persoon wilt doden, en dat je enkel een semiautomatische drone hebt. Je zit in een kamer die ver is verwijderd van het oorlogsgebied waar de verdachte zich bevindt, en je geeft de drone alle informatie over de persoon die je zoekt. De drone is in staat

om het gebied zelf te verkennen, en als de technologie aangeeft dat het zoekproces voorbij is, kun je het resultaat van de zoekactie beoordelen en vervolgens beslissen of de drone wel of niet moet vuren. Op basis van de informatie die je hebt verzameld, geef je het bevel om te vuren. Maar wat blijkt? De gedode persoon is niet de terreurverdachte en is dus ten onrechte gedood. Die fout heeft alles te maken met een productiefout, die tot een mankement in de werking van de drone heeft geleid. Natuurlijk houdt dat niet in dat jij in geen enkel opzicht moreel verantwoordelijk bent voor de dood van de verdachte. Er is dus geen *responsibility gap*, maar waarschijnlijk zullen de meeste mensen het gerechtvaardigd vinden dat je minder verantwoordelijk bent dan wanneer je een 9mm-pistool gebruikte. Dat heeft te maken met het feit dat de beslissing om te vuren is gebaseerd op informatie die niet van jezelf maar van de drone komt, informatie die bovendien niet correct is.

De afname van de oorzakelijke rol van de soldaat via de technologie gaat dus voor velen gepaard met een vermindering van de verantwoordelijkheid. Het overhevelen van een activiteit – het inwinnen van informatie – brengt met zich mee, niet dat de mens niet verantwoordelijk is, maar wel dat hij in mindere mate verantwoordelijk is. Dat voedt het vermoeden dat het overhevelen van *alle* beslissingen op AI-systemen tot de zogeheten *responsibility gap* leidt. Maar klopt dat vermoeden? Zo nee, waarom niet? Met die vragen ben ik aanbeland bij de kern van mijn analyse van het vraagstuk over morele verantwoordelijkheid en AI.

VOORWAARDEN VOOR VERANTWOORDELIJKHEID

Zoals eerder aangekondigd, wil ik een conservatief standpunt innemen. Hoewel autonome AI-systemen nieuw zijn, zijn de verschillen met klassieke niet-autonome technologieën niet van dien aard dat we voor fouten met de eerste *niemand* verantwoordelijk kunnen houden, en voor fouten gemaakt met de tweede wel iemand. De bestaande praktijk wordt niet onderbroken door nieuwe

technologieën; er is ook op het vlak van verantwoordelijkheid geen morele disruptie. Ik vermoed dat velen daarmee wel akkoord gaan, maar dat ze het tegelijk niet makkelijk vinden om haarfijn uit te leggen waarom ze precies akkoord gaan, waarom minstens iemand verantwoordelijk is voor de fouten gemaakt door een autonoom systeem. De filosoof kan daarbij helpen. Om kort te zijn: zeker iemand is verantwoordelijk, en dat komt omdat de klassieke voorwaarden voor verantwoordelijkheid ook nu zijn vervuld. Ik wees eerder al op het vermogen tot gewaarwording in de brede zin van het woord, maar welke andere voorwaarden moeten zijn vervuld opdat iemand verantwoordelijk kan worden gehouden? Ik onderscheid oorzakelijke verantwoordelijkheid, autonomie en kennis.

Het is vanzelfsprekend – ik stipte het enkele bladzijden terug al aan – dat morele verantwoordelijkheid oorzakelijke verantwoordelijkheid veronderstelt. Wie in het geheel niet is betrokken bij het ontstaan van het (ongewenste) resultaat van een handeling kan voor dat resultaat niet verantwoordelijk worden gehouden. In het kader van AI voldoen verschillende mensen aan deze voorwaarde: de programmeur, de fabrikant, de gebruiker. Dat betekent echter niet dat we de stelling van de verantwoordelijkheidskloof hebben ondermijnd. Niet elk oorzakelijk verband is relevant voor morele verantwoordelijkheid, niet elke betrokkenheid gaat gepaard met morele verantwoordelijkheid. Denk aan de wetenschapper in het laboratorium van daarnet. Als zij of hij valt en als gevolg daarvan de dood van een paar collega’s veroorzaakt, houden we die wetenschapper niet in morele zin verantwoordelijk, enkel in oorzakelijke zin.

Er is dus meer nodig. Morele verantwoordelijkheid vereist ook autonomie. Ik breng in herinnering dat je dat begrip op minstens twee manieren kunt begrijpen. Wanneer iemand het over autonomie heeft, dan kan men dat in negatieve zin invullen. Wie autonoom is in dat opzicht, kan geheel zelfstandig, onafhankelijk functioneren. Voor mijn redenering is alleen de tweede, positieve vorm relevant. Deze variant betekent dat je de zaken tegen elkaar kunt afwegen, en dat je op basis daarvan zelf een beslissing kunt nemen.

Het feit dat je in staat bent om te beraadslagen en te beslissen is echter niet voldoende om moreel verantwoordelijk te worden gehouden. Het kan bijvoorbeeld zijn dat je de gegronde beslissing neemt om de koning te vermoorden, maar wanneer de koning wordt vermoord, ben je daar niet per se verantwoordelijk voor, bijvoorbeeld omdat iemand anders dat doet net voor het moment dat jij de trekker wilt overhalen en onafhankelijk van jouw beslissing. Je bent alleen verantwoordelijk als je beslissing ook oorzakelijk is verbonden met de moord; de weloverwogen beslissing moet aan de basis van de moord liggen. Met andere woorden: wanneer iemand moreel verantwoordelijk is, veronderstelt dat een oorzakelijk verband tussen de autonomie en de handeling. Het is alleen door die link dat er betrokkenheid bij de handeling is en dat de autonomie relevant is voor morele verantwoordelijkheid.

Kennis is de laatste voorwaarde. Je kunt alleen verantwoordelijk worden gehouden als je over de nodige relevante kennis beschikt. Wie niet weet dat een handeling fout is, kan daar niet verantwoordelijk voor zijn. En als de gevolgen van een handeling onvoorzien zijn, dan kun je daar ook niet voor worden gestraft. Let wel, de afwezigheid van kennis pleit je niet per se vrij. Het kan namelijk zijn dat je bepaalde dingen niet weet, terwijl je die eigenlijk wel had moeten weten. Geformuleerd in termen van verantwoordelijkheid: als het je rolverantwoordelijkheid is om van bepaalde zaken op de hoogte te zijn, maar je bent het niet, en dat gebrek aan kennis leidt tot ongewenste resultaten, dan ben jij moreel verantwoordelijk voor het ongewenste resultaat. Om een eenvoudig voorbeeld te nemen: als de bestuurder van een auto door een rood licht rijdt en daardoor een ongeval veroorzaakt, dan is die bestuurder moreel verantwoordelijk voor dat ongeval, zelfs als blijkt dat hij of zij geen weet had van het verbod om door een rood licht te rijden. Het is namelijk je plicht als burger en autobestuurder – lees: je rolverantwoordelijkheid – om van die regel op de hoogte te zijn.

ALLES ONDER CONTROLE

Wie dus is betrokken bij het gebruik van een technologie, wie met een weloverwogen beslissing aan de basis staat van dat gebruik, en wie weet heeft van de mogelijke gevolgen van die technologie, die kun je moreel verantwoordelijk houden voor alles wat misgaat met de technologie. Dat is althans wat de klassieke analyse inhoudt.

Matthias echter, en dat is nu cruciaal, meent dat aan een extra voorwaarde moet worden voldaan. Als een handeling of bepaalde loop van gebeurtenissen eenmaal in gang is gezet, moet je daar volgens hem ook controle over hebben. Dus zelfs als je betrokken bent, bijvoorbeeld omdat je de beslissing hebt genomen dat de handeling of loop van gebeurtenissen plaats moet vinden, terwijl je er verder niets kunt aan doen op het moment dat ze werden geïnitieerd, dan zou het oneerlijk zijn jou te straffen wanneer dat alles uitmondt in een onwenselijk resultaat. Welnu, omdat AIsystemen geheel zelfstandig functioneren, op zo'n manier dat je geen invloed kunt uitoefenen op hun beslissingen, dan kun je volgens iemand als Matthias niemand verantwoordelijk houden voor de gevolgen.

Tussen haken: het is belangrijk in gedachten te houden dat deze redenering is gebaseerd op een aanname, namelijk dat iedere betrokkene zijn of haar rolverantwoordelijkheid heeft genomen. Stel, ik ben ontwerper. Ik moet dus nadenken over de mogelijke negatieve gevolgen van mijn ontwerp en erop anticiperen in de mate van het mogelijke. Doe ik dat niet, dan neem ik met andere woorden mijn rolverantwoordelijkheid niet, en precies dat leidt ertoe dat het systeem een fout veroorzaakt. In dat geval spreekt het voor zich dat ik verantwoordelijk kan worden gehouden voor dat onwenselijke gevolg. Die morele verantwoordelijkheid volgt rechtstreeks uit mijn gebrek aan rolverantwoordelijkheid, zelfs wanneer ik geen controle meer heb. Ook Matthias beaamt dat.

Er zijn zaken die de argumentatie van Matthias lijken te ondersteunen. Als je verantwoordelijk wordt gehouden voor een handeling, dan wil

dat doorgaans zeggen dat je controle hebt. Ik ben als CEO verantwoordelijk voor de slechte cijfers van mijn bedrijf, want ik had andere beslissingen kunnen nemen die het bedrijf meer ten goede kwamen. Omgekeerd heb ik over een groot aantal zaken geen controle en draag ik in deze context geen verantwoordelijkheid.
Ik heb bijvoorbeeld geen enkele controle over de weersomstandigheden en heb ook geen enkele verantwoordelijkheid voor de gevolgen van het goede of slechte weer.
Verantwoordelijkheid gaat dus vaak gepaard met controle, net zoals de afwezigheid van controle meestal gepaard gaat met de afwezigheid van verantwoordelijkheid. Toch is het onjuist dat je controle over een geïnitieerde handeling of loop van gebeurtenissen moet hebben om verantwoordelijk te worden gehouden, en dat geen controle hebben je verantwoordelijkheid wegneemt. Er wordt wel vaak gezegd dat we vandaag over zo goed als alles controle willen en dat controle voor ons, modernen, heel belangrijk is, toch vertelt onze morele praktijk iets anders. Controle is ook vandaag niet zó belangrijk dat het onze morele verantwoordelijkheid uitwist als we geen controle hebben over de dingen. Dat blijkt onder meer uit het volgende.

Beeld je in dat je in Gentbrugge woont en dat je met een paar vrienden een avond wilt doorbrengen in een café in Wondelgem, een tiental kilometers verderop in het noorden van Gent. Het regent. Omdat je niet nat wilt worden, besluit je de auto te nemen. Je vertrekt, maar na een aantal minuten heb je een epileptische aanval. Dat is niet de eerste keer, want je bent epilepsiepatiënt. Het gevolg van de aanval is dat je de controle over het stuur kwijt raakt, waardoor de auto van de weg raakt en uiteindelijk een fietser ernstig wordt verwond. Het is niet zeker dat je zult worden gestraft, laat staan dat je een strenge straf zult krijgen, maar wellicht zullen weinigen of zelfs niemand je niet verantwoordelijk houden voor de verwonding van de fietser, en dat ondanks het gebrek aan controle over het stuur van de auto. Waarom?

Allereerst bezit je alle relevante kennis. Je weet niet dat binnen een aantal minuten een aanval zal plaatsvinden, maar als patiënt weet je

wel dat het risico op een aanval bestaat en dat zoiets gepaard kan gaan met een ongeval. Verder ben je ook autonoom (in positieve zin). Je bent in staat om een aantal zaken tegen elkaar af te wegen – de wens om naar een café te gaan, niet nat te worden, het risico op een aanval – en om op grond daarvan een beslissing te nemen. Ten slotte stap je doelbewust in de auto. Daardoor ben je oorzakelijk verbonden met het ongewenste gevolg op een manier die de morele verantwoordelijkheid voldoende grondt. Wanneer je immers een beslissing neemt, wetende dat die tot ongewenste gevolgen kan leiden, dan is het gerechtvaardigd je als kandidaat voor straf te beschouwen op het moment dat het ongewenste gevolg zich ook echt voordoet. Nogmaals, het is niet zeker dat een straf zal volgen, maar wie een risico neemt, is verantwoordelijk voor dat risico, en kan dus worden gestraft wanneer blijkt dat het onwenselijke gevolg daadwerkelijk plaatsvindt.

Het is belangrijk om te weten wat je daar nu wel en niet uit kunt afleiden dat relevant is voor mijn betoog. Het enige wat je uit het voorgaande kunt concluderen, is dat geen controle hebben je morele verantwoordelijkheid niet absorbeert. Verantwoordelijkheid vereist geen controle, en dus kun je niet zeggen dat autonome AI gepaard gaat met een verantwoordelijkheidskloof *omdat* je geen controle hebt over de technologie. Uit het voorgaande kun je echter niet concluderen dat de idee van een *responsibility gap* in het geval van autonome AI niet klopt, dat er in alle gevallen iemand verantwoordelijk is voor de fouten veroorzaakt door die technologie.
Immers, misschien zijn in het geval van autonome beslissingssystemen de andere voorwaarden voor morele verantwoordelijkheid niet vervuld en moeten we toch concluderen dat het gebruik van autonome AI hand in hand gaat met een verantwoordelijkheidskloof, zelfs als blijkt dat controle niet is vereist voor morele verantwoordelijkheid. Maar is dat ook zo?

POTJE BREKEN, POTJE BETALEN

Naast mijn kritiek op Matthias wil ik dus nog een extra punt maken. Ik wil laten zien dat je het gebruik van die hypergeavanceerde technologie op dezelfde manier moet begrijpen als de situatie met de epilepsiepatiënt van daarnet. Er is minstens iemand verantwoordelijk wanneer autonome AI fouten maakt – misschien is er zelfs collectieve verantwoordelijkheid maar omdat het volstaat om één verantwoordelijke te vinden om de stelling van een verantwoordelijkheidskloof te ondermijnen hoef ik niet in te gaan op de idee van gedeelde verantwoordelijkheid. Om te laten zien dat het niet klopt dat niemand verantwoordelijk is, beroep ik me op een aantal eerder aangehaalde voorbeelden: een *killer robot* doodt een burger, een zelfrijdende auto rijdt een fietser omver.

Om te beginnen is het belangrijk om te weten dat beide dramatische ongelukken het resultaat zijn van een lange keten van gebeurtenissen die zich uitstrekt van de vraag naar productie, over de zoektocht naar financiering tot ten slotte de programmering. Als we op zoek zijn naar een kandidaat-verantwoordelijke kun je wellicht meerdere personen aanwijzen – ik denk bijvoorbeeld aan de ontwerper of producent – maar de meest voor de hand liggende persoon is de gebruiker: de commandant die beslist de *killer robot* in te zetten tijdens een conflict en de inzittende in de autonome auto. Het is gerechtvaardigd om hen als kandidaat voor straf naar voren te schuiven omwille van de volgende redenen, net zoals de epilepsiepatiënt verantwoordelijk is voor het letsel van de fietser.

Allereerst kennen beiden de gebruikscontext en weten ze wat de mogelijke ongewenste gevolgen kunnen zijn. Ze weten niet of een ongeluk zal plaatsvinden, laat staan waar en wanneer precies. Autonome auto's en wapens zijn immers gebaseerd op *machine learning*, wat met zich meebrengt dat het niet (altijd) te voorspellen is welke beslissing zal worden genomen. Maar het soort van ongelukken dat kan gebeuren, is niet onbeperkt. Het doden van burgers en het vernielen van hun huizen *(killer robot)*, het omverrijden van een fietser

of het inrijden op een groep mensen (zelfrijdende auto) zijn dramatisch maar wel te voorzien; als gebruiker weet je dat zulke zaken kunnen gebeuren. En als je het niet weet, dan schiet je tekort: je zou het moeten weten. Het is je plicht, je rolverantwoordelijkheid, om na te gaan wat de mogelijke negatieve gevolgen zijn van de dingen die je gebruikt.

Ten tweede zijn zowel commandant als inzittende voldoende autonoom. Ze zijn in staat om de voor- en nadelen tegen elkaar af te wegen: de kans op een oorlogsmisdaad en minder doden in eigen gelederen *(killer robot)*, de kans op verkeersslachtoffers en het kunnen werken tijdens de verplaatsing (zelfrijdende auto). Wanneer nu, ten derde, op basis van deze zaken wordt besloten om effectief over te gaan tot het gebruik van autonome auto's en wapens, in de wetenschap dat dat ongewenste gevolgen met zich mee kan brengen, dan is het gerechtvaardigd zowel de commandant als inzittende verantwoordelijk te houden voor de ongewenste voorziene gevolgen. Wie risico's neemt, accepteert de verantwoordelijkheid voor dat risico, dat hij of zij mogelijk zal worden gestraft in het geval dat het ongewenste voorziene gevolg zich ook daadwerkelijk voordoet.[73]

Het gebruik van autonome AI is dus op het vlak van verantwoordelijkheid in overeenstemming met een bestaande morele praktijk. Net zoals je mensen verantwoordelijk kunt houden voor het gebruik van niet-autonome technologieën, zijn mensen ook verantwoordelijk voor zaken waarover ze geen controle hebben maar waarmee ze wel op een relevante manier zijn verbonden. De autonomie van technologie wist dus niet alleen de rolverantwoordelijkheid van de gebruiker niet uit, wel integendeel, daarnaast lost het ook de morele verantwoordelijkheid niet op. Kortom, en daarmee pik ik de rode draad van dit hoofdstuk opnieuw op, ook wanneer het over morele verantwoordelijkheid gaat, leveren AI-systemen geen argument dat in het voordeel van de disruptiethese pleit. De weg die het systeem aflegt om een beslissing te nemen, mag dan misschien wel volkomen ondoorzichtelijk zijn voor de gebruiker,

het systeem creëert geen *responsibility gap* en dus ook geen nieuw moreel probleem.

Wie daarmee niet akkoord gaat, moet ofwel aantonen wat er mis is met de bestaande morele praktijk waarin we mensen verantwoordelijkheid toeschrijven ofwel het verschil aantonen tussen morele verantwoordelijkheid in het geval van autonome systemen en de dagelijkse morele praktijk. Met dat laatste doel ik op de *relevante* verschillen. Want natuurlijk zijn er verschillen tussen het gebruik van een autonoom systeem enerzijds en het rijden met een wagen als patiënt anderzijds. De vraag is echter of die verschillen van belang zijn wanneer het gaat over morele verantwoordelijkheid.

HANDELEN ONDER DWANG

Wanneer het over AI gaat, zijn er dus geen zeven maar zes hoofdzonden. En net zoals onder meer bias en privacy is morele verantwoordelijkheid geen argument dat de disruptiethese ondersteunt. Dat is echter alleen waar voor zover mijn analyse op de voorbije bladzijden steek houdt. Maar neem even aan, ter afronding van dit hoofdstuk, dat wat ik zopas heb beargumenteerd niet klopt: er is wél een verantwoordelijkheidskloof. Stel dat de analyse van daarnet op verschillende plaatsen hapert, en dat je wel degelijk niemand verantwoordelijk kunt houden voor de schade veroorzaakt door speelgoedrobot AIBO, de zelfrijdende auto van Google, het rekruteringssysteem van Amazon of de *killer robot* van het Amerikaanse leger. Zou dat in dat geval wel een argument voor de disruptiethese opleveren? Zou een *responsibility gap* in het geval van autonome AI met andere woorden uniek zijn?

Als de *responsibility gap* bestaat in het geval van AI, bestaat ze niet alleen in die context. Wanneer een bliksem inslaat op een boom die vervolgens op mijn auto valt, gaat dat wel gepaard met schade maar kun je daar niemand verantwoordelijk voor houden. Bestaat die *gap* ook wanneer mensen zelf oorzakelijk betrokken zijn bij een ongewenste gebeurtenis? Stel, je bent bankbediende en iemand

overmeestert je met een pistool. Met het pistool tegen je hoofd word je gedwongen om de kluis te openen. Hoewel je dat niet wilt, vrees je echter voor je leven en open je de kluis voor de overvaller die met een hele hoop geld gaat lopen. Omdat je wordt gedwongen op straffe van de dood zal uiteraard niemand jou verantwoordelijk houden. Maar is er geen *responsibility gap*: de overvaller moet namelijk worden gestraft.

Laat ik een wijziging aanbrengen. Je bent nog steeds de bankbediende, je opent opnieuw de kluis en je laat veel geld verdwijnen. En hoewel je nu niet door iemand anders wordt gedwongen, handel je ook nu toch niet autonoom. Je hebt voor het eerst waanvoorstellingen. Er is een stem die jou dwingt om de kluis te openen en met het geld aan de haal te gaan. Je bent dus oorzakelijk verantwoordelijk voor de roof, maar men kan daar niemand moreel voor verantwoordelijk houden. Althans niet als men alle feiten kent; als men weet in welke mentale toestand je je bevond, lost je verantwoordelijkheid in rook op. Bovendien beroofde je de bank 's nachts, toen niemand aanwezig was in de bank. En ook daarna, toen je het geld verstopte, handelde je geheel alleen. Er is niemand die jou ook maar een beetje heeft geholpen. Conclusie? Er is een onwenselijke situatie waarvoor je wel een verantwoordelijke wilt maar niet kunt aanwijzen, niet omdat je niet weet wie de verantwoordelijke is, maar omdat er geen verantwoordelijke is. Het zou volkomen ongerechtvaardigd zijn ook maar iemand te straffen voor wat is gebeurd. Mocht AI met andere woorden gepaard gaan met een verantwoordelijkheidskloof, dan zou zulk gegeven niet nieuw zijn.

MET HET OOG OP DE TOEKOMST

Laten we aannemen dat ik het ook nu bij het verkeerde eind heb. Het is niet alleen onjuist dat AI niet gepaard gaat met een verantwoordelijkheidskloof, die kloof bestaat ook *uitsluitend* in de context van AI. Zouden we daarmee een argument hebben dat de disruptiethese op ethisch vlak ondersteunt? Zou die vermeende unieke gap met andere woorden ook moreel problematisch zijn? Om

die vraag te beantwoorden, kijk ik naar twee rechtvaardigingen voor het bestaan van de praktijk van het toeschrijven van morele verantwoordelijkheid. De eerste heeft te maken met preventie, de tweede wijst op de symbolische betekenis van straffen.

Iemand berooft een bank, een soldaat doodt een burger, een autobestuurder negeert een rood licht: dat zijn telkens voorbeelden van onwenselijke situaties waarvan je als maatschappij niet wilt dat ze in de toekomst nog zullen plaatsvinden. Om dat te voorkomen, om ervoor te zorgen dat later de wet niet opnieuw wordt overtreden, bestaat zoiets als het toeschrijven van verantwoordelijkheid, een morele praktijk die is gestoeld op het psychologische mechanisme van klassieke conditionering. Na een overtreding wordt iemand verantwoordelijk gehouden, is hij of zij kandidaat voor een onaangename bejegening, en dat met als doel dat de overtreding in de toekomst niet meer zal plaatsvinden.

Dat doel, preventie, moet er vanzelfsprekend zijn, en het is duidelijk dat het middel, de verantwoordelijke straffen, vaak volstaat om het doel te bereiken. Maar toch hangt preventie niet per se samen met straffen, is het straffen van de verantwoordelijke niet noodzakelijk met het oog op preventie. Er zijn andere manieren dan straffen om ervoor te zorgen dat dezelfde fout niet opnieuw wordt gemaakt. Je kunt mensen leren zich aan de regels te houden, bijvoorbeeld door ze extra uitleg te geven, het goede voorbeeld te geven, enzovoort. Sterker nog, het is mogelijk dat onwenselijke situaties in de toekomst niet meer plaatsvinden zonder dat er een moreel verantwoordelijke is. Dat blijkt precies het geval te zijn in de context van AI.

Neem een algoritme dat de sollicitatiebrieven van alle vrouwen aan de kant legt of het platform Amazon Mechnical Turk dat je account ten onrechte blokkeert, waardoor je geen jobs meer kunt aanvaarden. Om te voorkomen dat zich dat in de toekomst herhaalt, is het vanzelfsprekend dat aan het AI-systeem wordt geknutseld door iemand met voldoende technische kennis, bijvoorbeeld de programmeur. Het is best mogelijk dat het systeem zoveel lagen heeft

waardoor de ontwerper geen zicht krijgt op het probleem en het dus niet kan herstellen. Maar het is ook mogelijk dat de programmeur wel succesvol kan interveniëren, in die mate dat het AI-systeem de fout in de toekomst niet meer zal maken. In dat geval is het technische werk voldoende voor het voorkomen van het probleem en heb je verder voor het gegeven doel, preventie, niemand meer nodig die een kandidaat voor straf is – ik herinner eraan dat dit de omschrijving van verantwoordelijkheid is. Met andere woorden, als het doel louter preventief van aard is, dan kan de uitsluitend technische tussenkomst van de ontwerper volstaan en is dus de vermeende afwezigheid van een moreel verantwoordelijke geen probleem.

RESPECT VOOR DE MENS

Er is nog een ander doel dat vaak wordt aangehaald om het toeschrijven van verantwoordelijkheid te rechtvaardigen. Dat doel heeft een symbolisch karakter. Het gaat namelijk over het respecteren van de waardigheid van een mens. Hangt dat doel vast aan het aanwijzen van een kandidaat voor straf? Zou een *responsibility gap* in het licht van dat objectief een probleem zijn?

In een liberale democratie heeft iedereen morele waarde. Welke eigenschappen je ook karakteriseren en wat je ook doet, je hebt *moral standing*, en die is voor iedereen gelijk, althans in principe. Die waarde hangt niet in de lucht, of beter gezegd, zij doet dat niet voor zover voorschriften aan die waarde worden vastgeknoopt die in overeenstemming met onze morele status zijn. Het principe dat ieder mens morele waarde heeft, houdt in dat jij rechten hebt en dat anderen plichten tegenover jou hebben. Je hebt onder meer recht op onderwijs en werk, en anderen mogen je niet beledigen en zonder goede reden pijn doen. Het is toegestaan dat een werkgever jou niet aanneemt op grond van relevante criteria, maar het druist flagrant in tegen jouw status van wezen met *moral standing* dat men jou tijdens een sollicitatiegesprek kleineert of belachelijk maakt.

Stel je voor dat dat laatste toch gebeurt. Dat is een probleem, want het is een ontkenning van het feit dat jij morele waarde hebt. Welnu, de praktijk van het toeschrijven van morele verantwoordelijkheid moet je minstens ten dele als een antwoord op zo'n probleem zien. Je moet dat als volgt begrijpen. Er vindt iets onwenselijks plaats – de waardigheid van een persoon wordt geschonden –, en als reactie daarop wordt iemand gestraft, of die persoon wordt op z'n minst aangewezen als kandidaat voor een straf. Dat wil zeggen: iemand wordt pijn gedaan en ervaart een onaangename gewaarwording, iets dat je uit jezelf niet wenst. Het doel van die straf, die onaangename ervaring, is nu dat wordt onderstreept dat de schending van de waardigheid een morele fout was, en dus dat de waardigheid van het slachtoffer wordt bevestigd. De straf heelt de wonde niet en maakt de fout ook niet ongedaan, maar ze heeft wel symbolisch belang. Ze doorstreept de ontkenning van de morele status die in de misdaad besloten lag.

De bevestiging van de morele waarde is duidelijk een goed, een doel dat je kunt realiseren aan de hand van een straf. Alleen is het maar de vraag of dat doel *uitsluitend* langs deze weg kan worden bereikt. Stel, een *killer robot* doodt een soldaat. Stel bovendien dat het waar is, in tegenstelling tot wat ik daarnet betoogde, dat niemand verantwoordelijk kan worden gehouden voor deze dood. Wil dat dan zeggen dat de morele waarde van de soldaat niet meer kan worden onderstreept? Het klopt dat het aanwijzen van een verantwoordelijke betekent dat de waarde van de soldaat serieus wordt genomen. Bovendien is het ongetwijfeld wenselijk dat, uit respect voor de waarde die elk individu heeft, iemand als kandidaat voor straf kan worden aangewezen. De bewering echter dat de verantwoordelijkheid voor de erkenning van de waardigheid *noodzakelijk* is, is onjuist. Je kunt ook recht doen aan de overledene zonder dat iemand verantwoordelijk wordt gehouden. Misschien wel het bekendste en meest voor de hand liggende voorbeeld daarvan is een begrafenis. De betekenis van dit ritueel ligt immers vooral in het feit dat het onderstreept dat de overledene waarde op zich heeft.

Voor alle duidelijkheid: ik beweer niet dat het toeschrijven van verantwoordelijkheid een zinloze praktijk is. Ook wil ik niet zeggen dat indien het gebruik van AI tot een gap zou leiden dat de onmogelijkheid om iemand verantwoordelijk te houden *nooit* een probleem zou zijn. Mijn punt is dat preventie en respect op zich geen reden zijn om te besluiten dat een verantwoordelijkheidskloof in de context van AI per se een moreel drama is.

Ter afronding

Wellicht zijn er nogal wat mensen – ik denk dan voornamelijk aan ondernemers en technologen – die de stelling verdedigen dat AI disruptieve technologie is omdat men daar belang bij heeft, vooral economisch belang. Ontwerpers en bedrijven willen een nieuwe technologie op de markt brengen, en om mogelijke gebruikers en kopers te overtuigen, beweert men dat het nieuwe ontwerp een bestaand domein op z'n kop zal zetten of ontwrichten. Indien die hypothese klopt, dan is wel duidelijk dat het hier over een welbepaalde invulling van de disruptiethese gaat, een over de positieve ontwrichtende effecten van AI. De interpretatie waarop ik me in dit hoofdstuk heb gefocust heeft een eerder negatieve invulling. Ik zoomde in op de problemen veroorzaakt door slimme technologie, en dan meer bepaald de morele problemen. Kort gesteld is de interpretatie dat AI geheel nieuwe morele problemen veroorzaakt, problemen die er nog niet waren toen er nog geen AI bestond. Zulke stelling is natuurlijk geen koren op de molen van techno-optimisten, maar past eerder in het wereldbeeld van alarmisten en technopessimisten. Ik beweer niet dat er in de toekomst geen nieuwe morele problemen zullen ontstaan, maar de ethische problemen waarmee AI vandaag kampt, zijn niet nieuw. Privacy, bias en transparantie bijvoorbeeld zijn de thema's waarover men het vandaag vooral heeft als het over AI-ethiek gaat, maar problemen met privacy, bias en transparantie zijn niet het privilege van AI. Kortom, AI is geen disruptieve technologie, althans niet voor zover het over ethiek gaat.

Dat is de stelling die ik in dit hoofdstuk naar voren heb geschoven en verdedigd, en die niet opgaat in de stroom van modieus alarmisme.

Wij geven vorm aan onze gebouwen en daarna geven onze gebouwen vorm aan ons.

Winston Churchill, 28 oktober 1943

3

De motor van de samenleving

In een interview met het blad *Playboy* in 1969 beweerde mediawetenschapper Marshall McLuhan, bekend van onder meer de boutade *the medium is the message*, dat computers in de nabije toekomst het leven van veel mensen zullen kunnen orkestreren. Machines, aldus de Canadese denker, zullen in staat zijn om de belangrijkste mediakanalen over te nemen, zelf berichten te schrijven en die onder de bevolking te verspreiden. Intussen weten we dat McLuhans voorspelling niet bij de haren getrokken was – denk aan de chatbot TAY (voluit *Thinking About You*) van Microsoft die in maart 2006 via Twitter seksistische en racistische berichten de wereld instuurde. Interviewer Eric Norden echter was op dat ogenblik, in 1969, toch wat onder de indruk van wat McLuhan beweerde. Hij vroeg of we als mensheid nog controle zullen hebben over de technologie. Of, zo wierp Norden op, zal technologie de mensheid controleren? McLuhan antwoordde als volgt: ‘Ik zie geen mogelijkheid van een wereldwijde rebellie van luddieten die alle machines in stukken zullen slaan, dus we kunnen net zo goed achteroverleunen en zien wat gebeurt en wat met ons zal gebeuren in een cybernetische wereld. Je afzetten tegen een nieuwe technologie zal de vooruitgang ervan niet stoppen.’[74]

Net zoals vele anderen die over technologie en AI spreken, schuwde McLuhan de sterke en scherpe beweringen niet (en ook optredens in de media niet overigens – hij had een kort gastoptreden in de film *Annie Hall* uit 1977 van Woody Allen). Maar zijn uitspraken in het intussen befaamde interview met het blootblad zijn wel veelzeggend. Wanneer hij zegt dat niets of niemand, zelfs de (neo-)luddieten niet, de onverstoorbare vernieuwing van technologie een halt kan toeroepen, dan is die uitspraak een illustratie van een stelling die net

als de neutraliteitsthese en disruptiethese erg populair is onder ingenieurs en wetenschappers – ‘luddiet’ verwijst naar Ned Ludd, die eind achttiende eeuw weefmachines zou hebben vernietigd als reactie op mogelijk verlies aan jobs door automatisering. Ik heb het hier over de zogeheten determinismethese, die over technologie en andere aanverwante zaken spreekt in termen van determineren, dat wil zeggen: in termen van noodzakelijk of onvermijdelijk. Het is ook die vaakgehoorde stelling die ik in dit hoofdstuk onder de microscoop leg.

Om misverstanden te vermijden is het goed om meteen duidelijkheid te scheppen. Er bestaan minstens vier interpretaties van de determinismethese. De interpretatie van McLuhan is daar één van. Ze luidt, zo kun je uit het citaat van hierboven afleiden, dat de ontwikkeling van technologie zich onvermijdelijk voortzet. Eens een technologie op de markt is gezet, kan het niet anders dan dat dat uitmondt in de productie van nieuwe technologie. De tweede interpretatie is verwant aan de eerste en schuift het volgende naar voren: alle technologie is noodzakelijkerwijs ontstaan; dat een technologie wordt uitgevonden is onvermijdelijk. Deze invulling vind je, net zoals de eerste overigens, bijvoorbeeld in *What Technology Wants* uit 2010 van Kevin Kelly, de voormalige hoofdredacteur van het technologietijdschrift *Wired*. De derde invulling van de determinismethese gaat niet zozeer over dingen die we ‘technologie’ noemen, maar over een technologische, instrumentele manier om naar de werkelijkheid te kijken. Ze luidt dat zo’n denkwijze alle domeinen van de samenleving heeft ingepalmd, zodanig dat we vandaag de dag op geen andere dan de instrumentele manier naar de dingen kunnen kijken. Wellicht de bekendste vertegenwoordiger van die interpretatie is de filosoof die ik in het eerste hoofdstuk al aanhaalde: Heidegger. In zijn bekende tekst *Die Frage nach der Technik* uit 1954 beweert hij ‘dat […] de techniek het noodlot van onze tijd zou zijn, waarbij noodlot betekent: het onontkoombare van een onveranderlijke afloop.’[75] De laatste interpretatie ten slotte gaat over de band tussen technologie en samenleving. Niet alles wat we vandaag de dag doen wordt bepaald door apparaten of de platformen

van Google en Amazon, maar technologie determineert wel interpersoonlijke verhoudingen en sociale processen. Die bewering vind je onder meer in het beroemde essay ‘Do Machines Make History?’ uit 1994 van Robert L. Heilbroner: ‘Ik denk dat we inderdaad kunnen stellen dat de technologie van een samenleving een bepaald patroon van sociale betrekkingen oplegt aan die samenleving.’[76]

VAN DRUKPERS TOT PROTESTANTISME

Straks leg ik die vier interpretaties van de determinismethese nog uitgebreider uit. In het verdere vervolg van de inleiding sta ik uitsluitend stil bij de laatste, die gaat over technologie en de sociale effecten ervan. Dat moet volstaan als opwarming.

De interpretatie van de determinismethese over sociale effecten doet onder meer denken aan *De Staat* van Plato. Daarin laat hij een zekere Glauco de mythe van Gyges vertellen. Gyges is een brave herder die een gouden ring vindt en merkt dat die magische eigenschappen bezit: wanneer je de zegel van de ring naar de handpalm draait, word je onzichtbaar. De herder gebruikt die ring om tot het hof door te dringen, de koningin te verleiden, de koning te doden en ten slotte de troon te bestijgen. Glauco dist de mythe op om duidelijk te maken dat ook een rechtschapen mens uiteindelijk de morele verboden met voeten zal treden wanneer daar voordeel uit te halen valt. Het lijdt geen twijfel dat mensen met een verdorven ziel de ring zullen gebruiken om hun belangen veilig te stellen, maar ook wie het doorgaans goed voorheeft met de wereld, zoals de herder Gyges, zal niet nalaten dat te doen. Kortom, de mogelijkheid om een artefact (de ring) te gebruiken, zal in alle gevallen tot hetzelfde resultaat leiden. Het feit dat men de ring naar de handpalm kan draaien, zal telkens uitmonden in een morele puinhoop, los van de menselijke inborst, of ruimer, los van de cultuur.

Toch werd die versie van de determinismethese pas populair rond het einde van de negentiende en het begin van de twintigste eeuw, en dat

onder invloed van onder anderen econoom en socioloog Thorstein Veblen. De stelling dat technologie de samenleving determineert, dook in verschillende contexten op, bijvoorbeeld in de reclamewereld. In april 1920 werd in de *Ladies Home Journal* een advertentie geplaatst met een foto van het nieuwste strijkijzer (de *Simplex Ironer*) naast de oude versie, en dit met de melding dat de nieuwe technologie een hele groep vrouwen gelukkiger zal maken.[77]

Het denkbeeld van technologie die de samenleving bepaalt, keert ook terug, al dan niet terecht, wanneer over de feodale samenleving wordt nagedacht. Sommigen beweren dat de introductie van het buskruit uit China niet anders kon dan mee de neergang inluiden van het sociale en politieke regime van de middeleeuwen. Dat zou onder meer te maken hebben met het feit dat ridders te paard een makkelijk doelwit waren voor met pistolen gewapende soldaten. Omwille van dat risico gebruikten ridders minder en minder de paarden en stijgbeugels die ze van boeren in ruil voor bescherming konden lenen. Wapentechnologie, zo luidt de bewering, zou op die manier onvermijdelijk hebben bijgedragen tot de ondermijning van het systeem van leenheer en leenman.

Een ander en vaak aangehaald voorbeeld van de determinismethese is de bewering dat de boekdrukkunst in de vijftiende eeuw wel tot de Reformatie moest leiden. In de periode voorafgaand aan die uitvinding was de Bijbel doorgaans alleen toegankelijk voor de clerus. Maar door de bekende innovatie van Johannes Gutenberg kregen steeds meer mensen toegang tot Gods woord, en dat moest volgens sommigen wel leiden tot de bloei van het protestantisme en het schisma in het christendom.

Aan het eind van de negentiende eeuw stond men vrij argwanend tegenover de eerste telefoon. Men dacht dat mensen door die technologie elkaar sowieso minder vaak fysiek zouden opzoeken, waardoor sociale contacten zonder enige twijfel oppervlakkiger zouden worden. Gezinnen zullen verspreid wonen en mensen zullen enkel nog spreken met gelijkgestemden, hun werk elektronisch doen

en elkaar alleen nog bij ceremoniële gelegenheden ontmoeten, zo luidde het toen. Tussen haken merk ik op dat deze reactie doet denken aan hoe sommigen op de opkomst van sociale media in het begin van deze eeuw reageerden.

Tot slot wordt de determinismethese over de sociale effecten van technologie ook in filosofische teksten verdedigd of wordt ze er regelmatig aangehaald. In *Misère de la philosophie* uit 1847 van Karl Marx staat bijvoorbeeld het volgende: ‘De handmolen geeft je de samenleving met de feodale heer, de stoommachine de samenleving met de industriële kapitalist.’[78] Een andere illustratie vind je in *Le système technicien* van Jacques Ellul. In die bekende studie uit 1977 kun je het volgende lezen: ‘De techniek is autonoom ten opzichte van de economie en de politiek. Techniek lokt sociale, politieke en economische veranderingen uit en conditioneert ze. De techniek is de motor van al het andere, ondanks elke schijn van het tegendeel en ondanks de menselijke trots, die pretendeert dat de filosofische theorieën van de mens nog steeds bepalende invloeden zijn en de politieke regimes van de mens nog steeds beslissende factoren in de technische evolutie.’[79]

DE TECHNOLOGISCHE CONDITIE

Op het eerste gezicht lijkt er wel wat te zeggen voor de determinismethese over de sociale effecten van technologie. Er zijn nauwelijks of geen plaatsen waar je geen uurwerken vindt. Ontwaak je uit je slaap, stap je in de auto, open je een laptop, kijk je naar een kerk: het lijkt erop dat je niet kunt ontsnappen aan digitale of analoge uurwerken. En meer algemeen gebruiken we technologie om te eten (magnetron), sporten (fiets), slapen (Bose Sleepbuds II), verhuizen (lift), studeren (laptop), genieten (vibrator), schrijven (pen), communiceren (telefoon), adverteren (sociale media), reizen (vliegtuig), genezen (MRI-scan) of eten (fornuis). De cyborg is wel een curiosum, maar tegelijk ook een uitvergroting van wat alle of de meeste mensenlevens karakteriseert, namelijk dat eenieders leven

door en door met technologie is vervlochten. De menselijke conditie is een technologische conditie. Is dat geen reden om te vermoeden dat technologie onvermijdelijk sociale effecten heeft?

Toch zijn er ook zaken die erop wijzen dat de relatie tussen samenleving en technologie gecompliceerder is. Ik denk aan de eerste computer, de Colossus. Die werd tijdens de Tweede Wereldoorlog ontworpen door ingenieur Tommy Flowers, die zich daarvoor beriep op het werk van wiskundige en informaticus Alan Turing. Dat was in een militaire context. Het doel was om zo de geheime codes van het Duitse leger te kunnen ontcijferen. Tijdens de middeleeuwen vond men dat in kloosters niet afzonderlijk maar gezamenlijk tot God moest worden gebeden, ook tijdens de nacht. Omdat mensen niet spontaan op hetzelfde ogenblik wakker worden, ontstond daardoor de behoefte aan een wekkermechanisme. De eerste mechanische klok zou rond de tiende eeuw gemaakt zijn door monnik Gerbert van Aurillac, die later ook bekendheid verwierf als paus Sylvester II. Dus naast de computer lijkt ook de wekker een argument tegen de determinismethese. Beide laten zien dat de technologie ontstaat onder invloed van een sociale context, die een hetzij religieus hetzij militair karakter heeft. Het zijn geen voorbeelden van technologieën die de samenleving determineren, het zijn voorbeelden van technologieën die een product zijn van de samenleving.

Twee andere voorbeelden: de fiets en ruimtevaart. In zijn vaak geciteerde studie *Of bicycles, bakelites, and bulbs: toward a theory of sociotechnical change* uit 1995 laat techniekssocioloog Wiebe Bijker zien dat er in het begin uiteenlopende soorten fietsen waren en dat die diversiteit samenhing met het verschil in smaak en voorkeur van mensen.[80] De zogeheten Hoge Bi, een fiets met een groot voorwiel, was vooral in trek bij jonge mannen, omdat je er een hoge snelheid mee kon halen en indruk mee kon maken op vrouwen. Vrouwen vonden die gevaarlijk en zelfs onzedig; voor hen werd een lager model met gelijke wielen ontworpen. Tot slot is er ook vanuit de ruimtevaart reden om te twijfelen aan de versie van de determinismethese over

sociale effecten. Naar het einde van de jaren 1980 verminderde de financiering voor de ontwikkeling van technologieën om de ruimte te verkennen. Dat had veel te maken met het einde van de Koude Oorlog, met de afname van de nationalistische spanning tussen de Verenigde Staten en de USSR, het toenmalige Rusland.

Daarmee hebben we intussen al een idee van het debat over de determinismethese, althans over één van de vier invullingen. Determineren machines de maatschappij? Bepaalt AI het uitzicht van onze samenleving? Besturen programmeurs de wereld? Als je daar ook de drie andere interpretaties van de determinismethese bij neemt, dan wordt duidelijk dat het straks ook over de volgende vragen zal gaan. Klopt het dat een technologische instrumentele zienswijze onze cultuur domineert? En zo ja, is dat wel zo problematisch als vaak wordt aangenomen? Daarnaast zullen ook deze vragen aan bod komen. Ontstaan technologieën onvermijdelijk? Leidt de ene technologie noodzakelijkerwijs tot de andere? Of is technologie eerder een antwoord op een vraag die vanuit de samenleving komt? Dat zijn de vragen waar dit hoofdstuk om zal draaien, een hoofdstuk dat opnieuw aan hoofdzakelijk twee kapstokken is opgehangen, aangevuld met een derde.

Allereerst moet ik op de volgende bladzijden een aantal zaken verhelderen. Wat betekent ‘determineren’ bijvoorbeeld wel en niet, en wat moet je precies verstaan onder ‘de determinismethese’? Ik verduidelijk wat de vier interpretaties wel en niet inhouden. Zonder het zien van die meerzinnigheid is het onmogelijk of op z’n minst erg moeilijk om het debat over de verhouding tussen technologie en samenleving goed te begrijpen en voeren. Ten tweede zal ik ook evalueren. Ik argumenteer dat geen enkele van de vier interpretaties klopt. Niet alle technologie ontstaat noodzakelijk; het is duidelijk dat niet elke technologische evolutie onvermijdelijk is; een instrumentele zienswijze is niet in beton gegoten en het is ook niet zeker dat dit altijd problematisch is; en als technologie sociale effecten heeft, volgen die niet altijd noodzakelijk. Tot slot laat ik zien waarom we volgens mij moeten nadenken over de vragen die ik stel. Als het bijvoorbeeld klopt

dat technologie de samenleving niet determineert maar zelf wordt bepaald door sociale factoren, dan houdt dat in dat de wereld er anders had kunnen uitzien. Ook al valt het nog te bezien of dat een reden tot veel optimisme is, hieruit kun je wel opmaken dat in dit hoofdstuk een thematiek terugkeert die veel ouder is dan het denken over technologie, namelijk die van noodzaak en contingentie. Is alles wat nu bestaat onvermijdelijk zo? Of kan het ook anders?

Determinisme gedetermineerd

Zoals aangegeven zijn er vier invullingen van de determinismethese, vier beweringen waarin over technologie of dingen die daarmee te maken hebben in termen van determineren wordt gesproken. Maar wat houdt ‘determineren’ precies in? Dat is niet altijd even duidelijk. Dat komt wellicht ten dele omdat die term, net zoals ‘rechtvaardigheid’ of ‘verantwoordelijkheid’, verschillende zaken kan betekenen, en omdat in de context van een filosofische reflectie op technologie ‘determineren’ een specifieke invulling heeft. We doen er dus goed aan om eerst uit te leggen wat die term hier wel en niet betekent.

Vaak worden ‘bepalen’ en ‘vastleggen’ als synoniemen voor ‘determineren’ gegeven. Maar dat helpt ons nog niet zoveel vooruit, want wat wordt daar precies mee bedoeld? Een eerste duidelijke betekenis van ‘determineren’ is ‘identificeren’. Het is in die zin dat bijvoorbeeld biologen de term gebruiken wanneer ze het over een plant hebben. Ze bedoelen met ‘determineren’ in dat geval dat moet worden nagegaan tot welke soort die plant behoort. De determinismethese heeft echter niets te maken met deze invulling, wel met de tweede. Wanneer je die andere betekenis gebruikt, verwijs je met ‘determineren’ naar een relatie tussen twee zaken, meer bepaald naar een oorzakelijke relatie. Het betekent dat iets, een proces bijvoorbeeld, een effect heeft op iets anders, of dat iets nieuws ontstaat ten gevolge van dat proces. Deze interpretatie wordt gebruikt

wanneer men zegt dat levensstijl een determinant van gezondheid is. De wijze waarop je leeft, determineert je gezondheid: het heeft een effect op je gezondheid, een effect dat al dan niet positief is.

De tweede betekenis van ‘determineren’ valt echter niet volledig samen met ‘veroorzaken’. Dat blijkt bijvoorbeeld uit het volgende. Wanneer je zegt dat je interesse voor poëzie een gevolg is van het feit dat je ouders vroeger thuis veel poëzie lazen, betekent dat niet hetzelfde als de bewering dat jouw interesse gedetermineerd is door de interesse van je ouders. ‘Determineren’ gaat wel over een relatie van oorzaak en gevolg, maar die omschrijving volstaat niet; een nauwkeurigere sterkere omschrijving is vereist. Die gaat als volgt. Met ‘determineren’ wordt verwezen naar een bepaald type van effect dat uitgaat van een determinant, namelijk een noodzakelijk effect. Het betekent dat een toestand onvermijdelijk tot een bepaald gevolg leidt; dat effect moet wel volgen uit die toestand; de oorzaak garandeert het effect. Zelfs als we in staat zouden zijn om de tijd terug te spoelen, zouden we zien dat precies dezelfde begintoestand exact hetzelfde effect zou sorteren. De relatie tussen koken en verdampen bijvoorbeeld heeft zo’n noodzakelijk karakter. Het is niet louter zo dat verdamping een effect is van het koken; het koken van water kan niet anders dan verdamping tot gevolg hebben. Of neem de gaswet van Boyle: wanneer het volume van de lucht in een fietspomp afneemt, stijgt onvermijdelijk de luchtdruk.

Enkele eeuwen geleden dacht wiskundige Pierre-Simon Laplace nog dat alles gedetermineerd is, dat alle oorzakelijke relaties noodzakelijk zijn, dat alle toestanden een onvermijdelijk gevolg zijn van vorige toestanden. Nu wordt daar nog weinig of geen geloof aan gehecht. Sommige oorzakelijke relaties hebben overduidelijk een noodzakelijk karakter, maar sommige duidelijk niet; bepaalde effecten zouden er evengoed niet kunnen zijn. Neem bijvoorbeeld de invloed van ouders op kinderen op het vlak van muziekkeuze. De interesse voor The Beatles van nogal wat mensen volgt uit het feit dat hun ouders vroeger veel naar The Beatles luisterden. Toch is dat effect niet in beton gegoten. Die interesse had er ook niet kunnen zijn, zelfs wanneer de

ouders eertijds ieder weekend *Abbey Road* uit 1969 op de platenspeler legden.

Het is hier nu niet van belang of ze allemaal even plausibel zijn of überhaupt steek houden, maar tot slot wil ik erop wijzen dat er zeker vier vormen van determinisme zijn, dat er dus vier vormen van determinisme zijn met telkens een andere oorzaak: een genetische variant, omgevingsdeterminisme, psychologisch determinisme en een technologische versie.[81] Het is uiteraard enkel die laatste die straks onze aandacht zal opeisen. De vier versies van de determinismethese die ik daarnet in de inleiding aanhaalde, hebben immers allemaal met de technologische versie van determinisme te maken.

Genetisch determinisme betekent dat een gen dat codeert voor een eigenschap garandeert dat een organisme die eigenschap zal vertonen. Wetenschapper Walter Gilbert verdedigde deze theorie. Hij zei ooit dat we het fenotype van een persoon kunnen voorspellen, louter door diens DNA op een cd-rom te analyseren. Tussen haken: vandaag zijn er nauwelijks nog wetenschappers die daar geloof aan hechten. Genen maken een effect hoogstens waarschijnlijk, ze garanderen het niet. Naast genetisch determinisme bestaat ook omgevingsdeterminisme. Dat wil zeggen dat je op grond van externe invloeden de ontwikkeling van mensen kunt vastleggen. John Watson, de grondlegger van het behaviorisme, moet je in die context situeren. In 1926 schreef hij dat als een ouder wenst dat zijn of haar kind arts wordt, dat hij ervoor kan zorgen dat het een arts wordt, en dat los van de talenten, afkomst of wensen van het kind. Een vorm van psychologisch determinisme vind je in populariserende inleidingen tot de freudiaanse psychoanalyse. Daar lees je dat de al dan niet bewuste verwerping door het kind van het zogeheten oedipuscomplex resulteert in een seksuele perversie. Als je het incestverbod niet aanvaardt, zo gaat de theorie, dan word je fetisjist, masochist, frotteurist. De determinismethese ten slotte, en dat is nu het belangrijkste, gaat over de technologische variant. Die these gaat met andere woorden over een relatie waarin een effect onvermijdelijk tot stand komt, en over de plaats die technologie, of zaken die daar nauw

mee verbonden zijn, in die relatie inneemt. Zoals al aangestipt, vallen de vier interpretaties van de determinismethese onder het technologisch determinisme. Dus niet enkel de bewering dat technologie noodzakelijk deze of gene sociale effecten heeft, is daar een voorbeeld van, maar ook de theorie die gaat over het instrumentele denken dat onze tijd zou karakteriseren. Het is met die laatste versie dat ik nu ook start.

De kolonisering van de leefwereld

Als de determinismethese op meerdere manieren kan worden begrepen dan houdt dat in dat de argumenten die in deze context worden gebruikt altijd een argument voor of tegen een *welbepaalde* invulling van de determinismethese zijn. Het is belangrijk dat voor ogen te houden, al is het maar wegens het gevaar van spraakverwarring. Iemand zou immers kunnen argumenteren voor de determinismethese en iemand tegen, maar telkens op grond van een andere interpretatie van die these, waardoor het een dovemansgesprek is.

Zoals al aangegeven zijn er vier invullingen van de determinismethese. Eén gaat over de sociale effecten van technologie, twee andere over de ontwikkelingsgeschiedenis van technologie, en dan meer bepaald het ontstaan en de evolutie van technologie. Ik verduidelijk en evalueer die verderop in dit hoofdstuk nog uitgebreid. Ik begin met de interpretatie van de determinismethese die aansluit bij een populaire, wijdverspreide opvatting van onze cultuur. Die luidt dat we vandaag alles enkel nog zien als een middel om onze doelen te realiseren en verlangens te bevredigen.

Die korte omschrijving laat wel onmiddellijk al zien waarom deze versie van technologisch determinisme in zekere zin een buitenbeentje is. Het gaat hier immers niet in de eerste plaats over technologieën, over dingen met een functie, maar over een manier

van kijken en denken. Toch laat men het onder technologisch determinisme vallen, omdat het gaat over een *technologische* manier om naar de dingen te kijken, en omdat die zienswijze in termen van determineren wordt beschreven. Er is dus reden om de dingen die over een technologische zienswijze worden gezegd in dit hoofdstuk te bespreken, maar er is ook reden om die dingen eerst en onafhankelijk van de andere versies van de determinismethese te bespreken.

HEIDEGGER EN HET FASCISME

De interpretatie van determinismethese waarmee ik start, vind je onder meer in de werken van de al genoemde denkers Ellul en Heidegger. Ik kies voor die laatste om de eerste versie van technologisch determinisme te schetsen, omdat hij een van de beroemdste (techniek)filosofen van de voorbije eeuw is en omdat hij veel invloed heeft gehad tijdens en na zijn leven. Niettemin doet de keuze voor de Duitse filosoof bij sommigen wellicht de wenkbrauwen fronsen. Was hij niet de filosoof die lid was van de Nationalsozialistische Deutsche Arbeiterpartei (NSDAP)? Moet je vanuit die wetenschap nog aandacht besteden aan het denken van de Duitse filosoof?

Het klopt dat Heidegger niet aan de juiste zijde van de politieke geschiedenis stond. Toch is dat geen reden om zijn denken in het algemeen en zijn teksten over technologie in het bijzonder links te laten liggen. Wat me interesseert, zijn ideeën, beweringen, opvattingen, theorieën. Zijn ze plausibel? Zijn ze juist? Het antwoord op die vragen mag niet worden beïnvloed door de auteur van een uitspraak. Ik haalde het in het eerste hoofdstuk al aan: een theorie is niet juist of onjuist, niet meer of minder fout, omdat ze door Simone de Beauvoir, Kim Clijsters of Barack Obama werd geformuleerd. Ook de morele eigenschappen of politieke keuzes van de auteur van een theorie zijn in deze context irrelevant. Uitspraken van een fascist kunnen waar zijn, zelfs al zijn ze van een fascist. Een fascist kan uiteraard onware dingen zeggen, maar die beweringen zijn niet fout *omdat* ze van een fascist komen.

Hoewel Heideggers foute politieke beslissingen dus geen reden zijn om niet in te gaan op zijn denken, is mijn vraag: klopt de populaire overtuiging dat we vasthangen aan een technologische, instrumentele blik op de wereld? Mijn punt is dat Heideggers versie van technologisch determinisme niet zonder problemen is. Om dat aannemelijk te maken, moet in de eerste plaats duidelijk zijn wat zijn versie inhoudt.[82]

DE ESSENTIE VAN DE DINGEN

Heideggers techniekfilosofie is deel van een striemende kritiek op de geschiedenis van de filosofie – hij heeft het uitsluitend over de westerse wijsbegeerte, de niet-westerse ontsnapt aan zijn aanval. Die geschiedenis laat Heidegger niet aanvangen bij Plato in de vierde eeuw voor het begin van onze jaartelling, maar het is volgens Heidegger wel fout gelopen met de filosofie sinds Plato. Sinds dan denken filosofen na over alles wat bestaat, en dan vooral over mensen. Zij doen dat op een specifieke manier: ten eerste, door te focussen op wat iets wel en niet is, en twee, door te zoeken naar een grond van alles wat bestaat. Ik verklaar me nader.

Van Plato via Thomas van Aquino tot Descartes, filosofen hebben zich altijd al afgevraagd wat de essentie van iets is, zij stellen zich sinds lang al zogenaamde ‘wat-is-vragen’. Toegepast op de mens betekent dit dat wordt gezocht naar een eigenschap die kenmerkend is voor alle organismen die we ‘mensen’ noemen. Daarnaast wil het ook zeggen dat filosofen sinds meer dan tweeduizend jaar zoeken naar unieke eigenschappen, dat wil zeggen: kenmerken die uitsluitend bij mensen voorkomen. Bij Aristoteles leidde dat bijvoorbeeld tot zijn bekende omschrijving van de mens als *animal rationale*. Volgens hem hebben alle mensen het vermogen om te redeneren en zou het ook dat vermogen zijn dat hen onderscheidt van dieren die geen mensen zijn.

De geschiedenis van de filosofie is echter meer dan het telkens opnieuw zoeken naar een essentie. Zeker tot aan het begin van de moderne filosofie rond de achttiende eeuw volstond het voor de meeste filosofen niet om te bepalen wat iets is. Daarnaast dachten ze ook na over iets dat al het overige overstijgt, dat wil zeggen: iets dat al het bestaande rechtvaardigt. In de loop van de geschiedenis is dat doorgaans God geweest; Hij is diegene die alles wat bestaat – van dieren tot planten – bestaansrecht geeft, ondersteunt. Mensen hebben grote of kleine verlangens, maar uiteindelijk bestaat alle leven ter wille van God.

Heidegger wilde radicaal breken met dat type van reflectie. Want filosofen mogen dan wel nagedacht hebben over God of over unieke en universele eigenschappen, daarmee is nog niet duidelijk op welke manier mensen in het leven staan. De vraag die Heidegger daarom naar voren schoof, en waar ook een groot deel van zijn oeuvre om draait, is niet *wat* iets is, maar *hoe* iets is. In zijn woorden: de kerntaak van de filosofie is niet om te zoeken naar een essentie, maar naar existentie. Hoe staan mensen in het leven? Wat is menszijn?

DE TECHNOLOGISCHE IMPERATIEF

Een belangrijk deel van zijn antwoord op de vraag naar hoe wij vandaag in het leven staan, is dat een flink deel van ons alledaagse doen en laten niet wordt gestuurd door expliciete reflectie, maar door een spontaan verstaan van de werkelijkheid. Wanneer ik tijdens het koken een ui moet snijden, vorm ik me niet eerst een voorstelling van een mes, van hoe het eruitziet en waartoe het dient. Nee, ik grijp op een onnadenkende manier naar het mes, omdat ik een mes onwillekeurig begrijp als iets dat dient om te snijden. Dit is een voorbeeld van wat ook meer in het algemeen het geval is, zo meent Heidegger. De wijze waarop we nu leven, wordt gestuurd door een spontaan begrip, niet alleen van mijn kookvaardigheden, maar van de werkelijkheid tout court.

Dat begrip heeft een historisch karakter. Heidegger heeft dat sterk beklemtoond. Alles verandert: samenlevingen, mensen, maar ook opvattingen. Het kader van waaruit we vandaag de dag de wereld bekijken, verschilt van een middeleeuws kader. Natuurlijk, het is onmogelijk om een moment aan te wijzen waarop het premoderne in het moderne begrijpen omslaat, maar dat neemt niet weg dat er een verschil is tussen hoe we nu de wereld benaderen en hoe dat vroeger werd gedaan. In de zestiende eeuw kon een Spaanse rechtbank nog oordelen dat een rivier niet mocht worden rechtgetrokken (en konden niet-menselijke dieren nog ter verantwoording worden geroepen op een rechtbank). Alles is immers een product van God, en dus zou het onrespectvol zijn aan Gods werken te morrelen. Tegenwoordig is zo'n oordeel zo goed als ondenkbaar, en dat komt doordat we de werkelijkheid niet meer begrijpen als een effect van een bedoeling, laat staan van een goddelijke wil.

Als in het premoderne tijdperk alles in religieuze zin werd verstaan, wat is er dan eigen aan een *moderne* blik? Heidegger, die op dat punt werd beïnvloed door schrijver Ernst Jünger, meent dat we vandaag door een technologische bril naar de wereld kijken. Wat bedoelde hij daarmee?

In het eerste hoofdstuk heb ik erop gewezen dat we de term 'waarde' vaak in instrumentele zin gebruiken. Iets heeft zulke waarde, wanneer het bijdraagt aan de realisering van een doel. Alles wat valt onder de noemer 'technologie' is ontworpen met een doel voor ogen en heeft dus per definitie instrumentele waarde (op voorwaarde dat het niet defect is). Wanneer Heidegger beweert dat wij de werkelijkheid technologisch opvatten, wil hij daarmee zeggen dat ons tijdsgewricht wordt gekenmerkt door een instrumentele blik op de realiteit. Sport, kunst, data, we zien die zaken volgens hem als een instrument. Sport is een middel om fit of aantrekkelijk te zijn, kunst een instrument om de wereld een geweten te schoppen, en de data die we op het web achterlaten dienen om AI-systemen te voeden of zijn een middel voor techbedrijven om winst te maken. We kijken naar de dingen in de wereld als instrumenten om onze projecten te verwezenlijken, wensen

te bevredigen, targets af te vinken. Het nutsdenken, dat is wat onze moderne tijd karakteriseert, althans volgens Heidegger. Het communisme en kapitalisme zijn verschillende politieke systemen, maar wat hen verbindt, is een instrumentele kijk op de werkelijkheid.

Het spreekt voor zich dat ontwerpers en AI-ontwikkelaars voornamelijk in termen van nut denken. Het is immers hun taak om technologieën te ontwerpen, zaken die een middel voor een doel zijn. Minder voor de hand liggend is dat ook op de terreinen waar niet alles in de eerste plaats rond technologie draait het nutsdenken is doorgedrongen. Nochtans is het dat wat volgens Heidegger ons tijdsgewricht kenmerkt. Het denken in termen van middelen en doelen beheerst niet alleen de techwereld, dat denken is alle domeinen van de samenleving binnengeslopen. We worden gekolonialiseerd door een instrumentele blik – om de terminologie van filosoof Jürgen Habermas te gebruiken. Van de politieke wereld tot de cultuursector, van de gezondheidszorg tot het onderwijs, overal is de instrumentele manier van denken aanwezig. Er is niets dat aan die blik ontsnapt, de mens incluis, en ook de natuur wordt geïnstrumentaliseerd. Steenkool dient de industrie, wind wordt gebruikt om elektriciteit op te wekken, bergen om te ontspannen na een druk werkjaar, mooie landschappen om een selfie op Instagram aantrekkelijker te maken, en de Rijn, aldus Heidegger in een beroemde passage, is een middel om de turbines van een waterkrachtstation aan te drijven.

Heidegger gaat echter nog verder. Uit het vorige kun je niet afleiden dat alle sferen worden beheerst door een technologische manier van denken. Want dat denken mag dan wel zijn doorgedrongen tot alle domeinen, dat wil nog niet zeggen dat een technologisch kader het enige is van waaruit naar de dingen wordt gekeken. Niettemin beweert Heidegger, om het citaat uit de inleiding opnieuw op te rakelen, dat het instrumentele denken als het ware als een noodlot boven ons hoofd hangt. Het denken in termen van middelen en doelen beheerst alle domeinen en heeft alle andere zienswijzen verdrongen. Het gevolg is dat iedere mens in de greep is van een blik waaraan niet te ontsnappen valt en dat het voor iedereen onmogelijk is om op een

niet-instrumentele manier naar de wereld te kijken. Pakweg de natuur kunnen we *uitsluitend* nog als middel verstaan, en niet als iets dat ook op zich waarde heeft. De woekering van de imperatief van het nutsdenken: daarover gaat Heideggers techniekfilosofie.[83]

TECHNOLOGISERING EN TECHNOCRATISERING

Het is duidelijk dat in Heideggers tekst geen technologieën centraal staan – vliegtuigen, auto's, keukenrobots – maar wel een technologische manier van denken, een kader, een blik. Heidegger vestigt niet de aandacht op onze technologische conditie, het feit dat ons bestaan is doordrongen van technologieën. Hij wijst op het denkkader dat in de moderniteit over de wereld is geschoven en op de technologische imperatief die zich overal en altijd laat gevoelen.

Toch is Heideggers techniekfilosofie, net zoals Elluls denken, een schoolvoorbeeld van de determinismethese, althans van een bepaalde invulling ervan. Om die interpretatie goed te begrijpen, breng ik in herinnering dat determineren over een oorzakelijke band gaat waarbij het gevolg noodzakelijk is, waarbij de oorzaak het gevolg garandeert. Heideggers technologisch determinisme laat zich daarom als volgt formuleren. Onze cultuur wordt beheerst door een technologische manier van denken – dat is de oorzaak. Er is een manier van denken die niet in handen is van een persoon, partij of bedrijf, maar die zich over alle domeinen van de samenleving heeft uitgespreid en de hele samenleving omspant, op zo'n manier dat er geen plaats meer is voor een andere dan de instrumentele manier van kijken. Het gevolg is dan dat wij, modernen, onvermijdelijk in termen van middelen en doelen denken, en dat we uitsluitend nog in die termen naar de wereld kunnen kijken. Gegeven het feit dat ons tijdsgewricht zo doordrongen is van een instrumentele kijk, moet het wel zo zijn dat we vanuit dat perspectief nadenken. Dat is zo als het over kunst of politiek gaat, maar evenzeer wanneer het de natuur of het onderwijs betreft.

Ter afronding wil ik nog twee zaken onder de aandacht brengen. Ten eerste geeft Heidegger geen verklaring voor het vermeende instrumentalisme van onze tijd. Hij beschrijft *wat* volgens hem het geval is, maar niet *waarom* dat zo zou zijn, niet wat de oorzaak van het primaat van de technologische blik zou zijn. Wel meent hij dat die blik niet het gevolg is van onze technologische conditie. Dat we enkel nog in instrumentele termen zouden kunnen denken, is geen gevolg van het feit dat onze wereld doordrongen is van technologieën. Eerder het omgekeerde is het geval, aldus Heidegger. We begrijpen de wereld enkel in termen van nut, en *daarom* worden steeds meer technologieën gemaakt; ons bestaan is sterk verbonden met technologie, juist *omdat* we uitsluitend in instrumentele zin zijn gaan nadenken.

Verder kun je Heideggers denken niet aanvallen door erop te wijzen dat er tegenwoordig bij beslissingsprocessen, bijvoorbeeld in de politiek, wel nog steeds rekening wordt gehouden met morele waarden of dat belangengroepen ook nu nog altijd een stem hebben. Dat komt omdat technologisering niet hetzelfde is als technocratisering. Heidegger meent dat onze cultuur doordrongen is van een technologische manier van denken en dat wil hier zeggen: door een instrumentele manier van denken. Die omschrijving valt echter niet samen met de karakterisering van onze cultuur als een technocratische cultuur. Dat laatste betekent dat de beslissingen worden overgelaten aan experts, mensen met erg gespecialiseerde kennis, die zich uitsluitend laten leiden door feiten en niet door waarden of belangen.

VAN HEIDEGGER NAAR GOOGLE

Nogal wat mensen voelen zich aangetrokken tot Heideggers betoog. In zekere zin is dat niet verwonderlijk. Er zijn immers ontzettend veel voorbeelden van technologisch, instrumenteel denken. Wanneer bijvoorbeeld kunstenaars subsidies aanvragen, moeten zij uitleggen waartoe hun projecten zullen leiden. Dient het om nog meer fondsen

te verwerven? Zal het ons land meer op de kaart zetten? Zullen daar andere sectoren dan de kunstsectoren van profiteren? Als je niet positief op die vragen kunt antwoorden, maak je nauwelijks of zelfs geen kans op financiering. Ook op andere terreinen zijn er tal van voorbeelden van nutsdenken: topsporters streven een selectie voor het nationale team na omdat dit hun kansen op een transfer naar het buitenland doet toenemen; studenten lopen in het buitenland stage om zo een betere positie op de arbeidsmarkt te verwerven; onderzoekers geven lezingen op conferenties met de bedoeling om hun curriculum vitae te versterken; er is kritiek op het doceren van Latijn, Grieks en kunstgeschiedenis in het secundair onderwijs, niet omdat het niet interessant is, maar omdat het tot niets of nauwelijks iets dient, zo menen sommigen.

Het kader dat Heidegger heeft geschetst is intussen al enkele decennia oud, maar toch passen ook hedendaagse fenomenen in dat kader. Niet alleen een selectie of stage, maar ook wij, mensen, en meer in het bijzonder onze data, worden als een middel gezien. De sporen die we voortdurend op het web achterlaten, zijn een instrument voor geweldgewin. Dat blijkt althans uit het bedrijfsmodel dat rond 2000 werd geïntroduceerd door de oprichters van Google: Sergey Brin en Lawrence Page, beter bekend als Larry Page.

In 1995 leerden beide heren elkaar kennen aan Stanford University, zowat de hofleverancier van techmensen aan Silicon Valley. Een jaar later ontwierpen ze het intussen bekende algoritme PageRank. Dat algoritme bracht orde in het tot dan onoverzichtelijke internet, door het zoeken naar informatie gerichter te laten verlopen. Als Page en Brin het daar toen bij hadden gelaten, dan zou de wereld er nu heel anders hebben uitgezien. Oorspronkelijk was dat ook de bedoeling. In een paper uit 1998 schrijven ze nog dat Google in de eerste plaats een zoekmachine voor academici is en dat adverteren in die context niet past. Maar van dat idee zijn ze snel afgestapt. Dat komt omdat ze van Google de zoekmachine, Google de geldmachine wilden maken. In 1999 probeerden ze het bedrijf van de hand te doen, onder meer aan Altavista en Yahoo. Maar toen dat mislukte, besloten Page en Brin om

toch over te schakelen naar *targeted advertising*: bedrijven kopen ruimte op Googles webpagina's die ze kunnen vullen met gepersonaliseerde boodschappen. Dat bleek een schot in de roos. In 2001 stegen de inkomsten van 19 naar 86 miljoen dollar, in 2004 zat men aan meer dan drie miljard dollar winst. Tegenwoordig zou je Google als een reclamebedrijf kunnen zien, samen met Facebook het grootste. Naar schatting 70% van alle onlineadvertenties worden op de platformen van beide bedrijven geplaatst. Al wil dat uiteraard niet zeggen dat online adverteren tot meer consumeren leidt – dat is althans wat onderzoek suggereert, zo stipte ik in het vorige hoofdstuk al aan.

In zo'n zakenmodel neemt het belang van data toe, en zijn wij eigenlijk niets anders dan een verzameling gegevens. Wij, mensen, 'inforgs'.[84] Want hoe meer gegevens men over jou heeft – hoe beter men weet wat je voorkeuren en wensen zijn, wat je wilt doen en zult doen –, des te verfijnder de algoritmen, en dus hoe beter bedrijven of andere adverteerders hun reclameboodschappen kunnen afstemmen op de potentiële consument. Dat is de reden waarom datahandelaars als Experian en Axiom bestaan, bedrijven die ons offline overal volgen (geotracking) en ons onlinegedrag onophoudelijk surveilleren, elke aankoop of 'like', en die deze gegevens vervolgens doorverkopen aan adverteerders. Het is ook de reden waarom onder meer Google Maps werd ontwikkeld en waarom platformen als Facebook voortdurend naar onze aandacht hengelen en ons zo veel mogelijk online willen hebben. We weten intussen overigens dat men in dat laatste goed geslaagd is (toegegeven, ook uit eerste hand). Wie een account op Facebook heeft, zou gemiddeld vijftig minuten per dag op het platform doorbrengen; Instagram en Snapchat zouden het met minder moeten doen: elke dag worden de apps respectievelijk twintig en dertig minuten gebruikt.[85]

Ook onze data zijn dus opgenomen in een instrumentele logica. Platformen willen onze gegevens om die aan adverteerders te verkopen; adverteerders zoeken naar de juiste consumenten en

hebben daarom zoveel interesse voor onze data. Daarmee hebben we opnieuw een voorbeeld van nutsdenken, een voorbeeld dat Heidegger, meer dan vijf decennia terug, zelf niet had kunnen bevroeden. Bewijst dat zijn gelijk? Als er zoveel casussen zijn die getuigen van een instrumentele logica, bewijst dat niet dat we gevangen zitten in die logica?

VOORBIJ HET NUT

Heideggers versie van de determinismethese kampt met zeker twee problemen. Het eerste heeft te maken met de overtuigingskracht van de argumenten. Het tweede probleem is dat er, ondanks de vele voorbeelden van technologisch denken, toch ook meer dan één casus Heideggers analyse tegenspreekt.

Wat is precies het eerste probleem? Wie het determinisme verdedigt, legt de lat hoog. De bewering is dat, gegeven een bepaalde beginsituatie, een bepaald effect *noodzakelijkerwijs* zal volgen. Het volstaat dan niet om aan te tonen dat er een oorzakelijk verband is, het moet ook aannemelijk worden gemaakt dat het niet anders kan dan dat het gevolg voortvloeit uit een bepaalde oorzaak. Dat geldt voor Heidegger, en uiteraard ook voor de varianten van technologisch determinisme die straks nog aan bod komen. Het probleem is echter dat in Heideggers *Die Frage nach der Technik* geenszins voldoende redenen worden gegeven om te besluiten dat het technologisch determinisme juist is. Natuurlijk geeft hij voorbeelden van hoe we op een instrumentele manier de dingen benaderen. Hij heeft het onder meer over het water van de Rijn om een waterkrachtstation aan te drijven, en meer algemeen, over de natuur die wordt gezien als een bron van energie. Maar die enkele voorbeelden volstaan niet om aan te nemen dat we *onvermijdelijk* in termen van nut denken. Mijn punt is dus niet – althans hier niet – dat het onjuist is wat Heidegger beweert, dat zijn versie van het technologisch determinisme niet klopt. De kritiek is dat zijn argumentatie voor zijn stelling tekortschiet. Om

de stelling te verdedigen die hij verdedigt had hij een stevigere onderbouwing moeten voorzien.

Een mogelijke reactie op die kritiek zou kunnen zijn dat je op zoek gaat naar wetenschappelijke evidentie. Hoewel Heidegger – die nochtans uitspraken doet over hoe wij vandaag naar de wereld kijken – zelf niet verwijst naar wetenschappelijke studies, zou je kunnen nagaan of uit onderzoek blijkt dat het instrumentele denken tot *alle* geledingen van de samenleving is doorgedrongen. Zulk onderzoek is praktisch gezien wellicht onhaalbaar, maar stel dat het toch wordt gedaan en dat het deze resultaten oplevert, zou dat Heideggers bespiegelingen ondersteunen?

Als iets noodzakelijkerwijs bestaat, dan is het per definitie in alle gevallen zo. Een vierkant heeft noodzakelijkerwijs vier hoeken, en dus is het in alle gevallen zo dat een vierkant vier hoeken heeft. Kun je één geval aanwijzen waarin het niet voorkomt, dan is het niet noodzakelijk. Het omgekeerde echter klopt niet: wanneer iets in alle contexten het geval is, dan kun je daaruit niet afleiden dat het ook noodzakelijk zo is. Stel, je ziet op een zaterdag dat het regent en ook de week erna constateer je op zaterdag dat het regent. Sterker nog, je ziet dat het op alle zaterdagen van het jaar regent. Daaruit kun je echter niet concluderen dat het noodzakelijkerwijs op zaterdag regent. Wanneer dus uit sociologisch of antropologisch onderzoek zou blijken dat het instrumentele denken zich over alle domeinen van de samenleving heeft verspreid, dan heb je weliswaar reden om te vermoeden dat het onvermijdelijk is maar je kunt het er niet met zekerheid uit afleiden. Zulke wetenschappelijke bevindingen zouden met andere woorden het gebrek aan voldoende ondersteuning van Heideggers stelling niet wegwerken. Het zou dan nog steeds aannemelijk moeten worden gemaakt dat het technologische denken *noodzakelijk* is, en niet louter iets dat in alle domeinen aanwezig is. Heidegger had op zulke kritiek kunnen anticiperen door gematigder te zijn en minder sterke beweringen te formuleren.

Er is daarnaast nog een tweede probleem. Niet alleen schiet Heideggers argumentatie tekort, ook is zijn centrale stelling niet juist. We zijn niet veroordeeld tot instrumentalisme, een andere zienswijze is mogelijk. Heideggers analyse is dus onvoldoende genuanceerd, spreekt in te grove termen, ziet de verschillen over het hoofd die er nochtans duidelijk zijn.

Dat de wereld ook vandaag nog steeds meer is dan een middel om onze doelen te realiseren, blijkt onder meer uit het feit dat mensen spelen, hobby's hebben, fundamenteel onderzoek verrichten of zorg dragen voor elkaar. Wanneer je zorgt draagt voor iemand wil dat zeggen dat je die persoon niet als een middel maar als een doel op zich beschouwt, of op z'n minst niet uitsluitend als een middel. Het kan zijn dat zorgen voor iemand je een goed gevoel geeft, maar dat is normaal gezien niet de reden voor de zorg. Het zorgen is doorgaans gestoeld op een niet-instrumentele blik die de ander ziet als iemand die op zich waarde heeft. Wie nu beweert dat er mensen zijn die anderen helpen omwille van het goed gevoel dat helpen hen geeft, heeft het wellicht bij het rechte eind, maar ondermijnt daarmee mijn kritiek op Heidegger niet. Er is meer dan één persoon – en wellicht geldt het voor de overgrote meerderheid – die zorgt omwille van de persoon zelf, en dat volstaat als argument tegen Heideggers determinisme. Of neem het koesteren van objecten, zoals het horloge van mijn overleden grootvader uit het eerste hoofdstuk. Dat voorwerp functioneert nog steeds naar behoren, maar toch absorbeert dat functioneren de waarde ervan niet. Het horloge heeft betekenis, niet alleen omdat het mij de juiste tijd toont, maar ook, en vooral, omdat dat het laatste geschenk was dat mijn grootvader mij heeft gegeven voor mijn verjaardag. Die waarde zal daar altijd aan blijven kleven, zelfs als mocht blijken dat het horloge volstrekt nutteloos is geworden.

Sommigen gaan weliswaar akkoord met die kritiek op Heidegger, maar menen dat het steeds meer gebruiken van technologie toch een bedreiging vormt voor een niet-instrumentele kijk. In zekere zin is dat begrijpelijk, want technologie is per definitie gemaakt met een doel voor ogen. Toch hoeft technologie geen bedreiging te zijn voor een

benadering die niet in termen van middelen en doelen oordeelt, beide kunnen ook naast elkaar bestaan. Sterker nog, sommige technologieën zijn precies de uitdrukking van een niet-instrumentele blik. Ik denk in de eerste plaats aan medische technologie, dingen die zijn gemaakt met het oog op zorg en die uitgaan van de idee dat mensen op zich waarde hebben. Jazeker, medische apparatuur is een middel om te zorgen voor de ander. Maar dat kadert in die context wel in een niet-instrumentale benadering van anderen. Voor de hand liggende voorbeelden zijn defibrillatoren, MRI-scanners, pacemakers, infuuspompen en stethoscopen, maar ook het AIsysteem dat ik in het begin van het boek vermeldde, ontworpen om bacteriële infectieziekten te bestrijden, past in dat kader.

Die voorbeelden nemen niet weg dat we vaak in termen van middelen en doelen denken, dat het niet onwaarschijnlijk is dat we dat meer doen dan vroeger en dat sommigen wellicht de *indruk* hebben dat het instrumentele denken woekert. En natuurlijk valt het niet uit te sluiten, althans in principe niet, dat er ooit een tijd komt waarin we enkel nog in technologische zin kunnen nadenken, al denk ik dat de kans nihil of erg klein is. Toch laten die voorbeelden zien dat het onjuist is dat de wereld enkel in technologische zin kan worden begrepen.

Wil dat nu zeggen dat het omgekeerde wel klopt? Moeten we daaruit afleiden dat het onmogelijk is om de wereld in *louter* instrumentele termen te verstaan? Nee. Dat blijkt onder meer uit hoe we met veel technologieën omgaan. Mijn laptop is gemaakt met allerlei doelen voor ogen, en ik begrijp en gebruik die alleen op die manier. Of om een actueel thema te nemen: onze data worden door techgiganten vaak alleen als een snelle weg naar geldgewin gezien. Dat die gegevens mij toebehoren en dat ik er controle over zou moeten hebben, wordt daarbij vaak uit het oog verloren. Tussen haken: dat laatste heeft natuurlijk te maken met economische belangen. Op een niet-instrumentele manier naar data kijken, bijvoorbeeld door persoonsgegevens aan het recht op privacy te koppelen, is voor *Big Tech* een rem op het streven naar winst.[86]

HET PROBLEEM MET INSTRUMENTALISERING

Alle vormen van technologisch determinisme zijn beschrijvend: ze geven weer hoe de wereld in elkaar steekt – dat is althans de opzet. Tot slot van mijn uiteenzetting over Heidegger wil ik erop wijzen dat hij daarover ook een oordeel velt. Dat de instrumentele imperatief zou woekeren, is onwenselijk; het zou beter zijn als dat denken niet zo zou overheersen.

De toon van Heideggers essay is dus overwegend pessimistisch. Op het moment dat hij die tekst schreef, was dat niet zo opmerkelijk. De werken van andere techniekfilosofen van het eerste uur, ik denk aan die van Karl Jaspers, waren ook vooral negatief, en verschillen in dat opzicht van die van veel techniekfilosofen na hen. Voor de *petite histoire*: dat Heideggers analyse vrij somber is, is ook niet verrassend gelet op zijn persoonlijke voorkeuren. Zo zou hij zich laatdunkend uitgelaten hebben over zijn buren omdat die veel televisie keken. En naar verluidt zou hij zich hebben gestoord aan een uitzending over de schilderkunst van Paul Klee, omdat de snelle camerabewegingen een diepgaande reflectie op de kunst van Klee zouden verhinderen.[87]

In een interview met *Der Spiegel* in 1966 zei hij dat enkel God ons kan redden. Toegepast op de thematiek die ons bezighoudt, zou je dat als volgt kunnen interpreteren. God kan de dominantie van een instrumentele blik in onze cultuur tegengaan. Hij, en bij uitbreiding religie of een meer religieus geïnspireerde kijk, kan de overheersing van het technologische denken doorbreken, met als gevolg dat mensen opnieuw voeling krijgen met de waarde die mensen en dingen op zich hebben, los van hun nut. Dat is althans het kritische potentieel dat Heidegger God en religie toedicht. Ik laat nu in het midden of die oplossing steek houdt. Het is in ieder geval wel zo dat die vermeende remedie veronderstelt dat de diagnose juist is. Wat Heidegger als oplossing naar voren schuift, gaat ervan uit dat het inderdaad zo is, ten eerste, dat we vasthangen aan een instrumentele blik, en ten tweede, dat de dominantie van zulke blik onwenselijk is. Daarnet heb ik al minstens een vraagteken geplaatst bij het eerste punt. De vraag

is nu of het tweede punt klopt. Ook hier denk ik dat Heidegger bijval geniet. Niet alleen menen nogal wat mensen met Heidegger dat we enkel nog instrumenteel kunnen denken, een niet gering aantal mensen, zo vermoed ik, vinden net als Heidegger dat het problematisch is dat we de dingen alleen maar als middelen kunnen zien. Heeft men daarin gelijk? Is zo'n negatieve visie overtuigend?[88]

Voor tal van zaken is het geen probleem om ze als een middel te zien. Alle technologieën zijn ontworpen om een doel te realiseren, en dus zou het absurd zijn te beweren dat het onwenselijk is om ze in instrumentele zin te begrijpen. Dat geldt ook voor artefacten die geen technologieën zijn. Kennis kan een doel op zich zijn, maar wetenschappelijke kennis als een middel zien om technologie te maken is vanzelfsprekend ook geen probleem, net zoals het volkomen onproblematisch is om naar muziek te luisteren als een middel om te ontspannen. Kunnen we hetzelfde ook zeggen over niet-artificiële zaken? Volgens Immanuel Kant moeten we iedere mens als een doel op zich beschouwen en mogen we niemand als een middel zien. Dat is op het eerste gezicht wel aannemelijk, maar toch niet geheel overtuigend – het staat in ieder geval op gespannen voet met de idee van werken in loondienst. Wanneer iemand wordt aangenomen om het bedrijf te reorganiseren, dan wordt die persoon wel geïnstrumentaliseerd, maar op zich is daar niets problematisch aan. En wat zou er verkeerd zijn aan het feit dat mensen planten zien als dingen die dienen om de woonkamer aantrekkelijk te maken?

Als je het instrumentele denken wilt bekritiseren, is het dus raadzaam om voorzichtig te zijn. Dat is geen kritiek op Heidegger, want hij beweert niet dat het op zich problematisch is om de dingen instrumenteel te benaderen. Wel meent hij dat het fout is om ze *uitsluitend* zo te begrijpen. En dat is vanzelfsprekend minstens ten dele terecht. Het is geoorloofd iemand aan te nemen om winst te maken, maar het is onaanvaardbaar om die persoon te zien als iemand die geen rechten heeft en die je geen loon verschuldigd bent. Terwijl het geen probleem is om katten als gezelschapsdieren te zien,

is het wel onverantwoord om ze te beschouwen als organismen die geen recht op eten en drinken hebben. Hetzelfde geldt ook voor niet-levende zaken, data bijvoorbeeld. Geanonimiseerde gegevens gebruiken om medische AI-systemen te trainen is niet per definitie afkeurenswaardig, integendeel. Je kunt moeilijk of zelfs onmogelijk bezwaren hebben tegen het voeden van AI met data die niet meer aan een persoon kunnen worden gekoppeld om zo beter kanker op te sporen. Maar het is onwenselijk wanneer medische informatie enkel wordt gezien als een middel om winst mee te maken door ze aan een datahandelaar te verkopen. Het probleem dat ik hier nu aankaart, is dus niet zozeer dat mijn foto op Facebook of Instagram zonder mijn toestemming wordt gebruikt door *Big Tech* (hoewel dat ook problematisch is), noch dat de context van de foto buiten beschouwing wordt gelaten. Ik wil erop wijzen dat het een probleem is dat de foto geen betekenis of gewicht heeft, tenzij als middel om het AI-systeem voor gezichtsherkenning te trainen of accurater te maken.

Niettemin rijst de vraag of Heideggers kritiek niet wat overtrokken is. Het mag dan wel onwenselijk zijn om in *bepaalde* gevallen enkel in instrumentele termen te denken, is dat ook in *alle* gevallen fout? Is een louter instrumentele blik per definitie onwenselijk?

Ik verwees er in de inleiding van het boek al naar. Om batterijen te maken is veel lithium nodig die men delft uit de mijnen in onder meer Nevada. Het probleem echter is dat dit delven niet zonder gevolgen is – het leidt onder meer tot grote putten gevuld met toxische zwarte modder. Daarnaast weten we natuurlijk ook dat steenkool verbranden om machines aan te drijven broeikasgassen uitstoot. Hoewel het effect van het gebruik van lithium en steenkool dus duidelijk onwenselijk is, is mijn vraag waarom het herleiden van beide zaken tot instrumenten een probleem zou zijn. Zeker, het is ontoelaatbaar om mensen en andere dieren uitsluitend in instrumentele zin te begrijpen, maar wat is het probleem als het bijvoorbeeld over dit soort van materiaal gaat, over lithium en steenkool? Je kunt dezelfde vraag ook stellen wanneer het over de zaken gaat die op basis van lithium en steenkool respectievelijk worden gemaakt en aangedreven:

technologieën. Waarom zouden we die als meer dan een middel moeten zien?

Het is niet duidelijk wat Heideggers antwoord daarop zou zijn. De reden is dat hij in zijn tekst geen argumenten geeft voor zijn bewering waarom een louter instrumentele kijk in alle gevallen fout is. We weten dus bijvoorbeeld niet waarom hij het afkeurt om naast organismen ook dingen als boormachines en wagens enkel als middelen te zien. Toch betekent dat niet dat er, los van Heidegger, geen argumenten zijn die Heideggers visie ondersteunen. Als we bijvoorbeeld toespitsen op technologie en AI dan worden door anderen in deze context doorgaans drie redenen naar voren geschoven. Volstaan die om te besluiten dat een uitsluitend instrumentele blik op technologie en AI altijd problematisch is?

Ten eerste wijzen sommigen erop dat technologieën altijd iemand toebehoren, dat ze altijd de eigendom van iemand zijn. Daardoor kun je ze niet louter als een middel zien; je moet ze ook met respect behandelen, je moet er voorzichtig mee omspringen. Die tegenwerping is weliswaar terecht wanneer het over het eigendom van anderen gaat, maar wat als het gaat over *mijn* bezittingen? Het tweede antwoord heeft enkel betrekking op AI. Het herleiden van zulke technologie tot instrumenten wordt afgewezen, omdat slimme technologie rechten zou hebben. Dit is niet de plaats om dat voorstel grondig te onderzoeken, maar stel dat het steek houdt. Dan volstaat ook dat niet. Niet alle technologie is immers slimme technologie. Vandaar opnieuw mijn vraag: waarom zou het fout zijn om domme dingen als mijn fiets of koffieapparaat als middelen te zien die een doel dienen, en louter als middelen? Een derde antwoord ten slotte luidt dat het niet goed is om technologieën uitsluitend als instrumenten te bekijken, omdat we daardoor de puur instrumentele houding ten aanzien van mensen versterken, of omdat het risico groot is dat dit zal gebeuren. Of voor wie vindt dat Heidegger ongelijk heeft en dat we vandaag mensen niet enkel als middelen zien: het is onwenselijk om technologie uitsluitend als middel te zien, omdat we daardoor andere mensen geen waarde op zich meer zullen toedichten, of dat de kans

groot is dat we dat niet meer zullen doen. Hier wordt dus niet beweerd dat een louter instrumentele kijk op technologie *op zich* problematisch is. Zulke kijk is volgens deze benadering enkel een probleem omdat het onwenselijke gevolgen heeft, in dit geval: de instrumentalisering van mensen, het versterken van interpersoonlijke instrumentalisering, of het niet geringe risico dat een van beide zaken gebeurt. Het is duidelijk dat dit op zich niet volstaat om een louter instrumentale blik op technologie te problematiseren. Om overtuigend te zijn is verdere ondersteuning nodig, ondersteuning die bij voorkeur uit empirisch onderzoek komt. Klopt het inderdaad dat een uitsluitend instrumentele kijk op de dingen wordt overgeheveld naar hoe we met mensen omgaan, of dat daar een groot gevaar voor bestaat? Heeft een bepaalde kijk op technologie een onwenselijk effect op onze kijk op mensen? Toegepast op een actueel debat: is het onwenselijk om een seksrobot louter als een middel voor de bevrediging van seksuele verlangens te zien, *omdat* het gevolg is dat andere mensen tot pure lustobjecten worden herleid, of omdat de kans groot is dat dit zal gebeuren?

Daarmee kan ik nu het deel over Heidegger, over de eerste versie van de determinismethese, afronden. Mijn besluit luidt als volgt. De bewering dat ons denken gedetermineerd is, klopt niet. Het is niet zo dat we enkel in technologische zin naar de wereld kunnen kijken. Maar zelfs als Heidegger gelijk zou hebben, is het problematisch dat hij niet uitlegt waarom dat onwenselijk zou zijn in *alle* gevallen, dus niet alleen als het gaat over mensen en niet-menselijke dieren, maar ook als het over pakweg steenkool of technologie gaat. De drie argumenten die door anderen in de context van technologie en AI worden gegeven volstaan in ieder geval niet om Heideggers bewering te ondersteunen. Het is bijvoorbeeld helemaal niet zeker dat de bezorgdheid gerechtvaardigd is dat een uitsluitend instrumentele kijk op technologie zou kunnen leiden tot een instrumentele omgang met mensen. Voor alle duidelijkheid: ik beweer niet dat er geen goede redenen bestaan voor een algemene afwijzing van een louter technologische blik, wel dat het niet de gegeven argumenten zijn.

Soorten technologisch determinisme

Ik beklemtoonde het eerder al: naast Heideggers versie zijn er nog drie andere varianten van de determinismethese die je vaak hoort onder technologen. Die gaan allemaal over een oorzakelijke relatie waar technologie deel van uitmaakt. Dat hebben ze uiteraard gemeen met de versie van Heidegger, want zoals ik al uitlegde, verwijst elke vorm van determinisme naar een oorzaak-gevolgrelatie. Toch verschillen ze alle drie ook van diens theorie. Het gaat in al deze gevallen niet over een technologische zienswijze, zoals bij Heidegger, maar over robots, laptops en telefoons – technologieën zoals ik die eerder heb omschreven: zaken die door mensen zijn gemaakt met een doel voor ogen, zaken die bovendien meestal materieel van aard zijn.

Een van de drie andere varianten van de determinismethese gaat over de vermeende onvermijdelijke sociale gevolgen van technologie. Die laat ik voorlopig nog even links liggen. Ik leg om te beginnen de twee andere interpretaties uit: de ene gaat over het ontstaan van technologie, de andere over de evolutie van technologieën op het moment dat een eerste versie van een technologie op de wereld is gezet. Ik schets nu eerst de contouren van die laatste twee versies.

ONTSTAAN EN EVOLUTIE

Leg je de focus op de interpretatie van de determinismethese over het ontstaan van technologie, dan is de stelling dat alle technologie wel moest worden uitgevonden. Technologie is in dat geval een noodzakelijk gevolg van iets, zonder dat wordt gespecifieerd wat de oorzaak precies is. Het kon niet anders dan dat een technologie werd ontwikkeld. Het lijkt er misschien wel op dat computers en sociale media ook niet hadden kunnen bestaan, bij nader inzien zijn ze het onvermijdelijke gevolg van pakweg de Tweede Wereldoorlog en het proces van individualisering dat zich sinds een aantal decennia sterk

heeft doorgezet. Het wiel en het internet, ze moesten op een bepaald moment wel worden gecreëerd.

Concentreer je je op de evolutie van technologieën, dan kan het gaan over de opeenvolging van verschillende technologieën of over de vernieuwing van een bestaande technologie. Duidelijk is dat het hier telkens over de onvermijdelijke band tussen twee technologieën gaat. Het ene ding leidt noodzakelijk tot het volgende ding; onder alle omstandigheden zou deze technologie tot de ontwikkeling van die andere technologie leiden. Een voorbeeld van deze vorm van technologisch determinisme is de volgende bewering. Het was niet te vermijden dat aan het eind van de achttiende eeuw door ingenieur Claude François Jouffroy d'Abbans de eerste stoomboot werd gebouwd – de zogeheten *Pyroscaphe* –, aangezien er reeds boten en stoommachines bestonden – boten werden al in de oudheid gebruikt en de eerste stoommachine werd in het begin van de achttiende eeuw door Thomas Newcomen uitgevonden. Die denktrant kenmerkt ook de documentaire 'The Machine That Changed the World' uit 1992 over de geschiedenis van de computer. De leidende idee van die documentaire is dat iedere stap in het ontwikkelingsproces tot een nieuwe doorbraak moest leiden.

Er zijn duidelijk verschillen tussen beide versies van de determinismethese, tussen die over het ontstaan en die over de evolutie van technologie. De ene versie gaat over één technologie, de andere over minstens twee technologieën. De ene versie stelt dat alle technologie noodzakelijk is, terwijl de andere stelt dat alle technologie onvermijdelijk tot andere technologie leidt. Die laatste bewering houdt in dat sommige technologie noodzakelijk is – de technologie namelijk die onvermijdelijk volgt uit een andere – maar niet dat *alle* technologie onvermijdelijk is, wat de determinismethese over het ontstaan van technologie wel beweert.

Toch wil ik ook de gelijkenis tussen beide vormen benadrukken. Zowel wanneer het gaat over het ontstaan van technologie als wanneer het over de evolutie ervan gaat, wordt beweerd dat technologie er in alle

gevallen zal komen. Dat houdt in dat ze geen resultaat is van overleg of compromis tussen allerhande groepen, dat ze geen effect van sociale processen, relaties of groepen is. Indien dat wel zo zou zijn, dan zou technologie er ook niet kunnen zijn, want relaties tussen mensen en groepen zijn niet in beton gegoten, net zoals een compromis dat niet is. Daarnaast betekent de onvermijdelijkheid van technologie ook dat overheden en andere stakeholders passief staan tegenover het ontstaan en de verdere ontwikkeling ervan. Dat kun je op de volgende cultuurhistorische manier duiden. Vanaf de verlichting werden verschillende zaken minstens voor een deel onderworpen aan sociale, democratische controle: ethiek, economie, politiek. Technologie echter is wellicht een van de laatste dingen die daaraan ontsnapt. God is dood, dat onderschrijven de meeste aanhangers van het technologisch determinisme. Maar dat betekent niet dat iedere transcendentie is verdwenen. Technologie heeft de plaats van God ingenomen, zo zou je de determinismethese kunnen samenvatten – althans een bepaalde invulling van die these. Was het eertijds God die zich aan de greep van het volk onttrok, nu is het de technologie die de menselijke controle overstijgt.

DON'T SHOOT THE MESSENGER

Straks zal ik de argumenten voor en tegen beide vormen van technologisch determinisme belichten. Ik wil nu echter al beklemtonen dat zulke reflectie praktische relevantie heeft. Als het namelijk waar is dat het ontstaan en de evolutie van technologie onvermijdelijk is, dan heeft dat gevolgen die niet onbelangrijk zijn. Voor overheden en andere stakeholders is dan geen rol weggelegd tijdens het ontwerp en maken van dingen. Als ze al invloed kunnen uitoefenen, dan is die voorbehouden aan de periode *na* de productiefase. Al wil dat uiteraard nog niet zeggen dat die invloed verwaarloosbaar is. Politici en groepen kunnen van belang zijn voor de verspreiding van technologie onder de bevolking en om de toegang van burgers tot technologie te controleren – vaak betekent het dat de toegang tot de technologie wordt verbreed, soms dat die wordt versmald, in het geval van

bijvoorbeeld militaire technologieën. Daarnaast kunnen geëngageerde burgers proberen om de al dan niet gunstige gevolgen van technologie te beïnvloeden. Als blijkt dat een technologie onwenselijke effecten heeft, dan kunnen overheden daarop anticiperen door ze trachten op te vangen; wenselijke effecten kunnen dan weer worden versterkt of extra belicht.

Er zijn daarnaast nog andere goede praktische redenen om beide vormen van determinisme van naderbij te bekijken. Die hebben te maken met het soort van interesse die mensen kunnen hebben voor technologisch determinisme. In de vorige hoofdstukken heb ik er al op gewezen dat je vanuit meerdere hoeken interesse kunt hebben voor beweringen. Je kunt in uitsluitend theoretische zin belangstelling hebben, maar er kunnen ook belangen spelen. Dat geldt ook voor de determinismethese die over het ontstaan en de evolutie van technologie gaat.

Aan de ene kant kan de belangstelling zuiver theoretisch zijn en ontwikkel je uitsluitend argumenten voor technologisch determinisme. Aan de andere kant kan het ook zijn dat sommigen, en het gaat dan in de eerste plaats om technologieontwikkelaars, het determinisme over het ontstaan en de evolutie van technologie verdedigen omdat ze zich willen verdedigen tegen kritiek op de door hen ontwikkelde technologie. Je kunt misschien wel gekant zijn tegen die ontwikkeling, zo verdedigt men zich, maar die kritiek ketst af op het feit dat die technologie er in alle gevallen toch zal komen. De verdediging van de determinismethese wordt hier gebruikt om negatieve commentaren te marginaliseren. Dat maakt het uiteraard relevant om ons dadelijk verder in het technologisch determinisme te verdiepen. Als blijkt dat het klopt, geef je daarmee munitie aan de techwereld; blijkt daarentegen dat het determinisme geen steek houdt, dan ontneem je technofielen en ontwikkelaars een mogelijk antwoord op de kritiek op hun ontwerpen.

Een ander mogelijk motief, opnieuw voornamelijk van ontwerpers, is van morele aard: men wil de eigen handen in onschuld wassen. Men

interesseert zich voor die theorie, omdat men ervan overtuigd is dat determinisme betekent dat je niet verantwoordelijk kunt worden gehouden voor de technologie. De redenering gaat dan als volgt: als ik een technologie ontwerp, een technologie die er sowieso zal komen, dan kan men mij niet moreel verantwoordelijk houden voor de negatieve gevolgen van het gebruik van de technologie. Straks kom ik daarop nog terug, maar veronderstel nu even dat het determinisme iemands morele verantwoordelijkheid inderdaad absorbeert. Het gevolg is dan dat in een deterministische wereld technologieontwikkelaars geen morele verantwoordelijkheid dragen en dat je hen dus niet kunt straffen voor onwenselijke effecten. Alleen al daarom is het relevant om straks uit te vissen of we inderdaad op het vlak van technologie in een deterministische wereld leven.

Om verwarring te voorkomen, wil ik nog wel benadrukken dat die menselijke motieven om zich voor het determinisme te interesseren geen reden zijn om die theorie aan te vallen of zelfs te verwerpen, maar wel om die te onderzoeken. Het is mogelijk dat het determinisme niet klopt, en het kan dat die theorie uit eigenbelang voortkomt, maar als de determinismethese fout is, dan is dat omdat er iets schort aan de theorie, en niet omdat de verdediging van die theorie voortvloeit uit bepaalde motieven. We moeten onze pijlen op de theorie richten, niet op de mens en de mogelijke belangen achter de theorie.

IN DE GREEP VAN EXPERTEN

Naast Heideggers versie van technologisch determinisme en de twee versies over de ontwikkelingsgeschiedenis van technologie is er nog een vierde, meteen ook de laatste die ik hier introduceer. Die luidt dat alle technologie onvermijdelijk sociale effecten heeft; die gevolgen zouden er in om het even welke context zijn. Mochten we een teletijdmachine hebben, dan zouden dezelfde sociale effecten opnieuw plaatsvinden. Ik schetste die interpretatie al in de inleiding van dit hoofdstuk.

Die bewering lijkt in zekere zin op de disruptiethese uit het vorige hoofdstuk, de stelling dat AI disruptieve technologie is. Beide focussen niet op technologie op zich, maar op de gevolgen van het gebruik van technologie. Een andere gelijkenis is dat het in de twee gevallen over technologie gaat, en dus niet over een manier van denken. Bovendien wordt nu ingezoomd op een oorzakelijk verband waarin technologie zelf geen effect is, maar een oorzaak van iets anders is. Toch zijn er ook duidelijk verschillen tussen de disruptiethese en de stelling over de sociale effecten. Die verschillen hebben te maken met het effect van technologie. De disruptiethese zoals ik die heb geïnterpreteerd, verwijst naar de gevolgen van AI op het vlak van de ethiek; de oorspronkelijke invulling heeft een economisch karakter en gaat over het effect van AI op de markt. Maar terwijl de disruptiethese in principe over gevolgen op *alle* vlakken kan gaan – over economische en morele maar ook sociale gevolgen –, ligt de focus in de context van het determinisme *uitsluitend* op sociale effecten, de gevolgen van technologie op sociale relaties, processen en rollen. Verder is het voor de determinismethese niet van belang of het sociale effect een disruptief karakter heeft of niet.

Doorgaans wordt die interpretatie niet apart vermeld of besproken. Men koppelt ze meestal aan twee andere betekenissen van de determinismethese, die over het ontstaan en de evolutie van technologie. Het kan niet anders dan dat technologie ontstaat en dat uit de ene de andere technologie voortvloeit, en dus kan het niet anders dan dat technologie sociale invloed uitoefent, zo luidt dan de stelling. Als dat klopt, dan wil dat zeggen, ten eerste, dat groepen van mensen, sociale relaties, rollen en processen in de samenleving worden beïnvloed door de beslissingen van een handvol experts op het vlak van technologie: ontwerpers, computerwetenschappers en AI-ontwikkelaars die streven naar steeds efficiëntere en rendabele technologie en die zich daarvoor beroepen op de exacte wetenschappen en ingenieurswetenschappen. Ten tweede wil het ook zeggen dat het leven van nogal wat mensen wordt beïnvloed door technologieën waar zijzelf tegelijkertijd geen enkele controle over

hebben. Als de combinatie van de drie vormen van determinisme inderdaad klopt, dan behoeft het geen verdere uitleg dat dit een onwenselijke situatie is. Maar klopt dat inderdaad?

In zekere zin is het begrijpelijk dat het determinisme over sociale effecten enerzijds en het determinisme over het ontstaan en de evolutie anderzijds met elkaar worden verbonden. De idee van sociale invloed lijkt naadloos te volgen uit een van beide andere vormen van determinisme. Mocht blijken dat de ontwikkeling van de ene technologie noodzakelijkerwijs tot andere technologie leidt, dan is het op het eerste gezicht moeilijk om zich die ontwikkeling zonder sociale effecten voor te stellen, of zou het op z'n minst niet verrassend zijn dat de onstuitbare voortgang van technologie ook sociale impact heeft. Toch zijn die verschillende interpretaties niet onlosmakelijk met elkaar verbonden. Ik kan me zonder veel moeite voorstellen dat technologieën elkaar onvermijdelijk opvolgen zonder dat die innovaties op sociaal vlak invloed hebben. En omgekeerd is het niet al te moeilijk om me een wereld voor te stellen waarin het ontstaan of de evolutie van technologie noodzakelijk sociale effecten heeft, terwijl dat ontstaan of die evolutie geen onvermijdelijk, noodzakelijk karakter heeft.

Die interpretaties staan dus los van elkaar, althans in theorie, en moeten bijgevolg afzonderlijk worden besproken. Ik leg daar de nadruk op omdat wie die interpretaties koppelt doorgaans enkel argumenteert voor de interpretatie over het ontstaan en de evolutie van technologie. Men lijkt ervan uit te gaan dat de redenen voor de interpretatie van de determinismethese over de ontwikkelingsgeschiedenis ook opgaan voor de interpretatie van sociale effecten, maar dat is dus onterecht. Vandaar mijn vraag: klopt het dat technologieën effecten hebben die zowel sociaal als onvermijdelijk zijn? Dat is de vraag waarop ik mij nu eerst focus. Op de volgende bladzijden evalueer ik dus de versie van de determinisme over sociale effecten. Het is pas een flink eind verderop in dit hoofdstuk dat ik de argumenten belicht voor en tegen de versies van

de determinismethese die te maken hebben met het ontstaan en de evolutie van technologie.

De sociale effecten van technologie

Omdat door de bank genomen weinig aandacht wordt besteed aan het determinisme met betrekking tot sociale effecten is het belangrijk om een aantal mogelijke maar vaak voorkomende redeneerfouten onder de aandacht te brengen. Ik doe dat aan de hand van de volgende twee zaken: enerzijds sociale media en de algoritmen die de fora van die media sturen, en anderzijds de polarisering van de samenleving. Dat laatste karakteriseert verschillende landen. We weten dat bijvoorbeeld in de Verenigde Staten de samenleving meer en meer uiteenvalt in Republikeinen en Democraten. De eerste groep pleit tegen migratie en is voor wapenbezit, de tweede groep is pro migratie en gekant tegen het bezit van wapens. Terzijde: naar verluidt zouden Levi's jeans meer door Democraten dan door Republikeinen worden gedragen.

ZIJN SOCIALE MEDIA POLARISEREND?

Vaak wordt beweerd dat beide oorzakelijk verbonden zijn: sociale media werken polarisering in de hand. De redenering gaat dan soms als volgt: sinds het eerste decennium van deze eeuw bestaan sociale media en daarna nam de polarisering toe, en dus moet het besluit wel zijn dat de polarisering een gevolg is van de sociale media. Laten we even aannemen dat de chronologie klopt: eerst sociale media, dan polarisering. Dat rechtvaardigt echter niet de conclusie dat sociale media bijdragen tot polarisering. Het is niet omdat twee gebeurtenissen elkaar in de tijd opvolgen, dat de gebeurtenis die eerst kwam de oorzaak is van de tweede. Wanneer wielrenner Remco Evenepoel een kruisteken maakt vooraleer hij een wedstrijd rijdt en hij nadien de wedstrijd wint, wil dat niet zeggen dat hij won *omdat* hij een kruisteken maakte.

Anderen argumenteren dan weer op deze manier. Het gebruik van sociale media gaat gepaard met de toename van de polarisering, en dus is technologie de oorzaak van die polarisering. Ook die conclusie is niet terecht, zelfs als het klopt dat er een samenhang is. Hier speelt de verwarring tussen een correlationeel verband en een oorzakelijk verband. Dat twee fenomenen samenhangen met elkaar houdt niet in dat er een oorzaak-gevolgrelatie is. Het is bijvoorbeeld zo dat de stijging van de misdaadcijfers samenhangt met de toename van de verkoop van ijsjes, maar dat betekent niet dat er meer misdaden zijn *omdat* er meer ijsjes worden verkocht, of omgekeerd. De stijging van beide wordt veroorzaakt door een derde factor: de toename van temperatuur in de zomer.

Voor alle duidelijkheid: ik beweer hier niet dat sociale media polarisering niet effectief in de hand werken. Het is mogelijk dat dit wel het geval is. Ik beweer alleen dat je dat niet kunt concluderen op basis van de volgende gegevens: dat polarisering in de tijd volgt op het gebruik van sociale media en dat er een correlatie bestaat tussen beide fenomenen.

Niet alleen de determinist moet oppassen voor slordige redeneringen, ook wie het aanvalt moet waakzaam zijn voor overhaaste conclusies. Dat blijkt bijvoorbeeld hieruit. Wanneer wordt nagedacht over de invloed van sociale media op polarisering wordt dikwijls een beroep gedaan op wetenschappelijke studies als de paper ‘The Editor vs. The Algorithm’ uit 2019 en het werk van econoom Ro’ee Levy.[89] Die studies ondersteunen de idee van sociale media als echokamer. Algoritmen leiden ertoe dat de diversiteit van artikelen in de newsfeed op bijvoorbeeld Facebook afneemt en dat mensen na verloop van tijd steeds minder tijd spenderen aan het lezen van verschillende soorten artikelen. Verder is het ook zo dat de algoritmen ons vaak artikelen aanbieden die in de eerste plaats onze overtuigingen bevestigen, eerder dan dat ze die uitdagen. Kortom, er is evidentie voor de bewering dat sociale media ideologische bubbels creëren.

Aan de andere kant wijzen wetenschappers er ook op dat er geen bewijs is voor de stelling dat algoritmen en sociale media bijdragen aan de polarisering van de samenleving. Ze mogen dan wel echokamers creëren, dat AI-systemen de spanning en tegenstelling tussen groepen doet toenemen werd tot op heden niet aangetoond.[90] Sommigen trekken op basis daarvan de conclusie dat er geen oorzakelijk verband is tussen AI en polarisering. Ook dat is niet gerechtvaardigd. Als je geen bewijs hebt dat iets bestaat, is dat nog geen bewijs dat het niet bestaat.[91] Dus zelfs al hebben we geen oorzakelijke band kunnen vaststellen, het is nog steeds mogelijk dat die band er wel degelijk is, maar dat wij die eenvoudigweg nog niet hebben achterhaald.

Verder kun je de versie van technologisch determinisme over sociale effecten ook niet aanvallen door je te beroepen op een studie die polarisering voldoende verklaart zonder daarbij te moeten verwijzen naar sociale media als oorzaak. Wie determinist is als het gaat over sociale media en polarisering beweert immers enkel dat sociale media noodzakelijkerwijs leiden tot polarisering, en niet het omgekeerde, namelijk dat alle polarisering noodzakelijkerwijs voortvloeit uit sociale media. In het verlengde daarvan: mocht uit wetenschappelijk onderzoek blijken dat polarisering door sociale media ontstaat, maar niet enkel en alleen door sociale media, dan is dat nog geen argument tegen het technologisch determinisme. Wie vindt van wel, vertrekt vanuit een fout begrip van dat determinisme.

De determinist oppert dat technologie onvermijdelijk sociale effecten heeft, niet dat die effecten *uitsluitend* door technologie ontstaan.

BEL ME, SCHRIJF ME

De voorbije paragrafen zijn enkel een pleidooi voor voorzichtigheid wanneer wordt nagedacht over de sociale effecten van technologie. Ze laten niet zien dat technologie geen sociale effecten heeft. Het zou ook minstens verrassend zijn mocht ik dat laatste verdedigen. Er zijn immers talloze voorbeelden van hoe technologie sociale processen,

rollen en relaties beïnvloedt. Die invloed kan wenselijk, onwenselijk of neutraal zijn. Door de microgolfoven eten gezinsleden minder vaak op hetzelfde moment aan dezelfde tafel, auto's onderstrepen sociale status, en door sociale media kunnen we makkelijker oude vriendschappen hernieuwen of nieuwe contacten leggen.

Een ander voorbeeld is het juridisch en politioneel gebruik van AI. In het eerste hoofdstuk introduceerde ik de term 'algoracisme' om de aandacht te vestigen op discriminatie door AI op basis van huidskleur en afkomst. In diezelfde context verwees ik ook naar het COMPASalgoritme, dat in de Verenigde Staten is gebruikt om de kans op recidive in te schatten. In 2016 bleek dat het algoritme witte mensen en mensen van kleur ongelijk behandelde: de eerste groep werd sneller vrijgelaten dan de tweede groep. Ook het gebruik van AI door politiediensten is niet zonder sociale gevolgen. Wanneer zulke technologie wordt gebruikt om te voorspellen waar wellicht delicten plaats zullen vinden en de politie vervolgens naar een hotspot uitrukt om te patrouilleren kun je in de Verenigde Staten maar beter iemand met een witte huidskleur zijn. Volgens cijfers van het Amerikaanse ministerie van Justitie is de kans dat je wordt aangehouden als je van kleur bent meer dan twee keer zo groot dan wanneer je wit bent. Mensen met een donkere huidskleur hebben vijf keer meer kans om zonder gegronde reden te worden aangehouden dan mensen met een witte huidskleur. Deze zaken illustreren de ongemakkelijke waarheid dat technologie niet enkel een mannelijke maar ook een witte hegemonie kan onderstrepen of zelfs versterken.[92]

Nog een voorbeeld: de telefoon. Omdat hij voortvloeit uit de telegraaf werd de telefoon kort na het einde van de negentiende eeuw alleen gebruikt voor korte zakelijke boodschappen. Dat veranderde in het begin van de vorige eeuw. Mede door de verlaging van de beltarieven raakte de telefoon steeds meer gewild, waarna hij tijdens het interbellum ook als een middel voor sociale communicatie werd gebruikt. Die verschuiving heeft duidelijk sociale invloed gehad. Een eerste bevinding is dat niet alleen het aantal kennissen maar ook het aantal gesprekken met eenzelfde persoon toenam. Verder zou de

telefoon ook invloed hebben gehad op sociale praktijken. Sommige praktijken, zoals korte onaangekondigde bezoeken, namen af; rituelen als de aankondiging van een bezoek werden dan weer door de telefoon geïntroduceerd. Tot slot zou ook het aantal fysieke ontmoetingen zijn toegenomen, wat wellicht komt doordat je met de telefoon sneller afspraken kunt maken.[93] Het pareert in ieder geval ook de populaire kritiek dat nieuwe communicatietechnologieën tot minder fysiek sociaal contact leiden.

Ook nu is het zaak om geen snelle, maar wel juiste conclusies te trekken. De betekenis van de determinismethese die ik nu aan het bespreken ben, gaat immers over sociale gevolgen die onvermijdelijk zijn. De casussen die ik zopas aanhaalde, wezen enkel op het sociale karakter van de effecten van technologie en zeggen niets over het al dan niet onvermijdelijke karakter van die gevolgen. Je mag daarom uit de gegeven voorbeelden niet concluderen dat de sociale effecten er ook niet hadden kunnen zijn, net zoals de conclusie niet gerechtvaardigd is dat de gevolgen noodzakelijkerwijs voortvloeien uit de technologie.

NOODZAKELIJK? MOGELIJK!

Stel nu dat alle technologieën sociale impact hebben. Zouden die gevolgen er zijn in om het even welke samenleving en in om het even welk tijdsgewricht? Volgen de effecten noodzakelijkerwijs? Is het voldoende dat technologie er is opdat de sociale effecten volgen? Hebben utopisten en alarmisten gelijk wanneer ze zeggen dat een technologie op sociaal vlak onvermijdelijk respectievelijk goede en slechte gevolgen heeft?

Het gevaar bestaat dat je daar positief op antwoordt *omdat* je de juiste voorspelling deed dat de sociale effecten zouden volgen. Ik heb voorspeld dat de gevolgen er zouden zijn, en aangezien de effecten zijn opgetreden, is de impact onvermijdelijk, zo zou je kunnen redeneren. Hoewel het begrijpelijk is dat die conclusie wordt getrokken, is ze niet terecht. Wanneer iets gedetermineerd is – lees:

wanneer een effect noodzakelijkerwijs volgt – dan kun je op voorhand voorspellen dat het effect zal volgen. Het omgekeerde is niet waar. Mijn vrouw voorspelt dat ik morgen angstig zal zijn wanneer ik in de wachtkamer van de tandarts zit. Wanneer dat de dag erna inderdaad zo is, wil dat niet zeggen dat dat per se zo is. Zij had daar wel goede redenen voor – ik ben immers doorgaans een beetje een angsthaas als het op de tandarts aankomt – maar het is ook mogelijk dat het nu anders was, bijvoorbeeld omdat ik buiten haar weten om bij een gedragstherapeut een aantal sessies heb gevolgd om mijn angst onder controle te houden.

Verder is het ook zo dat de effecten van technologie op andere dan de sociale vlakken niet per se onvermijdelijk zijn. Neem een auto met een benzinemotor. Gebruik je zulke technologie, dan worden er wel broeikasgassen uitgestoten maar ontstaat niet per definitie smog. Wanneer er smog is, is dat enkel wanneer er op een kleine oppervlakte veel benzineauto's rijden. De smog is dus wel een gevolg van de auto, maar geen onvermijdelijk gevolg.[94] Welnu, niet alleen ecologische maar ook sociale effecten volgen niet in om het even welke omstandigheden. Dat is wat opnieuw de telefoon leert. Die heeft enkel kunnen leiden tot de verbreding van de interpersoonlijke contacten omdat die na verloop van tijd betaalbaar is geworden en omdat mensen op de hoogte werden gebracht van het bestaan van de telefoon. Daarnaast werd het succes ook pas mogelijk omdat het aanvankelijk negatieve beeld van de telefoon verdween, omdat de productie en distributie toenam, omdat er veel schakelcentrales werden gebouwd, en omdat het aantal kabelverbindingen toenam. In een andere dan de huidige wereld – zonder goedkope abonnementen, zonder kabelverbindingen, met weinig distributiemogelijkheden – zou de telefoon niet die sociale effecten hebben.

Als technologieën sociale gevolgen hebben, zijn die dus niet per se noodzakelijk. Het volstaat niet altijd dat de technologie bestaat, opdat de sociale effecten zullen volgen. Er moeten eerst aan een aantal voorwaarden zijn voldaan vooraleer die effecten zich kunnen

voordoen, voorwaarden die van velerlei aard kunnen zijn: psychologisch, sociaal, economisch, politiek, materieel, enzovoort. Dat geldt niet alleen voor de telefoon, maar ook voor tal van andere technologieën, misschien zelfs alle. AI-systemen bijvoorbeeld hebben naast economische ook sociale effecten, maar die effecten kunnen er alleen zijn wanneer aan bepaalde voorwaarden is voldaan. Er zijn geen sociale gevolgen van slimme systemen als de scheepsindustrie in zeecontainers niet wereldwijd computers en mobiele telefoons verspreidt, als de tarieven voor gebruikers te hoog zijn, als de mobiele netwerken niet functioneren, als er onvoldoende technologische knowhow is bij gebruikers, enzovoort. Ook hier moet je dus zeggen: in een andere wereld of een ander tijdsgewricht – zonder bijvoorbeeld internationaal transport – zou AI op sociaal vlak niet de effecten hebben die het nu heeft.

Laten we echter aannemen dat mijn redenering niet klopt: als technologie sociale effecten heeft, volgen die effecten toch wel in alle omstandigheden, los van plaats en tijd. Mogen we dan besluiten dat de determinismethese juist is, deze interpretatie althans? Dat valt te betwijfelen. Alle technologieën zijn per definitie ontworpen met een doel voor ogen en hebben dus ook effect als ze doen waarvoor ze zijn gemaakt. Toch is het verre van zeker dat alle technologieën een invloed hebben op sociaal vlak, laat staan dat die noodzakelijk zou zijn. Denk aan de boormachine uit het eerste hoofdstuk of de elektrische tandenborstel uit het vorige. Het is bijzonder moeilijk of zelfs bijna onmogelijk om effecten op sociale relaties, processen of instellingen te bedenken die zouden volgen uit het gebruik van die twee technologieën. Het is niet uitgesloten dat ook zulke basale dingen sociale impact hebben, maar het zou niet verbazen mocht blijken dat die impact er niet is.

Ik rond af. Niet alleen Heideggers versie van het technologisch determinisme klopt niet, ook de determinismethese over sociale effecten kampt met meerdere problemen. Hoewel het niet zeker is, maar wel erg waarschijnlijk, dat niet elke technologie sociale gevolgen heeft, is het wel duidelijk dat niet alle sociale gevolgen noodzakelijk

zijn. Een minder sterke maar vooral ook juistere stelling is dat alle technologieën *mogelijk* zulke gevolgen hebben. Die conclusie laat ruimte voor de uitspraak dat het erg waarschijnlijk is, maar niet noodzakelijk, dat bijvoorbeeld de algoritmen van Instagram sociale gevolgen hebben en dat die gevolgen waarschijnlijker zijn dan de gevolgen van pakweg een elektrische tandenborstel. En verder laat het ook toe dat *sommige* technologieën onvermijdelijk impact hebben op sociaal vlak. Mocht in de toekomst of nu blijken dat enkele technologieën inderdaad sowieso sociale impact hebben, dan spreekt dat mijn conclusie niet tegen. Maar zoals ik eerder al beklemtoonde: om dat aannemelijk te maken volstaat het niet om te wijzen op sociale effecten, het moet in dat geval ook duidelijk zijn dat die effecten onvermijdelijk zijn.

Het ontstaan van technologie

Het is tijd om even terug te blikken. Dit hoofdstuk gaat over een populaire stelling die net als de neutraliteitsthese en disruptiethese vaak wordt verdedigd door politici, filosofen, computerwetenschappers en ingenieurs, namelijk de determinismethese. Er bestaan daar zeker vier versies van en intussen weten we dat de versie van Heidegger en die over sociale effecten niet juist zijn. Maar er zijn nog twee andere versies, die ik wel al uitlegde maar die ik voorlopig nog niet evalueerde. De ene gaat over het ontstaan van technologie, de andere over de evolutie van technologie eens ze op de markt is gezet. Ik heb er eerder op gewezen dat de versie van het determinisme over de onvermijdelijke sociale effecten vaak in één adem wordt genoemd met de twee versies van het technologisch determinisme over het ontstaan en de evolutie van technologie. Het is echter belangrijk om die uit elkaar te houden. Als het niet klopt dat technologie onvermijdelijk sociale gevolgen heeft, wil dat niet per se zeggen dat ook de twee andere versies fout zijn. We doen er dus goed aan ook die andere varianten van naderbij te bekijken.

Vooraleer ik dat doe, wil ik eerst onder de aandacht brengen dat het denken over het ontstaan en de evolutie van technologie vaak samenvalt met een poging om technologie op dezelfde manier als de natuur te begrijpen. Het onderzoek hiernaar ontstond in de tweede helft van de negentiende eeuw en is geïnspireerd door de evolutietheorie die toen haar eerste successen boekte. In zijn essay 'Darwin Among the Machines' uit 1863 bijvoorbeeld oppert schrijver Samuel Butler dat het ontstaan en de evolutie van organismen en machines volgens dezelfde darwinistische principes moeten worden opgevat. Net zoals organismen komen en gaan, komen en gaan machines ook, en dat heeft te maken met een betere of slechtere aanpassing aan de omgeving dan de concurrenten, aldus Butler. Deze gedachtegang wordt nog steeds verdedigd, bijvoorbeeld in het al vermelde werk *What Technology Wants* van Kelly.

Een naturalistische blik op technologie gaat regelrecht in tegen een visie die in de achttiende eeuw populair was onder filosofen en wetenschappers. Binnen zo'n kader werd de natuur in een technologisch vocabularium beschreven – denk aan de metafoor van God als horlogemaker en de kosmos als een uurwerk. Maar wie het kijken naar technologie als naar de natuur zeer eigenaardig of zelfs absurd vindt, moet weten dat een aantal technologieën zijn gemaakt naar het model van dingen uit de natuur. Een eenvoudig voorbeeld is prikkeldraad – ik breng in herinnering dat ik 'technologie' breed opvat. Prikkeldraad werd aan het eind van de negentiende eeuw in Illinois in de Verenigde Staten uitgevonden naar het voorbeeld van de Osagedoorn: een boom met takken die sterke doornen heeft.[95] Een meer hedendaags voorbeeld is CRISPR/ Cas9, de technologie waarmee kan worden ingegrepen in het DNA van mensen en die in 2015 door *Science* is uitgeroepen tot de doorbraak van het jaar. Het is gebaseerd op het feit dat bacteriën moleculen zoals het befaamde Cas9 gebruiken om zich te verdedigen tegen virussen.[96]

TWEE KEER HET WIEL

Ik begin met de interpretatie van de determinismethese over het ontstaan van technologie, daarna focus ik op de evolutie van technologie, op de invulling over de opeenvolging en vernieuwing van technologieën. Alle technologie moest wel worden uitgevonden – daarover gaat het nu. Gezien de omstandigheden kon het niet anders dan dat een technologie ontstond. Het feit dat een technologie tot stand komt, is noodzakelijkerwijs zo. Onderhandelingen en discussies tussen groepen van mensen kunnen daaraan niets veranderen. Dat is de stelling van de determinist over het ontstaan van technologie. Je hoort ze dikwijls, zeker onder mensen die nauw zijn betrokken bij het maken van technologie. Maar klopt ze ook?

Om die opvatting aan te vallen, wordt soms verwezen naar technologieën die, zo wordt het argument dan geformuleerd, ‘niet uit noodzaak zijn ontstaan’. Het wiel is daarvan een voorbeeld, althans zoals het in Mexico en Midden-Amerika na de vierde eeuw na Christus werd uitgevonden. Dat was in de periode voorafgaand aan de kolonisering door de Spanjaarden, toen men nog niet wist dat het wiel in Europa en Azië al veel eerder was uitgevonden, in het vierde millennium voor het begin van onze jaartelling. Het punt is nu dat de reden voor de uitvinding verschilt. In Europa en Azië vond men het wiel uit om zware en grote voorwerpen te kunnen verplaatsen die door mensen of niet-menselijke dieren niet of moeilijk konden worden gedragen. Aan de andere kant van de Atlantische Oceaan was dat niet zo. Daar werd het uitgevonden voor miniatuurvoorwerpen die ofwel als offermateriaal ofwel als speelgoed dienden, en niet als reactie op een behoefte. In Amerika vloeide de ontwikkeling van het wiel dus niet voort uit noodzaak.[97] Dat geldt ook voor het ontstaan van een bepaalde type vierwieler aan het eind van de negentiende eeuw: de auto. Die ontstond niet omdat daar behoefte aan was, bijvoorbeeld omdat er geen of onvoldoende paarden meer waren. Het was eerder omgekeerd: de behoefte aan auto’s ontstond na de uitvinding van de auto.

Critici beweren dat die voorbeelden een probleem zijn voor het determinisme, althans voor de interpretatie waarop ik me hier focus. Die theorie gaat immers over het noodzakelijke karakter van het ontstaan van technologie, terwijl uit de voorbeelden net blijkt dat de technologieën niet altijd worden gemaakt uit noodzaak. Hoewel ik ook meen dat dit type van determinisme onvoldoende gegrond is – dadelijk meer daarover –, mist deze aanval haar doel. Dat komt omdat 'noodzaak' telkens anders wordt begrepen. Zowel de kritiek op het determinisme als het determinisme zelf zijn geformuleerd in termen van noodzaak, maar dezelfde term wordt in het geval van de kritiek anders ingevuld dan in het geval van het determinisme. Terwijl 'noodzaak' voor de determinist 'onvermijdelijk' betekent, verwijst de kritiek met 'noodzaak' naar een behoefte, het feit dat men iets nodig heeft, dat men nood aan iets heeft. De kritiek mist met andere woorden haar doel. Dat zou niet zo zijn als de determinist zou zeggen dat technologie voortvloeit uit noodzaak of uit een behoefte, maar dat is niet wat de determinist beweert.

GELIJKTIJDIGE EVOLUTIE

Wil dat nu zeggen dat deze versie van determinisme wél stand houdt? Is het onvermijdelijk dat een technologie op de wereld wordt gezet? Dat hangt af van de kracht van het argument voor die theorie. De stelling dat alle technologie wel moest ontstaan is gebaseerd op de idee van gelijktijdige evolutie – het is ook het enige argument. Het betekent dat iets op verschillende plaatsen ontstaat, terwijl daartussen geen verband bestaat. Ik stel voor om dat wat van naderbij te bekijken.

Het fenomeen komt voor in de organische wereld. Antivriesstoffen, die ervoor zorgen dat het bloed blijft stromen ook wanneer het erg koud is, werden niet één maar twee keer ontwikkeld: eenmaal door vissen aan de Zuidpool, eenmaal door vissen in de Noordelijke IJszee. Enkele andere voorbeelden: navigatie door weerkaatsing van het geluid wordt niet alleen door vleermuizen gebruikt, maar ook door dolfijnen en twee soorten vogels: de Zuid-Amerikaanse vetvogel en

de Aziatische gierzwaluw; drijvende zwemblazen vind je bij weekdieren en kwallen; en naast kikkers hebben ook kameleons tongen ontwikkeld om vanaf een afstand hun prooi te strikken. Het simultaan ontstaan van een eigenschap komt ook voor in de plantenwereld. In Noord-Amerika hebben op verschillende plaatsen paddenstoelgeslachten los van elkaar zwammen ontwikkeld, en zeven soorten planten werden onafhankelijk van elkaar insecteneters om voldoende stikstof te hebben.[98]

Gelijktijdige evolutie is een fenomeen dat ook buiten de natuur voorkomt, bijvoorbeeld in het domein van de wetenschap. In 1979 ontdekte men een gen dat codeert voor het zogeheten eiwit p53. Dat was een erg belangrijke vondst in de strijd tegen kanker, omdat het gen een negatieve invloed heeft op celdeling en dus het ontstaan van tumoren verhindert. Die ontdekking werd gedaan op verschillende plaatsen en onafhankelijk van elkaar: in onder meer Londen, New York en Parijs. Daarnaast, en dat is belangrijk, is er ook gelijktijdige evolutie in de techwereld. Het is met andere woorden vaak onterecht wanneer boeken of websites over de geschiedenis van technologie een uitvinding aan slechts één persoon linken.De blaaspijp werd twee keer uitgevonden: één keer in Azië, één keer in Amerika. Het klopt dat Thomas Edison aan het eind van de negentiende eeuw in de Verenigde Staten de gloeilamp uitvond, maar het is ook waar dat de lamp in diezelfde periode ook werd uitgevonden in Engeland en Rusland, door respectievelijk Joseph Swan en Alexander Lodygin. Op 14 februari 1876 vroegen zowel Graham Bell als Elisha Gray een patent aan voor de telefoon die ze los van elkaar hadden gemaakt. En hoewel we vooral Bell met de telefoon associëren, diende Gray haar aanvraag zelfs drie uur eerder in. Sterker nog, in 1860 had Antonio Meucci in Italië al een patent verworven op de telefoon, maar onder meer omdat hij het Engels niet machtig was werd zijn patent niet verlengd in 1874, twee jaar voor Bell en Gray daar wel in slaagden. Vaak worden aan het ontstaan van de fotografie uitsluitend de naam Louis Daguerre en Frankrijk verbonden. Ook dat is onterecht, want in Engeland ontdekte William Fox Talbot in diezelfde periode en los van

Daguerre dezelfde principes die aan de basis van de fotografie lagen. De inkjetprinter werd twee keer in dezelfde periode uitgevonden: zowel in de labs van Canon in Japan als door technologiebedrijf HewlettPackard in de Verenigde Staten. Een laatste voorbeeld: de transistor werd kort na de Tweede Wereldoorlog uitgevonden zowel in de Verenigde Staten in de Bell Labs van telefoonbedrijf AT&T als in Parijs door twee Duitse fysici.[99]

TOEVAL EN NOODZAAK

Die bevindingen over de geschiedenis van technologie zijn voor sommigen niet enkel leuke weetjes. Men haalt ze ook aan als een argument voor het determinisme over het ontstaan van technologie. Nu blijkt dat technologieën los van elkaar in diezelfde periode zijn ontstaan, zo redeneert men, kunnen we besluiten dat het niet anders kan dan dat die technologieën zijn ontstaan. Er zit noodzaak in de ontwikkeling van de gloeilamp, telefoon, fotografie, inkjetprinter, transistor en andere technologieën. Rechtvaardigt het argument de conclusie?

Het eerste probleem is dat niet alles op meerdere plaatsen tegelijk en los van elkaar ontstaat. In de dierenwereld is de bombardeerkever uniek: het is het enige organisme dat chemicaliën kan combineren tot een giftige straal die op vijandige dieren kan worden gespoten. (Bombardeerkevers laten ook winden van maar liefst 100 graden Celsius.) Maar ook in de wereld van de artefacten is niet elk ontwerp een voorbeeld van gelijktijdige evolutie. Zeker, veel technologieën ontstaan op hetzelfde ogenblik en onafhankelijk van elkaar, maar dat geldt niet voor *alle* technologieën. De computer, het vliegtuig en de personenauto: die technologieën werden niet tegelijkertijd op verschillende plaatsen uitgevonden. Indien het dus zou kloppen dat gelijktijdige evolutie aantoont dat technologie niet anders kon dan ontstaan, dan kun je het noodzakelijke ontstaan van technologie niet veralgemenen, althans niet op basis van gelijktijdige evolutie.

Daarnaast is deze vraag cruciaal: toont gelijktijdige evolutie aan dat iets noodzakelijk is? Het antwoord daarop is ontkennend. Dat gelijke dingen parallel naast elkaar ontstaan, kan toeval zijn; ik kan me een wereld voorstellen waarin het toeval is dat dingen tegelijk maar onafhankelijk ontstaan. Wanneer dus blijkt dat iets op verschillende plaatsen is ontstaan zonder dat daartussen een verband is, volgt daaruit niet per se dat het wel moest zijn ontstaan. Je kunt gelijktijdige evolutie dus niet aanvoeren als een sluitend argument voor het determinisme. Niettemin is gelijktijdige evolutie wel een reden om te vermoeden dat het geen toeval is dat iets bestaat. En stel nu dat dit vermoeden klopt: het tegelijk ontstaan van iets op verschillende plaatsen is geen toeval. Volgt daar dan uit dat het noodzakelijk is?

Uitspraken kunnen alleen waar of onwaar zijn. Er zit niets tussenin: ofwel is een uitspraak waar, ofwel is ze onwaar. Dat is anders in het geval van noodzaak en toeval. Beide staan niet tegenover elkaar zoals waar en onwaar tegenover elkaar staan. Dat moet je als volgt begrijpen. Wanneer iets noodzakelijk is, wil dat zeggen dat het onvermijdelijk is. Als iets niet noodzakelijk is, dan had het er evengoed niet kunnen zijn. Focus je nu op toeval, dan blijkt dat alles wat toevallig is niet noodzakelijk is. Maar, en dit is belangrijk, wanneer iets niet toevallig is, wil dat nog niet zeggen dat het noodzakelijk is. Neem bijvoorbeeld het feit dat voetballen en fietsen mijn hobby's zijn. Dat is uiteraard niet toevallig, want ik ben opgegroeid in een land waar beide de populairste sporten zijn. Maar dat wil niet zeggen dat het noodzakelijk is dat voetbal en fietsen mijn hobby's zijn. Het is mogelijk dat ik me toeleg op tennis, bijvoorbeeld omdat ik goed kan tennissen en niet goed kan voetballen en fietsen (quod non). Kortom, wanneer blijkt dat het geen toeval is dat verschillende technologieën tegelijk en onafhankelijk van elkaar ontstaan, kun je daaruit niet afleiden dat het ontstaan ook noodzakelijk is.

Het is belangrijk om duidelijk te zijn over mijn punt. Ik beweer in de vorige paragrafen niet dat het ontstaan van technologie niet noodzakelijk is, dat we sociale controle hebben over technologie. Ik heb de stelling niet ondermijnd dat elke technologie die ooit werd

ontworpen wel moest worden ontworpen. Mijn kritiek is hier alleen dat de argumentatie op basis van gelijktijdige evolutie tekortschiet, dat je het determinisme niet kunt verdedigen op basis van die argumentatie. Gelijktijdige evolutie bewijst niet dat iets noodzakelijk is, zelfs wanneer blijkt dat evolutie niet toevallig is. Maar zelfs als je uit gelijktijdige evolutie noodzakelijkheid zou kunnen afleiden, zou je dat niet voor alle technologieën kunnen doen. Niet alles is immers tezelfdertijd en los van elkaar ontworpen.

De evolutie van technologie

Even recapituleren. We begonnen met Heidegger en zoomden daarna in op de sociale effecten van technologie. Vervolgens focusten we op de versie van de determinismethese die over het ontstaan van technologie gaat. Tijd dus om nu in te gaan op het andere aspect van de ontwikkelingsgeschiedenis van technologie. Dat gaat over de evolutie van technologie, over de opeenvolging of vernieuwing van technologieën op het moment dat een technologie werd ontwikkeld. Dat is meteen ook het vierde en laatste voorbeeld van technologisch determinisme.

De centrale gedachte is deze: technologie leidt onvermijdelijk tot vernieuwing van een bestaande technologie; het kon niet anders dan dat een nieuwe soort technologie voortvloeit uit een oudere technologie. Een voorbeeld van deze vorm van determinisme is te vinden op de achterflap van *The New Digital Age* uit 2013 van Jared Cohen en Eric Schmidt, de gewezen CEO van Google: ‘Maar klagen over de onvermijdelijke toename van de omvang en reikwijdte van de technologiesector leidt ons af van de echte vraag. Veel van de veranderingen die we bespreken zijn onvermijdelijk. Ze komen eraan.’[100] Het is niet verwonderlijk dat deze zinnen afkomstig zijn van het voormalig hoofd van een techgigant. Dat producten sowieso tot nieuwe producten leiden, dat er een onstuitbare opeenvolging is van oude naar nieuwe technologieën is niet alleen een overtuiging die al

even meegaat, het is ook een van dé credo's van de techindustrie in het algemeen en de AI-wereld in het bijzonder.

HET QWERTY-TOETSENBORD

Straks focus ik op de argumenten. Klopt die versie van de determinismethese of niet? Eerst wil ik ingaan op de verklaring die voor de onvermijdelijke voortgang wordt gegeven. Die verklaring moet je als volgt begrijpen. Technologieën zijn als planeten – dat is de centrale bewering. Ze bewegen zich onvermijdelijk voort en worden niet bewogen door collectieve tussenkomsten en democratisch overleg tussen groepen. Dat hadden we eerder gezien; deze vorm van determinisme komt op dat vlak ook overeen met de versie van het determinisme over het ontstaan van technologie. In deze context echter wordt meestal gezegd wat wél de motor van technologische voortgang is. Vaak worden twee zaken naar voren geschoven. De eerste verklaring is het streven naar de verbetering van technologie. Die verbetering gaat niet over ethiek, maar moet je in technische zin begrijpen. Het verwijst naar de wil om technologie efficiënter of rendabeler te maken. Het proces van onvermijdelijke technologieontwikkeling, zo beweert men, wordt gedreven door het streven naar een zo laag mogelijke hoeveelheid verbruikte energie. Ten tweede verwijst men naar wetenschappelijke kennis. Er kan efficiëntere of rendabelere technologie worden gemaakt, maar vernieuwing hangt natuurlijk niet in de lucht. Die is gebaseerd op wetenschap. Het is door wetenschappelijk inzicht dat men in staat is om nieuwe efficiëntere of rendabelere technologie te maken, om technologie op de markt te brengen die technisch gezien steeds meer vernuft is.

Is dat overtuigend? Neem even aan dat het klopt dat technologie onvermijdelijk uitmondt in nieuwe technologie. Wil dat ook zeggen dat de verklaring die daarvoor wordt gegeven juist is? Wordt alle innovatie inderdaad uitsluitend gedreven door het streven naar technische

verbetering door een beperkte kring van hypergespecialiseerde ontwerpers en computerwetenschappers?

Neem het QWERTY-toetsenbord van computers, dat zo is genoemd naar de volgorde van de eerste zes letters links bovenaan het bord. Die lettervolgorde dateert uit de tijd dat er nog geen computers bestonden en toen nog met typemachines werd gewerkt. Het doel was om het snelle typen van de typisten af te remmen. Wanneer men te snel typte, zouden de mechanische letterhamertjes botsen en blokkeren. Dat werd verhinderd door het QWERTY-toetsenbord. Door deze letters links bovenaan het bord te plaatsen zouden er geen problemen ontstaan met de typemachines. Die technische, praktische overweging om te kiezen voor zo'n bord verviel toen werd overgeschakeld op typemachines zonder hamertjes: computers. Niettemin bleef men ook voor die nieuwe technologie QWERTYtoetsenborden maken – en ook nu bestaan nog steeds zulke borden. Dat had alles met gewoonte te maken. Omdat typisten nu eenmaal gewoon waren om met een dergelijk toetsenbord te werken, paste men het ontwerp van computers aan die gewoonte aan. Conclusie? Het verleden kan een schaduw werpen over technologieën, waardoor een nieuw ontwerp niet noodzakelijk uitsluitend door het streven naar efficiëntie worden gedreven.[101]

Laat hierover geen twijfel bestaan: ik beweer niet dat er geen enkele technologieontwikkeling is die voldoende kan worden verklaard door het verlangen naar efficiëntie. Mijn punt is ook niet dat in de gevallen waarin efficiëntie niet volstaat als verklaring, zoals in het voorbeeld van het toetsenbord, dat die waarde dan *geen* rol speelt in het ontwerpproces. Ik wil er hier enkel op wijzen dat niet elke innovatie *uitsluitend* wordt verklaard door het verlangen naar efficiëntie. Geloven dat elke technologische vernieuwing enkel wordt gedreven door dat verlangen is een te eenzijdige benadering. Het is een blik op technologie die wellicht wordt verdraaid door het perspectief van waaruit naar de technologie gekeken wordt, en dat is in dit geval het kader van de expert ingenieurswetenschappen of

computerwetenschappen. Er kunnen met andere woorden nog andere niet-technische zaken een rol spelen in het ontwikkelingsproces, zaken als ethiek bijvoorbeeld, zelfs wanneer blijkt dat er een onstuitbare voortgang is van de ene naar de andere technologie. En wellicht komen die andere zaken van mensen die geen AI-ontwikkelaar, computerwetenschapper of ontwerper zijn: overheden, vakbonden, sociale bewegingen, enzovoort. Je kunt in dat verband denken aan een collectief dat opkomt voor bijvoorbeeld de rechten van mensen met lichamelijke beperkingen en dat ervoor strijdt dat een ontwerp ook voor hen toegankelijk is, zelfs als dat zou betekenen dat de technologie minder efficiënt is.

DE STOOMMACHINE

Wat met de andere verklaring? Is innovatie altijd gestoeld op wetenschap? Vertrekt het efficiënter of rendabeler maken van technologie in alle gevallen vanuit wetenschappelijk inzicht afkomstig uit voornamelijk domeinen als de exacte wetenschappen en ingenieurswetenschappen? Hoewel dat vaak zo is, is dat niet altijd het geval.

De overtuiging dat nieuwe efficiëntere technologie gebaseerd is op wetenschap doet denken aan de populaire idee dat technologie toegepaste wetenschap is – net zoals muziek toegepaste wiskunde zou zijn. Die opvatting ontstond in het midden van de negentiende eeuw en werd de decennia daarna steeds meer gedeeld. Ze lag aan de basis van het bekende motto van de Wereldtentoonstelling in Chicago in 1933 – *Science Finds, Industry Applies, Man Adapts* – en werd in 1966 door wetenschapsfilosoof en fysicus Mario Bunge verdedigd in een tekst met de weinig verrassende titel ‘Technology As Applied Science’. De centrale idee is dat alle technologie gebaseerd is op wetenschappelijke kennis, en dan vooral kennis uit de natuurwetenschappen. Sterker nog, alle technologie vertrekt noodzakelijk vanuit wetenschap. Er zou in het geheel geen sprake zijn van technologie als er geen wetenschap zou zijn. Er moet eerst kennis

zijn die afkomstig is uit wetenschappelijk onderzoek en die vervolgens wordt gebruikt om technologie te ontwerpen. Geen technologie zonder wetenschap, net zoals er volgens de drie monotheïstische godsdiensten geen leven zonder God is.

Op het eerste gezicht lijkt dat correct. Tal van technologieën hadden niet kunnen bestaan zonder wetenschap: geen telefoon zonder kennis van het elektromagnetisme, geen AI zonder computerwetenschappen en geen spaceshuttle zonder fysica en scheikunde. Maar zoals ik in de inleiding al schreef: technologie bestaat al heel lang, en belangrijker nog, technologie is ouder dan wetenschap. De Inuit in Noord-Canada en Groenland bijvoorbeeld leefden ergens rond het tweede millennium voor onze jaartelling en maakten parka's, laarzen, handschoenen, iglo's, messen en kajaks om zo het hoofd te kunnen bieden aan het schrale poolklimaat. En een van de oudste vormen van technologie dateert van 30.000 voor Christus: de pijl-en-boog. Deze zaken werden gemaakt zonder enige vorm van wetenschappelijke kennis.

Sommigen vinden dat wellicht niet overtuigend. Zijn kleding en pijlen-boog dan technologieën? Ik breng in herinnering dat ik, in het spoor van anderen, een brede opvatting van 'technologie' hanteer en dat er geen goede reden is om enkel recente uitvindingen als technologie te zien. Maar zelfs als je het met dat laatste niet eens bent, zijn er voorbeelden van minder oude vondsten die laten zien dat niet alle technologie toegepaste wetenschap is, en meer in het bijzonder, dat het maken van nieuwe efficiëntere of rendabelere technologie niet altijd is gestoeld op wetenschappelijke kennis. Een bekend voorbeeld is ook hét symbool van de eerste industriële revolutie: de stoommachine.

In de geschiedenis van de technologische uitvindingen is het begin van de achttiende eeuw een belangrijke periode. Door Thomas Newcomen werd toen de eerste stoommachine gemaakt, die op dat ogenblik werd gebruikt om water uit de mijnen naar boven te pompen. Niettemin kampte dat eerste model met enkele problemen, waarvan

de lage rendabiliteit het grootste was. Dat probleem werd opgelost door James Watt, die in 1775 de eerste moderne stoommachine maakte. De oplossing van de Schot bestond uit het installeren van een extra vat, waarin de stoom werd gecondenseerd. Hierdoor bestond de stoommachine uit twee delen: de condensator, die voortdurend koel werd gehouden, en de boiler, die onveranderlijk een hoge temperatuur had, en dus niet meer telkens moest worden opgewarmd, zoals het geval was bij de eerste stoommachine. De gevolgen van deze ingreep waren van historisch belang. De resultaten van de machine stegen met minstens 20%, waardoor Watts machine gebruikt kon worden voor onder meer zwaar vrachtverkeer.

Het punt is nu dat voor die verbetering geen wetenschappelijke kennis werd gebruikt. Watt wist dat zijn model beter was dan het ontwerp van Newcomen, maar hij kon niet uitleggen welk principe aan de basis ligt van het feit dat een stoommachine met boiler en condensator rendabeler is dan een machine zonder condensator. Hij was tot zijn ontdekking gekomen door te knutselen en experimenteren, niet door bestaande kennis toe te passen. Die kennis was er eenvoudigweg niet. Men kwam het principe pas enkele decennia later op het spoor, meer precies in 1824. Het was met name fysicus Sadi Carnot die de werking van Watts machine begreep. De hogere rendabiliteit had te maken met het feit, zo ontdekte Carnot, dat warmte zich verplaatst van een lichaam met een hoge naar een lichaam met een lagere temperatuur. Kortom, de idee dat je voor technologie en innovatie altijd wetenschap nodig hebt, mag dan wel een diepgewortelde overtuiging zijn, ze klopt niet. Natuurlijk baden de meeste dingen gemaakt door mensen niet in onwetendheid, maar soms is technologieontwikkeling het gevolg van wat zoeken en wroeten, en niet van kennis, laat staan van wetenschappelijke kennis.[102]

IN DE LIJN VAN DE GESCHIEDENIS

Mijn verhaal over het toetsenbord en de stoommachine is gebaseerd op de aanname dat de opeenvolging en vernieuwing van

technologieën inderdaad onstuitbaar is. Ik ben er op de voorbije bladzijden met andere woorden van uitgegaan dat de laatste versie van de determinismethese, die over de evolutie van technologie, klopt. Het is nu tijd om die aanname wat van naderbij te bekijken. Ik concentreer me op de drie bekendste argumenten: stapsgewijze evolutie, patronen in technologieontwikkeling en de wet van Moore. Bieden die voldoende ondersteuning aan de laatste versie van het technologisch determinisme? Dat is de vraag waarop ik me nu focus en die ik, dat kan ik nu al verklappen, negatief beantwoord.

Het eerste argument is gebaseerd op een historische blik op technologie, anders dan het fenomeen van gelijktijdige evolutie, dat ook een geografische kijk veronderstelt. Een dergelijke blik wijst op lijnen in de geschiedenis van technologie, op uitvindingen die in elkaars verlengde liggen en uit elkaar voortvloeien. Neem schepen. Het eerste schip was vergelijkbaar met de kajak of kano zoals we die nu kennen: een stuk hout dat in beweging werd gebracht door erin te gaan zitten en met de handen in het water te peddelen. Toen de eerste scheepvaarders rechtop gingen staan in het schip en ondervonden dat ze bleven drijven omdat de wind in hun kleren hing, was dat de aanleiding voor een nieuwe variant: het schip met zeil. In de loop van de geschiedenis werden verder steeds kleine wijzigingen aangebracht, waardoor er wel sprake was van vernieuwing, maar zonder dat er een duidelijke breuk was tussen de modellen.[103]

Een ander voorbeeld is de ontkorrelmachine. Dat is een technologie die dient om katoenvezels van zaden te scheiden, waardoor met die vezels kleding kan worden gemaakt. De standaardversie luidt dat ze in 1793 werd uitgevonden door Eli Whitney. Op een katoenplantage in Georgia in de Verenigde Staten zou hij gemerkt hebben dat het niet eenvoudig is om de zaden van kortvezelig katoen te verwijderen, en zou het besef gegroeid zijn dat dat reinigingsproces veel tijd kost en erg eentonig is – het werk werd verricht door AfroAmerikaanse slaven. (Niet enkel AI is gestoeld op uitbuiting.) Hij zou vervolgens geobserveerd hebben hoe mensenhanden de vezels van het zaad

scheiden en die techniek vertaald hebben naar de ontkorrelmachine. Tot hier de standaardversie. In werkelijkheid hoefde Whitney niet alles uit te vinden. Hij kon zich beroepen op de Indische ontkorrelmachine die in de twaalfde eeuw al in Italië werd gebruikt en in het begin van de achttiende eeuw in de Verenigde Staten werd geïntroduceerd. Niet dat Whitney niets hoefde te veranderen. De Indische variant was gemaakt voor langvezelig katoen, terwijl zijn machine er een moest zijn voor kortvezelig katoen. Maar Whitney kon dus wel terugvallen op een bestaand model dat hij licht aanpaste tot de versie die in 1793 begon te circuleren en de katoenteelt in het zuiden van de Verenigde Staten sterk veranderde.[104]

Het feit dat zich een lijn aftekent tussen artefacten bewijst volgens sommigen dat er noodzaak in technologieontwikkeling zit.
Aangenomen dat dat terecht is, geldt dat in ieder geval niet voor alle dingen. Niet elk ontwerp sluit naadloos aan bij een voorganger, niet elke technologie ligt op een continuüm. De straalmotor en radar bijvoorbeeld zijn geen nieuwe, licht gewijzigde versies van al bestaande technologieën. Maar stel nu eens dat dat wel zo was, dat alle technologieën deel uitmaken van een stapsgewijze, graduele evolutie. Kun je daaruit dan afleiden dat elke nieuwe versie van een ontwerp er had moeten komen?

Het is verleidelijk om daarop bevestigend te antwoorden. Als je de geschiedenis kunt voorstellen als een ononderbroken licht stijgende lijn dan roept dat al snel de idee van onvermijdelijkheid of noodzaak op. Verder impliceert determinisme dat technologieën in elkaars verlengde liggen. Als blijkt dat een technologie onvermijdelijk tot een nieuwe technologie leidt, dan houdt dat in dat die nieuwe versie aansluit bij de vorige, of het zou op z’n minst niet verrassend zijn mocht dat zo zijn. Het omgekeerde is echter niet juist. Het is mogelijk dat technologieën die uit elkaar voortvloeien dat ook noodzakelijk doen, maar dat technologieën uit elkaar voortvloeien volstaat op zich niet om te besluiten dat die ontwikkeling er moest komen. Ik kan me bijvoorbeeld inbeelden dat een overheid beslist om een AI-systeem

voor gezichtsherkenning te laten ontwikkelen dat nauw aansluit bij een al bestaand ontwerp. Het feit echter dat de technologie uit een beslissing voortvloeit, een beslissing die ook niet genomen kon worden, maakt dat de ontwikkeling niet onvermijdelijk is, zelfs als er een duidelijke lijn in de technologieontwikkeling valt te ontwaren.

PATRONEN IN DE ONTWIKKELING

Misschien biedt het tweede argument voor de laatste versie van de determinismethese meer soelaas. Dat argument doet denken aan het fenomeen van gelijktijdige evolutie van hiervoor. Het wijst op patronen in de opeenvolging van oude en nieuwe technologieën. Een uitvinding volgt op een andere uitvinding, en dat is zo, niet enkel op één plaats, maar op verschillende plaatsen. Enkele voorbeelden: eerst werden steensplinters ontdekt, vervolgens was men in staat om vuur te maken; de uitvinding van het mes volgt op het vuur; het maken van metaal wordt voorafgegaan door pottenbakken; na metaal komt elektriciteit; globale grootschalige communicatie volgt op de uitvinding van elektriciteit, enzovoort. Die volgorde zie je in verschillende culturen, op verschillende plaatsen, in verschillende landen, en dat wil volgens sommigen zeggen dat een keten van uitvindingen wel moest plaatsvinden. Het terugkeren van eenzelfde opeenvolging van technologieën toont dat die evolutie onvermijdelijk is, zo gaat de redenering.

Mijn antwoord daarop is gelijkaardig aan mijn repliek op de redenering waarin wordt verwezen naar gelijktijdige evolutie. De vaststelling dat eenzelfde volgorde op meerdere plaatsen voorkomt, rechtvaardigt op zich niet de bewering dat die keten van ontwerpen noodzakelijk is. Het valt niet uit te sluiten dat het gelijke verloop toeval is. Maar zelfs als het geen toeval is, wil dat niet zeggen dat de evolutie noodzakelijk is. Er kan een goede reden zijn waarom op verschillende plaatsen de ene technologie na de andere komt, zonder dat die keten onvermijdelijk is. Het feit bijvoorbeeld dat textiel overal pas ontstaat nadat het naaien is uitgevonden, komt omdat er geen textiel zonder

naaien kan zijn. Stel echter dat het meermaals voorkomen van dezelfde volgorde volstaat als argument voor noodzakelijkheid. Dan nog kun je niet besluiten dat *elke* keten van uitvindingen onvermijdelijk is. Sommige uitvindingen die uit eerdere uitvindingen voortvloeien komen op verschillende plaatsen voor, maar dat is niet voor alles zo.

Voor de duidelijkheid wil ik dit onderstrepen. Mijn stelling is hier nu niet dat er geen noodzaak in technologische evolutie zit. Ik beweer alleen dat je die conclusie niet kunt trekken op basis van de vaststelling dat op verschillende plaatsen dezelfde volgorde van technologieën terugkeert.

DE WET VAN MOORE

Het derde en laatste argument voor de versie van het determinisme over de vernieuwing van technologie is het populairste. Het is afkomstig van ingenieur Gordon Moore, de medeoprichter van technologiebedrijf Intel Corporation. In 1960 was Moore aanwezig op een conferentie waar hij collega Doug Engelbart een boeiend verhaal hoorde vertellen over chips: een samenstelling van elektronische componenten zoals transistors die worden geïntegreerd op een plakje silicium. Ze worden gebruikt voor onder meer computers, auto's, mobiele telefoons, betaalkaarten en huisdieren. Engelbart stelde zich de vraag of schaalverkleining ook voordelig zou zijn in de context van elektronica. Hij had eerder geobserveerd dat zoiets gold voor vliegtuigen: hoe kleiner ze zijn, hoe beter ze vliegen. Zou dat ook zo zijn voor technologieën die werken met chips? Dat was wat Engelbart zich afvroeg.

Het was niet Engelbart zelf maar Moore die de vraag probeerde te beantwoorden. Hij begon tal van gegevens bij te houden: de prijs van een transistor, de hoeveelheid transistors per siliciumplakje, de verwerkingssnelheid van chips, enzovoort. Op 19 april 1965 presenteerde hij zijn resultaten in de intussen wereldberoemde paper 'Cramming more components onto integrated circuits' in het tijdschrift *Electronics*. Daaruit bleek dat chips het eerste jaar uit vier

componenten bestonden, het jaar nadien uit acht, en in 1965, het jaar van de publicatie, telde hij er al meer dan zestig. Moore had met andere woorden vastgesteld dat het aantal componenten van chips jaarlijks verdubbelde. Als je die bevindingen toen extrapoleerde naar de toekomst, dan zou dat betekenen dat er in 1978 op één chip een half miljoen componenten zouden zitten. Dat was een ietwat overdreven optimistische voorspelling. Moore herzag in 1975 daarom zijn voorspelling: niet jaarlijks maar tweejaarlijks zou het aantal transistors verdubbelen. Die formulering staat intussen bekend als de befaamde wet van Moore.[105]

Moores bevindingen zijn koren op de molen van techniekoptimisten, spreken tot de verbeelding, en ook vandaag zijn er nog zaken die ze ondersteunen. Om maar iets te noemen: smartphones worden alsmaar lichter en de beelden die we erop te zien krijgen hebben een steeds betere kwaliteit. Daarnaast weten we ook dat op de chip A8 van de iPhone 6 uit 2014 op 89 vierkante millimeter maar liefst twee miljard transistors zitten; de chip A7 uit 2013 had daar enkel de helft van. Toch rijst de vraag of de wet van Moore steun biedt aan het determinisme. Er zijn minstens drie problemen.

Ten eerste kun je niet alle technologische vernieuwingen in een wet vatten. Ja, er bestaan nog andere wetten die in deze context relevant zijn. De wet van Kryder – zo genoemd naar ingenieur Mark Knyder, voormalig directeur van Seagate – luidt dat de hoeveelheid data die per vierkante centimeter van een magnetische harde schijf wordt opgeslagen elke dertien maanden verdubbelt, waardoor ook de prijs daalt. Maar tal van andere innovaties, en wellicht zelfs de meeste, hebben geen wetmatig karakter. Wie dus meent dat alle innovaties noodzakelijk zijn, kan dat niet op basis van een wet besluiten, want je kunt niet alle innovaties beschrijven aan de hand van een wet.

Ten tweede zijn er fysieke grenzen aan de miniaturisering. Het mag dan wel waar zijn dat gedurende verschillende decennia het aantal transistors tweejaarlijks verdubbelde, die toename kan zich niet eindeloos blijven doorzetten. Er zijn grenzen aan technologische

innovatie, niet omdat ze wegens morele redenen niet zouden mogen, maar omdat de vernieuwingen fysisch gezien onmogelijk zijn.

Het derde probleem is fundamenteler. Stel, alle technologieën verbeteren in sterke mate. En neem even aan dat er geen beperkingen zijn: het aantal verbeteringen dat kan worden aangebracht, is eindeloos. Hebben we dan een sterk argument dat in het voordeel van het determinisme pleit? Met andere woorden: als iedere technologische innovatie beantwoordt aan een wet, zijn die dan allemaal ook noodzakelijk?

Op het eerste gezicht moet je daar bevestigend op antwoorden. Tal van bekende wetten – de wet van de zwaartekracht en de wet van Boyle bijvoorbeeld – gelden om het even waar en wanneer; de fenomenen die ze beschrijven, zijn noodzakelijk. Toch zijn wet en noodzaak niet onlosmakelijk verbonden; *doorgaans* gaan wetten over gebeurtenissen die in alle gevallen plaatsvinden, maar niet alle wetten gaan over een onvermijdelijke gang van zaken. Een voorbeeld is de wet van vraag en aanbod. Die luidt dat de prijs van een product afhankelijk is van de verhouding tussen vraag en aanbod. We weten dat de prijs zal dalen als het aanbod groot en de vraag klein is; is het aanbod klein en de vraag groot, dan stijgt de prijs. Maar, en dat is nu cruciaal, die gang van zaken die door de wet van vraag en aanbod wordt beschreven zou er ook niet kunnen zijn. We kunnen ons een wereld voorstellen waarin wordt afgesproken dat de prijs van een goed of dienst wordt afgesteld door een algoritme of dobbelsteen, en niet door het gedrag van koper en verkoper. Dat lijkt me onwenselijk, maar er is niets dat aangeeft dat dit volstrekt onmogelijk is. Kortom, dat een toestand in de wereld een wetmatig karakter heeft, houdt niet per se in dat die toestand ook een noodzakelijk karakter heeft.

Hetzelfde geldt ook voor de wet van Moore. Neem even aan dat die wet klopt. Iedere twee jaar verdubbelt het aantal componenten op geïntegreerde circuits, waardoor technologieën niet enkel lichter en dunner worden, maar ook goedkoper en efficiënter. Wil dat echter zeggen dat de werkelijkheid waarnaar Moores wet verwijst

onvermijdelijk is, dat het onmogelijk is dat die er ook niet had kunnen zijn? Nee, en wel hierom. Het is niet moeilijk me een wereld in te beelden van cyberutopisten en technofielen die er alles aan doen om betere chips te maken zodat de wereld er minstens in technologisch opzicht op vooruitgaat. Aan de andere kant kan ik me ook voorstellen dat de wereld wordt bestuurd door enkele leden van de amish die niet veel ophebben met technologische verbetering, die van mening zijn dat steeds snellere, efficiëntere, lichtere en goedkopere technologie onwenselijk is en ons niet veel goeds zal brengen, en die daarom technologische innovatie trachten af te remmen. In dat geval zou er geen wet van Moore zijn. Nogmaals, ik denk niet dat het onwenselijk is dat experts naar rendabiliteit en efficiëntie streven, integendeel – al mag verbetering niet tot die twee zaken worden herleid. Maar als de wet van Moore gestoeld is op de overtuiging van enkele *believers* dat technologie steeds verder moet worden verfijnd, dan laat dat wel zien dat wat door de wet van Moore wordt beschreven niet onvermijdelijk is.

Wat is nu de conclusie? Ondersteunen de drie argumenten – stapsgewijze evolutie, patronen in technologieontwikkeling en de wet van Moore – de laatste versie van technologisch determinisme? Wie uit de vorige bladzijden besluit dat er geen noodzaak zit in technologische vernieuwing gaat te overhaast te werk. Ik heb helemaal niet aangetoond dat het determinisme over de evolutie van technologieën niet klopt, en dus ook niet dat we sociale controle over technologieontwikkeling hebben. Het is op dit punt in mijn betoog nog steeds mogelijk dat technologische vernieuwing noodzakelijk is. Ik heb enkel laten zien dat de stelling dat technologische vernieuwing noodzakelijk is niet wordt ondersteund door de drie aangehaalde argumenten. Een bepaalde gang van zaken mag dan wel worden beschreven door een wet, daaruit volgt niet dat die keten van gebeurtenissen noodzakelijk is. Die conclusie kun je ook niet trekken op basis van het terugkeren van een volgorde van uitvindingen op verschillende plaatsen en de lijn die je in een opeenvolging van innovaties kunt zien.

ONVERMIJDELIJK EN VERANTWOORDELIJK

Het technologisch determinisme over het ontstaan van technologie en de versie over technologische evolutie heb ik uitgelegd en geëvalueerd. Maar wat is de relevantie daar nu van? Waarom was het zinnig om daar geruime tijd bij stil te staan?

Ik heb het al aangehaald: een van de mogelijke redenen waarom er interesse voor het technologisch determinisme bestaat, is dat het de makers van technologie volgens sommigen zou ontslaan van hun morele verantwoordelijkheid. Als technologie onvermijdelijk is, zo gaat de redenering, dan zijn de makers ook niet moreel verantwoordelijk voor onwenselijke gevolgen. Want je kunt iemand toch niet straffen voor een slecht gevolg van de technologie die jij hebt ontworpen als blijkt dat de technologie er toch sowieso zou komen? Dat is althans de gedachtegang van sommige technologen en ondernemers. Op basis van de voorgaande bladzijden kun je niet zeggen dat die redenering fout is, wel dat je je niet kunt beroepen op het technologisch determinisme om je morele verantwoordelijkheid te ontlopen, eenvoudigweg omdat de argumenten voor het determinisme tekortschieten. Als het niet vaststaat dat technologie onvermijdelijk is, kun je niet zeggen dat je niet verantwoordelijk bent, *omdat* de technologie onvermijdelijk is. Let wel, er zijn situaties waarin niemand verantwoordelijk is. Iemand kan iets moreel fout doen maar ontoerekeningsvatbaar zijn, en dus ook geen kandidaat voor straf zijn. Het determinisme daarentegen kan niet worden ingeroepen om je morele verantwoordelijkheid uit te wissen.

Stel nu echter dat de argumenten volstaan. De wet van Moore bijvoorbeeld ondersteunt de determinismethese. Houdt de onvermijdelijkheid van technologie dan in dat er geen moreel verantwoordelijke is? Kun je enkel moreel verantwoordelijk zijn voor een technologie wanneer die niet onvermijdelijk is? Nee. Om dat te staven, gebruik ik dit gedachte-experiment.

Stel je voor dat ik een ingenieur ben en dat ik een moreel onwenselijke technologie kan maken: een AI-systeem voor sollicitaties dat mensen met een anderstalige afkomst uitsluit. Ik ben echter niet door en door verdorven, en twijfel regelmatig of ik dat systeem wel effectief in elkaar zal steken. Mijn buurman echter belichaamt het kwade en is bovendien ook chirurg. Hij plant tijdens mijn slaap een chip in mijn hoofd die kan detecteren wanneer ik een definitieve beslissing zal nemen over het al dan niet maken van de technologie. Als ik zou beslissen om die niet te maken, dan zou de chip het van mij overnemen en ervoor zorgen dat ik buiten mijn wil om toch de technologie maak. De volgende ochtend word ik wakker en beslis ik om het AI-systeem in elkaar te knutselen. Ik maakte het systeem met volle overtuiging en de hersenchip heeft zich er niet mee hoeven bemoeien. Het is duidelijk dat ik moreel verantwoordelijk ben voor de technologie. Niettemin is de technologie onvermijdelijk. Als ik die immers niet had gemaakt, dan had de chip dat wel gedaan; de technologie zou er dus in alle gevallen zijn gekomen. Natuurlijk zou ik niet moreel verantwoordelijk zijn wanneer ik zou hebben besloten om de technologie niet te maken en wanneer die chip het van mij had overgenomen. Om de termen uit het vorige hoofdstuk aan te halen: ik zou dan wel oorzakelijk, maar niet moreel verantwoordelijk zijn. Maar dat is het punt niet. Van belang is dat ik zélf heb besloten om het foute AIsysteem te maken en dat ik heel goed wist waarvoor ik koos. Dat maakt mij moreel verantwoordelijk, ondanks het feit dat de technologie er sowieso zou zijn gekomen, ondanks de onvermijdelijkheid van het systeem.[106]

Dus zelfs wanneer we aannemen dat een argument als de wet van Moore het determinisme ondersteunt, pleit dat technologieontwikkelaars niet vrij van hun morele verantwoordelijkheid. Maar wat als ik fout ben? Wat als het determinisme toch inhoudt dat je moreel niet verantwoordelijk bent? Wil dat dan zeggen dat ontwerpers en computerwetenschappers, technologen en AI-ontwikkelaars hun handen in onschuld kunnen wassen? Als je meent van wel, dan komt dat omdat je gelooft dat het

technologisch determinisme inderdaad juist is, ondanks het feit dat de argumenten van daarnet tekortschieten. Maar is dat wel terecht? Dat is de vraag die ik in het laatste deel zal beantwoorden, overigens negatief. Het determinisme over het ontstaan en de vernieuwing van technologie klopt niet, althans niet voor alle technologieën, noch voor de slimme noch voor de domme.

Technologie als sociale constructie

We zijn aanbeland bij het slot van dit hoofdstuk.[107] Ik heb vier vormen van technologisch determinisme besproken. De eerste versie ging over een instrumentele manier van denken, de tweede over de sociale effecten van technologie, de derde over het ontstaan van technologie en de vierde ten slotte over technologische evolutie.

Toen ik het over de twee eerste betekenissen van de determinismethese had, heb ik telkens twee dingen gedaan. Allereerst heb ik laten zien dat de ondersteuning voor de twee versies van technologisch determinisme niet volstaat, daarna werd ook aangetoond dat beide vormen van determinisme niet kloppen. Ook wanneer werd nagedacht over de twee andere versies van de determinismethese – over het ontstaan en de evolutie van technologie – werden de argumenten voor het determinisme aangevallen. Maar dat betekent nog niet dat beide vormen van determinisme onjuist zijn. Zoals eerder al werd aangestipt: als blijkt dat de argumenten voor de bewering dat iets bestaat tekortschieten, wil dat nog niet zeggen dat het niet bestaat. Het werk is dus nog niet afgerond; er moet nog worden aangetoond dat zowel het ontstaan als de evolutie van technologie niet noodzakelijk is, dat technologie wel degelijk een kwestie is van discussie, overleg, beslissen en compromissen sluiten. Ik beroep me daarvoor op deze twee voorbeelden: algoritmische planningssoftware – slimme technologie – en de automatische spinmachine – domme technologie.

DE UURROOSTERS VAN STARBUCKS

In 2014 schreef journaliste Jodi Kantor voor *The New York Times* het geruchtmakende artikel ‘Working Anything but 9 to 5’.[108] Daarin vertelt ze het verhaal van onder anderen Jannette Navarro, een alleenstaande moeder van 22 jaar die op de rand van de armoede leeft. Haar werkgever, Starbucks, betaalt haar als barista slechts negen dollar per uur. Een ander probleem is dat haar werk haar stress bezorgt, veel stress. Dat komt doordat het Amerikaanse bedrijf sinds kort algoritmische planningssoftware gebruikt. De samenstelling van de werkroosters voor de werknemers van Starbucks gebeurt niet meer door managers, maar door een AIsysteem dat bepaalt wanneer precies iemand moet werken, een systeem dat Starbucks kocht bij het bedrijf Kronos Incorporated. Op basis van data van het gedrag van consumenten uit het verleden voorspellen algoritmen op welke dag de meeste klanten zullen komen in het koffiehuis en op welk moment van de dag het meest zal worden geconsumeerd. Een grote kans op passage betekent dat er veel barista’s nodig zijn; is de voorspelling dat er weinig klanten zullen komen, dan zijn er minder werknemers nodig. Dat is voordeliger voor Starbucks zelf, want door een dergelijke planning van de werkroosters hoeven minder werknemers te worden betaald op het moment dat ze eigenlijk niet nodig zijn. Maar het wil ook zeggen dat iemand als Navarro haar leven nauwelijks een week van tevoren kan plannen. Ze moet altijd stand-by staan. Voorspelt het algoritme op maandag dat er zaterdag veel klanten zullen zijn, dan moet Navarro zich schikken naar de berekening door het algoritme, de macht van het getal. Het is een voorbeeld van hoe AI-systemen onmenselijk kunnen zijn in de dubbele zin van het woord: er komt geen mens aan te pas bij de waarschijnlijkheidsberekening en de gevolgen zijn onleefbaar voor de werknemers.

De introductie van AI is hier duidelijk gestoeld op het streven naar efficiëntie, wat niet uitzonderlijk is. Maar automatisering kan ook gegrond zijn in andere zaken dan efficiëntie. Het doel kan repressie zijn: men wil de arbeiders het zwijgen opleggen. Dat is de verklaring voor het overschakelen van de spinmachine die met de hand wordt

aangedreven naar de geautomatiseerde spinmachine in de eerste helft van de negentiende eeuw. Laat ik dat even toelichten.

Losse vezels van wol of katoen in elkaar draaien tot draad waarmee kan worden genaaid of geweven is een erg traag en eentonig proces. Vanuit dat opzicht was de uitvinding van de spinmachine aan het eind van de achttiende eeuw zeer welkom. Maar natuurlijk waren er nog wel arbeiders nodig om de machines te bedienen. Dat werd gedaan door de spinners; hun expertise was nodig om de spinmachines te laten werken. Deze arbeiders zaten daarom in een goede onderhandelingspositie toen ze met de fabriekseigenaars rond de tafel zaten. Ze konden gunstige arbeidsomstandigheden, meer en langere pauzes en hogere lonen bedingen. Maar, en dat is cruciaal, dat was niet naar de zin van het kapitaal, de ondernemers. Die beriepen zich daarom op wel erg drastische ingrepen om de macht van de spinners te breken. Na een staking van drie maanden verzamelden ze ingenieurs met de vraag om een nieuwe technologie uit te vinden: de automatische spinmachine. Die zou de oude handmatige machine moeten vervangen, waardoor de expertise van de spinners niet meer nodig zou zijn en ze geen hoge eisen meer konden stellen. Kortom, men had een nieuwe technologie op het oog om de arbeiders eronder te houden. In 1824 kregen de fabrieksdirecteuren wat ze wilden: uitvinder Richards Roberts maakte de eerste automatische spinmachine. De spinners werden niet massaal ontslagen; ze zorgden nog voor het herstel en onderhoud van de machines. Maar hun sterke onderhandelingspositie stortte wel ineen, net zoals de bereidheid om het werk neer te leggen.[109]

TINA, THERE IS NO ALTERNATIVE

De automatische spinmachine is een voorbeeld van hoe een oude technologie door een nieuwe wordt vervangen, de planningssoftware van Starbucks is een illustratie van de introductie van een nieuwe technologie. Beide casussen tonen dat de twee vormen van determinisme over technologieontwikkeling onjuist zijn. Ik breng in

herinnering wat die naar voren brachten. Alle technologie ontstaat onvermijdelijk, los van tijd en plaats; de evolutie van de ene naar de andere technologie is noodzakelijk, die zou er ook zijn als we de tijd zouden kunnen terugdraaien. Eerder heb ik laten zien dat de argumenten voor beide interpretaties niet volstaan, maar door de twee casussen weten we nu ook dat geen van beide interpretaties klopt, dat niet elk ontstaan van technologie en iedere evolutie gedetermineerd is. Waarom?

Roberts' uitvinding en het algoritme van Starbucks zijn duidelijk een effect van een welbepaalde sociale relatie. De automatische spinmachine werd op verzoek van de fabrieksdirecteuren uitgevonden om de macht van de spinners te breken; het AI-systeem werd ingevoerd met efficiëntie als reden, om geen werknemers te hoeven betalen wanneer ze geen nut hebben. Beide technologieen zouden buiten die context niet bestaan; mocht die context er niet zijn, dan zouden beide technologieën ook niet zijn ontwikkeld. Het punt is nu dat beide sociale relaties ook het product zijn van de industriële samenleving die zelf geen natuurwet volgt, die op haar beurt niet noodzakelijk is. Niet elke samenleving is geënt op het onderscheid tussen aan de ene kant het kapitaal en de vrije ondernemers, en aan de andere kant het werk en de werknemers. Bovendien is de industrialisering in het licht van de geschiedenis van vrij recente datum. Ze heeft niet altijd bestaan en kwam op vanaf het begin van de achttiende eeuw.

Mijn stelling luidt daarom als volgt. De spinmachine en planningssoftware zijn geen technologieën die wel moesten worden gemaakt. Ik spreek me hier niet uit over de wenselijkheid ervan, maar een wereld zonder beide was en is niet volstrekt onmogelijk. Wanneer we de tijd zouden kunnen terugdraaien, dan kun je er niet van op aan dat ze altijd opnieuw zouden bestaan. Die conclusie is gebaseerd op de volgende redenering. In een samenleving die niet is gecentreerd rond arbeid en kapitaal zouden die technologieën niet zijn uitgevonden. En aangezien zo'n samenleving niet onvermijdelijk is, zijn de spinmachine en het AI-systeem van Starbucks dat ook niet. Dat

is het geval voor deze technologieën, waarschijnlijk ook voor veel andere, en misschien zelfs voor alle technologieën. Dat laatste kun je niet uitsluiten, net zoals het niet onmogelijk is dat sommige wel onvermijdelijk zijn.

De versie van het determinisme over het ontstaan en de ontwikkeling van technologie klopt dus niet. Let wel, ik beweer hier nu niet dat geen enkele technologie onvermijdelijk is, wel dat niet alle technologieën dat zijn. Sommige technologieën passen niet binnen het kader van het determinisme, wat natuurlijk wel een reden is om te vermoeden dat er nog andere voorbeelden zijn die niet in het wereldbeeld van de determinist passen. Dat is in ieder geval wat de casussen van de automatische spinmachine en de planningssoftware van Starbucks duidelijk maken. Maar je kunt daar ook de *Big Five* aan toevoegen. Neem Google. De twee die het bedrijf oprichtten, Page en Brin, konden in plaats van de techgigant uit de grond te hebben gestampt ook docenten geworden zijn aan hun alma mater. Ze konden ervoor gekozen hebben om de zoekmachine niet te commercialiseren. En het was niet onvermijdelijk dat ze opteerden voor een ander model dan het bedrijfsmodel dat hen steenrijk maakte, een model dat niet is opgehangen aan *targeted advertising*. Het zou er in ieder geval voor hebben gezorgd dat we de voorbije tien jaren niet zoveel aan privacy hadden ingeboet.

Stel nu echter dat mijn redenering hapert. Het kon niet anders dan dat de industriële samenleving ontstond en dat die is gebaseerd op het onderscheid tussen werkgever en werknemer. Zou dat dan ook de doodsteek betekenen voor mijn aanval op het determinisme? Is het gevolg dat de technologieën uit mijn twee voorbeelden wel degelijk een onvermijdelijk karakter hebben?

Het antwoord daarop is ontkennend. Het mag dan wel zo zijn dat de spinmachine en de AI van Starbucks niet zouden bestaan buiten de industriële samenleving, de relatie tussen die technologieën en industrialisering is niet dezelfde als de relatie die wordt beschreven door de gaswet van Boyle. De spinmachine en planningssoftware

vloeien niet voort uit de industriële samenleving zoals de stijging van de luchtdruk volgt uit de afname van het volume van de lucht in een fietspomp. Het is onvermijdelijk dat de luchtdruk stijgt, maar het is niet in stenen gebeiteld dat de spinmachine of het algoritme worden gebruikt. Dat komt doordat de technologie het gevolg is van een keuze, een beslissing van een of meerdere personen in een bepaalde sociale rol: de werkgever, het management, enzovoort. Dat bleek duidelijk in het geval van de spinmachine, maar dat is ook zo in het voorbeeld met Starbucks. Toen Howard Schultz, de voormalige CEO van het bedrijf, hoorde van de verhalen die door *The New York Times* naar buiten werden gebracht, bood hij zijn excuses aan en zei hij dat het opstellen van de werkroosters voor herziening vatbaar is. Zeker, de keuze om een automatische spinmachine of planningssoftware te gebruiken hangt niet in de lucht; ze is gestoeld op redenen: het onderdrukken van protest en het streven naar efficiëntie bijvoorbeeld. Maar een gegronde beslissing is nog een steeds een beslissing, en dat houdt in dat er meerdere opties waren en dat het dus ook anders had kunnen lopen.

DE SCOT-BENADERING

Het is goed om duidelijk te zijn over de achtergrond van mijn verhaal dat ik op de vorige bladzijden heb geschetst. Dat verhaal sluit aan bij een stroming uit de tweede helft van de vorige eeuw die bekend staat als het sociaal constructivisme en die is verbonden met het werk van onder anderen Michel Foucault. Meer in het bijzonder valt mijn betoog binnen de zogeheten SCOT-benadering, voluit de *Social Construction of Technology*. Een van de bekendste vertegenwoordigers van die benadering is de al eerder vermelde Bijker, wiens sociaal constructivistische kijk op de ontwikkeling van de fiets erg invloedrijk is geweest.

Het sociaal constructivisme werpt geen astronautische, afstandelijke blik op technologie. Integendeel, het werkt op een concreet niveau en onderzoekt daar de ontstaansgeschiedenis van technologieën. Men

legt de sociale factoren bloot die hebben geleid tot het bestaan van datgene waarvan wordt beweerd dat het een sociale constructie is. Dat wil zeggen: de interpersoonlijke relaties, groepen en sociale rollen worden gedetecteerd die vorm hebben gegeven aan de ontwikkeling van technologie. De voorbeelden van Starbucks en de automatische spinmachine passen binnen die context, net zoals Bijkers onderzoek toont dat de verschillen in voorkeuren tussen mensen hebben geleid tot verschillende soorten fietsen.

Andere voorbeelden zijn de studies die de financieringsstromen voor AI aan het licht brachten.[110] Concreet: het is bekend dat veel technologieën die geen AI zijn een militaire oorsprong hebben – het internet, de Aviator-zonnebril, maandverband –, maar intussen weten we ook dat ook het leger heeft bijgedragen aan de hoogtijdagen van AI. Het event dat vaak wordt gezien als het begin van AI – het Summer Research Project on Artificial Intelligence in 1956 aan het Dartmouth College in New Hampshire – werd gefinancierd door het Office of Naval Research van het Amerikaanse ministerie van Defensie. Ook het *Defense Advanced Research Projects Agency* (DARPA), dat ik eerder in dit boek al aanhaalde, is daar onderdeel van. Haar hoofdopdracht is om onderzoeksgelden te beheren. Robert Sproull, voormalig directeur van het instituut, zei ooit dat zeker twee generaties computerwetenschappers werden gefinancierd door DARPA, en dat dus onder meer technologie die taal begrijpt afkomstig is van geld van het ministerie van Defensie. Hetzelfde instituut heeft ook bijgedragen aan de ontwikkeling van zelfrijdende auto's. In 2004 werd de eerste *DARPA Grand Challenge* georganiseerd. Dat is een wedstrijd voor auto's die zelfstandig meer dan tweehonderd kilometer moeten rijden in de woestijn tussen Nevada en Californië. Het winnende team wordt beloond met een smak geld – het eerste jaar met één miljoen Amerikaanse dollar, het jaar erop verdubbelde dat bedrag. Ten slotte is het wellicht ook geen toeval dat Amazon een van haar meest recente hoofdkwartieren in de buurt van het Pentagon liet bouwen.

Op een sociaal constructivistische manier naar technologie kijken betekent dus dat je de aandacht vestigt op de sociale oorzaak van technologie, en niet op de sociale gevolgen. De vraag is niet op welke sociale rollen en relaties technologie een goed of slecht effect uitoefent. Het is andersom. Binnen een dergelijk kader luidt de vraag welke sociale rollen en relaties aan de basis van technologie liggen. Voor wie is de technologie gemaakt? In welke context is ze ingebed? Wie besliste? Welke belangen speelden een rol? Toch kan die benadering ook verbonden zijn met het perspectief van waaruit wordt gefocust op de gevolgen van technologie, en wel op de volgende manier.

Stel, een bedrijf gebruikt een AI-systeem dat werknemers voortdurend volgt met de bedoeling om te weten of ze efficiënt werken. Het probleem echter is dat nogal wat werknemers zich daar zeer ongemakkelijk bij voelen, zodanig dat sommigen lijden onder die onafgebroken *surveillance*. Welnu, voor sommigen zijn die negatieve gevolgen van het gebruik van het controlesysteem op de werknemers een reden om zich voor het sociaal constructivisme te interesseren. Op grond van de onwenselijke invloed wil men de wordingsgeschiedenis van de technologie achterhalen, wil men uitpluizen welke partijen betrokken waren bij de keuze voor de technologie. Daar kan een psychologische reden voor zijn. Om onvrede met een probleem te kanaliseren helpt het voor sommige mensen om de oorsprong van een probleem te kennen. Maar er kunnen ook morele motieven spelen. Men kijkt naar de sociale geschiedenis, in de hoop een verantwoordelijke te vinden voor het probleem dat zich nu stelt.

De band tussen het perspectief dat focust op de sociale gevolgen en een sociaal constructivistische invalshoek kan ook een activistisch karakter hebben. Neem opnieuw het controlesysteem van daarnet. De technologie is niet alleen vervelend, ze sorteert ook onwenselijke effecten op een aantal werknemers. De druk om zo hard en efficiënt als mogelijk te werken is zo groot dat het gevaar dreigt dat sommigen niet meer kunnen functioneren en voor langere tijd thuis zullen moeten

blijven. Voor een aantal is het daarom duidelijk: de technologie zou er beter niet zijn. Omwille van de slechte effecten, wil men terug naar een werkcontext zonder AI en *surveillance*. De vraag die dan opduikt, is of dat wel kan. Een bedrijf zonder de technologie is wenselijk, maar is het ook mogelijk? Sommigen werpen een sociaal constructivistische blik op technologie met die vraag in het achterhoofd. Men redeneert dan als volgt. Wanneer je een goed zicht krijgt op de sociale context waarbinnen het systeem werd ontwikkeld, dan weet je ook dat het niet onvermijdelijk was dat de technologie toen werd gemaakt. En als je weet dat de technologie er ook niet had kunnen zijn, dan kan het ook nu anders. Dat is althans de achterliggende redenering, die is gestoeld op de aanname dat er een band is tussen vermijdbaar en veranderbaar. Men neemt aan dat iets nu anders kan, omdat het toen anders kon geweest zijn. Straks ga nog even in op die aanname.

TUSSEN DROOM EN DAAD

Op de openingsbladzijde van zijn studie *The Social Construction of What?* uit 1999 merkt filosoof Ian Hacking op dat in de voorbije decennia over heel veel zaken is gezegd dat het sociale constructies zijn: gender, emoties, wetenschap, ziekte, vluchtelingen, en tal van andere zaken.[111] Zelfs feiten worden door sommigen sociale constructies genoemd, en in 2020 opperde de beroemde filosoof Giorgio Agamben nog dat de corona-epidemie een uitvinding – lees: een sociale constructie – van de (bio)politiek is. Ik betwijfel sterk of dat allemaal even plausibel is, maar toch wil ik daar de automatische spinmachine en planningssoftware van Starbucks aan toevoegen. Beide technologieën zijn sociale constructies en dus hadden ze er evengoed niet kunnen zijn. Ze komen in dat opzicht overeen met juridische of politieke wetten, maar verschillen van een fenomeen als de stijging van de luchtdruk wanneer het volume van de lucht in een fietspomp afneemt.

Om onduidelijkheid te vermijden, wil ik dit onderstrepen: wellicht zijn veel technologieën sociale constructies, maar ik beweer niet dat dat zeker zo is voor *alle* technologieën. Mijn stelling is dat sommige sociaal geconstrueerd zijn, en dat dus niet elke technologie gedetermineerd is. Die bewering mag dan glashelder zijn, toch wil ik nog even de aandacht vestigen op een vijftal mogelijke misverstanden. Het eerste heeft te maken met het verschil tussen kunnen en willen, het laatste met de band tussen determinisme en constructivisme.

Eén. Eerder heb ik erop gewezen dat er naast de technologische variant nog andere vormen van determinisme zijn, waaronder de genetische variant. Ik voeg daar nu het sociaal determinisme aan toe. Het wil zeggen dat de wensen die samenhangen met sociale rollen, processen en relaties in alle gevallen bevredigd zullen worden. Wanneer het over dit type determinisme in de context van technologie gaat, dan is duidelijk dat het altijd gepaard gaat met sociaal constructivisme. Als de wens van een groep mensen dat een bepaalde fiets wordt gemaakt sowieso wordt vervuld, dan is die fiets een sociale constructie. Het omgekeerde is echter niet waar. Sociaal constructivisten kunnen sociaal deterministen zijn, maar dat is niet per se het geval. Je kunt van mening zijn dat technologieën sociale constructies zijn, omdat ze de uitkomst zijn van de wil van een aantal mensen in een sociale relatie. Maar tegelijk kun je ook beseffen dat de wens dat een bepaalde technologie wordt gemaakt niet kan worden vervuld, bijvoorbeeld omwille van fysische beperkingen of omdat het organisatorisch onmogelijk is.

Een andere manier om hetzelfde te zeggen is dat er een kloof bestaat tussen willen en kunnen. Veel technologieën die wel kunnen worden ontwikkeld, zijn door veel mensen niet gewild. Ik denk in de eerste plaats aan de atoombom of *killer robots* uit het vorige hoofdstuk, maar er zijn ook minder dramatische voorbeelden. Een app die de ontwikkelaars toegang geeft tot de gegevens van al je contacten: het is mogelijk, maar de meesten onder ons vinden dat terecht onwenselijk. Andersom kunnen veel mensen een technologie

wensen, terwijl men die eigenlijk niet kan maken. Denk aan de wet van Moore. Een aantal technofielen kan wel verlangen dat ook in 2030 het aantal transistors enorm blijft toenemen, maar dat zal niet gebeuren wegens de fysieke grens waarop de ontwerpers onherroepelijk stoten. Niet alles wat mogelijk is, is gewenst, maar ook niet alles wat gewenst is, is mogelijk. Om nog eens de bekende woorden van schrijver Willem Elsschot uit zijn gedicht *Het Huwelijk* uit 1910 vanonder het stof te halen: tussen droom en daad staan niet alleen wetten, maar ook praktische bezwaren.

Twee. De bewering dat een technologie onvermijdelijk uit een oude volgt, geldt zeker niet voor alle dingen. Tussen de oude en nieuwe spinmachine stond een sociale relatie in een specifiek tijdsgewricht die er niet had kunnen zijn, waardoor ook de geautomatiseerde versie niet onvermijdelijk is – dat zagen we eerder al. Maar wat als dat nu voor alle technologieën zou gelden? Moeten we dan besluiten dat eerdere technologieën *geen* rol spelen, dat de ontwikkeling van nieuwe ontwerpen altijd losstaat van bestaande ontwerpen?

Denk aan de begindagen van de pc.[112] Toen werd die bestuurd door het MS-DOS-programma dat een werkgeheugen van 640 kb had. Hoewel die capaciteit toen groot was, bleek die al snel tegen haar grenzen aan te lopen. Toch heeft men pas veel later een geheel nieuw systeem ontwikkeld: Windows 95 van Microsoft. In eerste instantie ontwikkelde men een programma dat gebaseerd was op het oorspronkelijke model maar dat een uitgebreider werkgeheugen had. Dat sluit aan bij wat we eerder over het schip en de spinmachine hebben gezien. Voor de ontwikkeling van veel technologie (maar niet voor alle) gaat men uit men van een bestaande technologie en maakt men vervolgens een nieuwe efficiëntere versie die toch niet volstrekt anders is. Dat heeft doorgaans met economische motieven te maken. Als geld in een technologie wordt gepompt en men daarna een efficiëntere maar geheel nieuwe technologie zou ontwerpen, dan zou dat die investeringen tenietdoen. Om dat te voorkomen wordt het

bestaande model als vertrekpunt voor de nieuwe efficiëntere versie genomen.

Dus zelfs als alle technologieën sociale constructies zijn, wil dat niet per se zeggen dat technologieontwikkeling *uitsluitend* het resultaat is van een wens van een aantal mensen en van wat op fysisch of organisatorisch vlak mogelijk is. Technologieontwikkeling in het heden kan ook worden bepaald door technologieontwikkeling in het verleden. Al wil ik wel nog eens benadrukken dat het helemaal niet zeker is, integendeel, dat de eerste technologie onvermijdelijk uitmondt in de tweede. Een meer plausibele bewering is dat een bestaande technologie de ontwikkeling van een nieuwe versie niet noodzakelijk maar wel waarschijnlijk maakt, en dat dit te maken heeft met de financiële investeringen in de eerste technologie.

Drie. Vaak worden de begrippen ‘echt’ en ‘geconstrueerd’ tegenover elkaar geplaatst. Wat geconstrueerd is, is niet echt; enkel wat geen geconstrueerd karakter heeft, is echt. Als een technologie een sociale constructie is, dan houdt dat dus vanuit die optiek in dat die niet echt is. Dat is uiteraard niet waar. De planningssoftware van Starbucks bijvoorbeeld is de uitkomst van een sociale relatie, maar is wel ingebed in hardware met een tast- en meetbaar karakter. Bovendien zijn de effecten van het AI-systeem, ondanks het feit dat de technologie een sociale constructie is, sterk voelbaar. Dat is althans wat blijkt uit de verhalen die journaliste Kantor in *The New York Times* optekende, verhalen die zijn geweven rond stress, onzekerheid en vermoeidheid. Of denk aan de voorbeelden die in de verschillende hoofdstukken terugkeerden: de ongelijke behandeling van mensen door de gebiaste slimme systemen die worden gebruikt door rechtbanken, politie, werkgevers, enzovoort. Dat zijn onwenselijke effecten van systemen die sociale constructies zijn.
Maar dat maakt de effecten niet minder echt.

Vier. Ik heb daarnet de spinmachine en planningssoftware aangehaald om te laten zien dat de twee laatste versies van technologisch determinisme – die over het ontstaan en de evolutie

van technologie – niet kloppen. Omdat ik nadien heb aangetoond dat beide technologieën sociale constructies zijn, zou nu de indruk kunnen ontstaan dat technologisch determinisme en sociaal constructivisme elkaar uitsluiten. Dat is uiteraard juist wanneer het gaat over de laatste twee varianten van determinisme. Als een technologie een sociale constructie is, dan had die evengoed niet kunnen bestaan, en dat is precies het tegenovergestelde van wat de determinist beweert. Niettemin kunnen beide ook samengaan. Ik verwijs dan naar de tweede vorm van determinisme, die luidt dat een technologie onvermijdelijk sociale effecten heeft. Zeker niet alle technologieën hebben noodzakelijke sociale effecten – dat liet ik eerder al zien – maar laten we aannemen dat minstens één technologie dat wel heeft. Als je daar op inzoomt, dan blijkt dat zo'n ontwerp het resultaat kan zijn van een sociaal proces, en dat het er bijgevolg ook niet had kunnen zijn. Het is met andere woorden niet uitgesloten dat een technologie onvermijdelijke gevolgen heeft op sociaal vlak, terwijl die tegelijkertijd het resultaat is van een welbepaalde sociale relatie in een bepaald tijdsgewricht, en dus zelf niet onvermijdelijk is.

Vijf. Een technologie kan dus én een sociale constructie zijn én iets anders determineren, maar een technologie kan ook sociaal geconstrueerd en zelf gedetermineerd zijn, zij het hier niet in de zin dat de technologie wel had moeten worden gemaakt. Om dat te illustreren wend ik me tot een technologie die net als het internet en de computer militaire wortels heeft en vandaag in zeker dertig landen wordt gebruikt om elektriciteit op te wekken: de kerncentrale.

Het lijdt geen twijfel dat kerncentrales, anders dan pakweg natuurverschijnselen, sociale constructies zijn. Ze zijn het resultaat van een beslissing van minstens drie stakeholders – politici, industriëlen en wetenschappers –, een keuze die men ook niet had kunnen maken. Dus het feit dat er kerncentrales worden gebruikt, is niet onvermijdelijk; dat kon ook anders geweest zijn. Maar, en dat is nu het punt, op het moment dat de beslissing werd genomen om deze technologie te ontwikkelen kan die voor een deel niet anders zijn dan

hoe ze is. Ik heb het hier over wat een kerncentrale tot een kerncentrale maakt: via de splijting van uranium energie opwekken die wordt omgezet in elektriciteit. Met andere woorden: *dat* de technologie bestaat, is niet noodzakelijk, maar *wat* een kerncentrale is, is dat wel – minstens ten dele. Dat komt doordat de ontwikkeling van de kerncentrale, anders dan de stoommachine, gebaseerd is op wetenschap, en preciezer gezegd, op kennis uit de kernfysica. En die kennis wordt op haar beurt gedicteerd door hoe de werkelijkheid op fysisch vlak in elkaar steekt. Natuurlijk kon het domein van de nucleaire fysica in een andere wereld met andere financieringskanalen ook niet zijn ontstaan. Als discipline is ze een sociale constructie, maar de kennis die ze voortbrengt, is dat niet. Wetenschappelijke kennis van fysici volgt de structuur van de werkelijkheid en kan bijgevolg niet anders zijn dan ze is. En dat heeft vanzelfsprekend gevolgen voor technologie die op die kennis is gebaseerd en waarvan het bijgevolg onjuist is te beweren dat ze niet onvermijdelijk is.

Een technologie kan dus én gedetermineerd én sociaal geconstrueerd zijn. Wie denkt dat dit een flagrante contradictie is, moet het volgende voor ogen houden. Dat een technologie geconstrueerd is, betekent dat ze voortvloeit uit een beslissing van een groep mensen. Het geconstrueerde karakter verwijst dus naar de wordingsgeschiedenis. Wanneer ik hier beweer dat een technologie gedetermineerd is, heb ik het over de technologie zelf, en niet over het ontstaan ervan. De structuur van een kerncentrale kan niet anders zij dan ze is, en dat volgt uit de opbouw van de werkelijkheid, een opbouw die de ijzeren wet van de kernfysica aan de technologie oplegt.

CONSTRUCTIVISME EN ACTIVISME

Het is duidelijk dat er in mijn verhaal over sociaal constructivisme iets op het spel staat. Ik heb die benadering geïntroduceerd omdat het de twee laatste versies van de determinismethese op z’n minst uitdaagt en zelfs ondermijnt. Daarnaast is mijn verhaal ook in praktische zin

relevant. Dat komt omdat de determinismethese vaak opduikt in al dan niet politieke discussies.

Neem de *killer robots* uit het vorige hoofdstuk, een technologie waarover tot op het niveau van de Verenigde Naties regelmatig wordt gediscussieerd. Een vaak terugkerend argument van de voorstanders is dat ze er sowieso zullen komen. De redenering luidt dat *killer robots* in alle gevallen in de toekomst zullen worden gebruikt, en dus kun je ze evengoed nu toelaten. Wanneer de internationale gemeenschap daarin zou meegaan, dan kan dat mogelijk ernstige reële gevolgen hebben. Zo bestaat de kans op een humanitaire ramp – robots kunnen fouten maken en talloze burgers doden – en is het risico dat de drempel om oorlog te voeren wordt verkleind, aangezien door het gebruik van zulke technologie minder militair personeel effectief de wapens hoeft op te nemen. Wanneer je nu echter aan de hand van de automatische spinmachine kunt aantonen dat het technologisch determinisme niet noodzakelijk klopt, dan heb je een reden om de argumentatie van de voorstander van *killer robots* minstens ter discussie te stellen. Als zulke technologie niet gedetermineerd is, waarom zou dat dan zo zijn voor hypergeavanceerde AI-systemen? Die vraag dwingt de voorstander om zijn redenering te onderbouwen, wat in het nadeel van de voorstander kan zijn. Dat is misschien weinig of niet relevant wanneer het over technologieën zonder (veel) impact gaat, maar het is anders wanneer het bijvoorbeeld over volautomatische wapensystemen gaat die mensenlevens kunnen eisen.

Sommigen voegen daar het volgende aan toe. Een sociaal constructivistisch perspectief is nuttig, omdat het inhoudt dat ook de huidige situatie kan worden veranderd. Als een technologie die nu bestaat ook niet had kunnen bestaan, dan is een samenleving zonder die technologie vandaag mogelijk, zo luidt de redenering. Klopt dat?

Het is duidelijk dat deze interesse voor het sociaal constructivisme vanuit een bepaalde hoek komt. Het gaat hier doorgaans over mensen die negatief staan tegenover een technologie en die precies daarom

willen dat ze verdwijnt. Dat negatieve oordeel kan met verschillende zaken te maken hebben, zaken die over de verschillende hoofdstukken heen regelmatig terugkeerden. We hebben gezien dat zowel de ontwikkeling als het gebruik van AIsystemen on-wenselijke ecologische effecten heeft. Slimme technologie trainen bijvoorbeeld gaat gepaard met de uitstoot van minstens 200.000 kilogram CO_2. Mensen kunnen ook kritisch staan tegenover een technologie omwille van politiek misbruik dat ervan wordt gemaakt of kan worden gemaakt. Neem het voorbeeld dat ik in het eerste hoofdstuk al aanhaalde. In de aanloop naar de Amerikaanse presidentsverkiezingen in 2016 werd via anonieme accounts op Twitter en Facebook foutieve informatie verspreid over Hilary Clinton, moslims en Mexicaanse immigranten. Die accounts waren in handen van het bedrijf Internet Research Agency, beter bekend als de *Trolls from Olgino*, dat in Sint-Petersburg is gevestigd en als doel had het beïnvloeden van de uitslag van de verkiezingen. Een andere mogelijke bron van een negatief oordeel heeft te maken met reële of mogelijk onwenselijke gevolgen op persoonlijk vlak. Daar zijn veel voorbeelden van: nogal wat technologie is een regelrechte aanslag op onze privacy; het gebruik van sociale media kan leiden tot symptomen van depressie; sociale media kunnen worden gebruikt voor *cyberbullying*, en we weten dat dit kan leiden tot angst en tot zelfmoord(gedachten).

Die problemen zijn te ernstig om ze links te laten liggen en de bezorgdheid erover is terecht, maar toch is de redenering niet zonder problemen. Wie vindt dat een technologie beter verdwijnt omdat zij gepaard gaat met negatieve gevolgen, pleit eigenlijk voor een wereld zonder (veel) technologie. Er zijn immers weinig of geen technologieën die volstrekt risicoloos zijn. Alles of bijna alles wat we maken brengt gevaren met zich mee, en dus is een risico op zich onvoldoende om te pleiten voor het verdwijnen van de technologie. Bovendien houdt het geen rekening met de positieve effecten. Natuurlijk weet ik dat er dingen worden gemaakt met volstrekt onwenselijke doelen voor ogen, en natuurlijk besef ik dat er onenigheid bestaat over het al dan niet wenselijke karakter van de

effecten van sommige technologieën. Tegelijkertijd kun je niet ontkennen dat er veel uitvindingen zijn waarover nagenoeg iedereen het eens is dat het goed is dat ze bestaan. Ik denk in de eerste plaats aan technologie om ziekten op te sporen of te genezen. Pleiten voor het verdwijnen van een technologie *uitsluitend* op basis van de negatieve gevolgen is dus te kort door de bocht, niet alleen omdat je dan eigenlijk veel dingen zou moeten veroordelen, maar ook omwille van de positieve gevolgen. Er is een *trade-off* tussen de verschillende gevolgen nodig, en geen eenzijdige blik op enkel de onwenselijke effecten.

Ik neem even aan dat die afweging is gemaakt en dat men besluit dat een bestaande technologie er beter niet zou zijn. Kan een constructivistische kijk dan steun bieden aan het streven van de activist naar een wereld zonder die technologie? Kun je uit het inzicht dat een technologie die nu bestaat maar er evengoed niet had kunnen zijn, afleiden dat een samenleving zonder die technologie mogelijk is?

Als je daar bevestigend op antwoordt, is dat begrijpelijk. Tal van zaken die nu het geval zijn maar dat evenzeer niet geweest konden zijn, kunnen vandaag nog worden uitgewist. Als de verlichting van mijn auto slecht werd geïnstalleerd door de garagehouder omdat die gehaast was, dan was dat vermijdbaar. Maar die fout kan nog steeds worden rechtgezet – de garagehouder kan de verlichting nu onmiddellijk correct installeren. Een ander voorbeeld: dat er vandaag de dag kerncentrales worden gebruikt, is het gevolg van een keuze in de vorige eeuw die toen ook niet gemaakt had kunnen worden maar waarop men vandaag nog kan terugkomen. Toch wil ik benadrukken dat het ene niet noodzakelijk uit het andere volgt. Het is niet omdat iets anders gelopen kon zijn, dat je het nu nog ongedaan kunt maken. Veranderbaar volgt niet per se uit vermijdbaar. Denk aan terreur. Iemand met slechte bedoelingen kan een dodelijke aanslag plegen. De dood van de slachtoffers was vermijdbaar, maar toch is hun dood onherroepelijk.

Dat laatste geldt ook voor technologie. Door recente ontwikkelingen op het vlak van robotica en AI is het mogelijk om een technologie te ontwikkelen die, op het moment dat ze in gebruik is genomen, niet meer kan worden tenietgedaan. Dat ontwerp volgt dan wel uit een beslissing en kon er dus evengoed niet geweest zijn, maar nu de technologie er toch is, is dat onherroepelijk zo. Kortom, als je vindt dat een wereld zonder een bepaalde technologie een betere wereld zou zijn en als je bovendien weet dat die technologie een sociale constructie en dus niet onvermijdelijk is, dan is die kennis op zich geen reden tot optimisme. Het is niet omdat een bestaande technologie er evengoed niet had kunnen zijn dat een wereld zonder die technologie nu mogelijk is. Het kan zijn dat het gebruik van een technologie niet onvermijdelijk is en nog steeds kan worden teruggedraaid, maar het is niet zo dat iets evenzeer niet had kunnen bestaan en nu *dus* uitgewist kan worden.

De activist die zich verzet tegen een technologie zou daarop kunnen antwoorden dat dit alleen geldt voor hypergeavanceerde AIsystemen, en niet voor alle slimme technologieën, laat staan voor technologie in het algemeen. Het merendeel van de bestaande technologieën wordt gecontroleerd door bedrijven en overheden, en dat is een reden voor hoop. Wij, activisten, kunnen immers druk uitoefenen op die instanties, kunnen burgers informeren en mobiliseren, waardoor de stekker mogelijk uit een technologie wordt getrokken. Die opmerking is dan wel terecht, toch wil ik de volgende twee zaken onderstrepen.

Ten eerste: dat een technologie er ook niet had kunnen zijn, houdt niet in dat het bestaan van die technologie nu kan worden teruggedraaid – dat zagen we daarnet –, maar uit het feit dat nu een punt kan worden gezet achter het gebruik van een technologie kun je ook niet concluderen dat men daadwerkelijk zal stoppen met de technologie. Er kunnen immers allerlei redenen zijn die ervoor zorgen dat iets niet *zal* veranderen terwijl het wel *kan* veranderen. Denk aan auto’s. In het licht van de geschiedenis is dat een jonge technologie, een ontwikkeling die er ook niet had kunnen zijn en waarover men kan beslissen dat het gebruik ervan moet worden stopgezet. Mijn punt is

nu niet dat dat laatste wenselijk is, maar wel dat het geenszins onmogelijk is dat in de toekomst wordt besloten dat auto's beter uit het straatbeeld zouden verdwijnen. Tegelijk is duidelijk dat een dergelijke beslissing al dan niet terecht op veel weerstand zou stuiten, bijvoorbeeld omwille van economische belangen van autofabrikanten, maar ook op basis van het feit dat overheden gedurende verschillende decennia sterk hebben ingezet op een samenleving met auto's door straten en snelwegen aan te leggen. Als die bezwaren worden gedeeld door belangrijke stakeholders dan is de kans reëel dat auto's niet zullen verdwijnen, zelfs al kunnen ze verdwijnen.

Ten tweede: wanneer wordt besloten om een technologie die kan worden uitgeschakeld daadwerkelijk uit te schakelen, wil dat niet zeggen dat er een korte lijn is tussen wens en werkelijkheid. Het realiseren van de mogelijkheid tot verandering kan bijzonder traag en moeizaam verlopen. Neem opnieuw de auto. Als men zou besluiten dat die moet verdwijnen, dan is duidelijk dat dat niet van morgen op vandaag zal gebeuren. Die traagheid kan te wijten zijn aan een psychologisch mechanisme als gewoontevorming of de neiging om alles bij het oude te willen houden. Ze is daartoe echter niet te herleiden. Er is ook het feit dat het reilen en zeilen van een niet onbeduidend aantal mensen en bedrijven momenteel afhangt van het gebruik van auto's, dat onze huidige vorm van samenleven sterk is verweven met die transporttechnologie. Een punt zetten achter het gebruik van een technologie zou dus op verschillende vlakken ingrijpende gevolgen hebben, en dat alleen al is een reden om te vermoeden dat het verlangen van de activist naar verandering minder snel werkelijkheid zal worden dan gewenst.

Samenvattend onderstreep ik graag nog dit. Natuurlijk is een sociaal constructivistische blik op technologie relevant, niet alleen om het determinisme aan te vallen, maar ook op praktisch vlak. Tegelijk zijn er ook redenen om het praktische belang van die benadering niet te overschatten. Uiteraard wil ik niet zeggen dat verandering onwenselijk is en dat verandering, zelfs als ze wenselijk is, er toch niet zal komen. Wel heb ik in de laatste paragrafen de aandacht willen vestigen op

zaken die de verandering op het vlak van technologie, zelfs als ze mogelijk is, kunnen vertragen of zelfs verhinderen. Dat moet je niet interpreteren als een uiting van conservatisme noch als een afkeer van activisme, maar als een expressie van realiteitszin.

Ter afronding

Uiteraard spreken niet alle ingenieurs en ondernemers in termen van determineren wanneer het over technologie en AI gaat, maar dat is wel zo voor een niet gering aantal ingenieurs en ondernemers. Het is belangrijk om te weten dat je in die context vier soorten beweringen kunt onderscheiden. Vooreerst is er de theorie, verdedigd door onder anderen Heidegger, dat we vandaag de dag enkel nog in instrumentele zin naar de wereld kunnen kijken. De tweede versie van de determinismethese luidt dat technologie op sociaal vlak onvermijdelijk sociale effecten heeft. De derde en vierde variant gaan over de ontwikkelingsgeschiedenis van geschiedenis, dat wil zeggen: over het ontstaan van technologie en over de evolutie van technologie. In dit hoofdstuk heb ik die vier stellingen belicht en zo helder als mogelijk uitgelegd. Daarnaast heb ik ook telkens een evaluatieve blik geworpen op die vier versies van het technologisch determinisme, met telkens twee vragen voor ogen. Volstaan de argumenten die worden gegeven om het technologisch determinisme te ondersteunen? Zo niet, zijn er redenen om de stelling te verwerpen? Die tweede vraag is niet onbelangrijk. Als immers blijkt dat een argument niet het werk doet dat het moet doen, dan kun je daaruit niet afleiden dat de stelling onjuist is, daar er andere beslissende argumenten kunnen zijn. Wanneer het nu over technologisch determinisme gaat, is mijn besluit dit. Voor geen van de vier versies volstaan de gegeven argumenten. Sterker nog, iedere versie kan door minstens één argument onderuit worden gehaald. Toegepast op de determinismethese over het ontstaan van technologie: gelijktijdige evolutie toont niet aan dat het ontstaan van een technologie onvermijdelijk is en bovendien kon de

planningssoftware van Starbucks er evengoed niet geweest zijn. Wil dat laatste nu zeggen dat geen enkele technologie onvermijdelijk is? Nee. Niet elke technologie is onvermijdelijk, dat is wat ik aan de hand van de software van Starbucks beweer, en dat is niet gelijk aan de bewering dat geen enkele technologie onvermijdelijk is.

De ontwikkeling van volledige AI zou het einde van de menselijke soort kunnen inluiden.

Stephen Hawking in een interview met de BBC in 2014

Nawoord

Het einde der tijden

Ik eindig met een apocalyptisch scenario. Wanneer het over technologie en AI gaat, wordt vaak het beeld opgeroepen van superintelligente entiteiten die de menselijke vermogens en beperkingen verregaand overstijgen. In zekere zin is dat begrijpelijk, want het kan naast prikkelend ook interessant zijn daarover te speculeren. Aan de andere kant is dat opmerkelijk. Zulke entiteiten bestaan momenteel niet en het ziet er niet naar uit dat die in de nabije toekomst zullen bestaan, laat staan dat ze ooit zullen bestaan. Bovendien is er een andere vorm van slimme technologie – *narrow AI* – die wel al bestaat en die nu al tal van morele problemen met zich meebrengt – ik had het over onder meer bias en veiligheid. Ook opmerkelijk is dat aan zulke *superintelligence* vaak een erg specifieke invulling wordt gegeven. Men omschrijft die AI niet zelden als een kwaadaardig systeem dat de mensheid kan en zal vernietigen. Dat springt in het oog omdat er een tegengestelde interpretatie van *superintelligence* mogelijk is. Je kunt zulke hyperintelligente entiteiten namelijk ook zien als morele systemen die de mensheid willen redden, helpen en verbeteren, bijvoorbeeld door iedereen boven de armoedegrens te tillen, door alle ziekten en lijden de wereld uit te helpen of door de hongersnood op te lossen. Vanwaar die tendens om de voorstelling van superintelligente technologie in de richting van het kwaad te duwen? Daarnaast bestaat echter nog een derde optie, die zich tussen de eerste en tweede bevindt. Het gaat over superintelligente systemen die, net zoals in het tweede geval, een moreel goed karakter hebben, maar die, net zoals in het eerste geval, de mensheid willen en zullen uitroeien. Beter nog, het gaat nu over machines met een moreel karakter die precies *daarom* de menselijke soort willen vernietigen.

Dat lijkt misschien op het eerste gezicht tegenstrijdig maar dat is bij nader inzien niet het geval. Het hangt af van hoe breed of hoe smal je ethiek in deze context opvat. Vertrek je van een smalle invulling – een interpretatie die focust op de menselijke soort – dan is het moeilijk of zelfs onmogelijk om dat te verzoenen met de extinctie van onze soort. Hanteer je een brede invulling – vertrek je van een blik op het totale ecosysteem waarvan de mens slechts een onderdeel is – dan kun je dat in overeenstemming brengen met het einde van de mens, die soort die in sterke mate verantwoordelijk is voor de grote onwenselijke effecten op het ecosysteem. *Superintelligence* die bekommerd is om de planeet, om het voorbestaan van nietmenselijk leven, om biodiversiteit en het ecologisch evenwicht, wel, zulk systeem kan ernaar streven om haar morele doelen te realiseren, niet door de mensheid te redden maar door die integendeel uit te roeien. De dood van de mens die het ecosysteem nieuw leven in blaast, zie hier de grondgedachte van een scenario dat in film en fictie nauwelijks aan bod komt.

Eindnoten

1 Soper, S. (2021). Fired by Bot at Amazon: ‘It’s You Against the Machine’. *Bloomberg* (June 28, 2021). https://www.bloomberg.com/news/features/202106-28/fired-by-bot-amazon-turns-to-machine-managers-and-workers-arelosing-out.

2 Criado Perez, C. (2019). *Onzichtbare vrouwen. Waarom we leven in een wereld voor en door mannen ontworpen*. (Vert. Heuvelmans, T., de Jong, S., Molegraaf, M. & Ridder, S.) Amsterdam: Prometheus, pp. 189-201.

3 Zuidema, T. (2020). Computer strijdt mee tegen corona. *Eos Wetenschap Special*, pp. 66-69.

4 Basalla, G. (1993). *Geschiedenis van de technologie*. (Vert. Kolthoff, B.) Utrecht: Uitgeverij Het Spectrum, pp. 105-108.

5 Trafton, A. (2020). Artificial intelligence yields new antibiotic. *MIT News* (February 20, 2020). https://news.mit.edu/2020/artificial-intelligence-identifiesnew-antibiotic-0220.

6 Crawford, K. (2021). *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven and London: Yale University Press, pp. 3641.

7 Schor, J. B. (2020). *After the Gig. How the Sharing Economy Got Hijacked and How to Win It Back*. Oakland: University of California Press, pp. 105-121.

8 Auteur onbekend (2021). Twee op de drie Vlamingen is nog onbekend met AI. *Knack* (18 juni 2021). https://datanews.knack.be/ict/nieuws/twee-op-de-drievlamingen-is-nog-onbekend-met-ai/article-news-1748229.html.

9 Marx, K. (2010). *Het kapitaal. Kritiek van de politieke economie*. (Vert. Lipschits, I.) Amsterdam: Boom, p. 349.

10 De volgende paragrafen zijn gebaseerd op mijn lectuur van Bartoletti, I. (2020). *An Artificial Revolution. On Power, Politics and AI*. London: The Indigo Press; Lambert, D. (2019). *La robotique et l’intelligence artificielle*. Namur: Éditions jésuites, pp. 9-51; Russell, S. (2019). *Human Compatible. Artificial Intelligence and the Problem of Control*. New York: Viking; Belpaeme, T. (2019). *Artificial Intelligence. Op weg naar de artificiele mens?* Gent: Academia Press; Dignum, V. (2019). *Responsible Artificial Intelligence. How to Develop and Use AI in a Responsible Way*. Dordrecht: Springer; Tegmark, M. (2017). *Life 3.0. Mens zijn in het tijdperk van kunstmatige intelligentie*. (Vert. Paalman, W. & van der Waa, F.) Amsterdam: Maven Publishing.

11 Voor wie meer over deze vragen wil lezen, kan ik de studie *Prediction*

Machines. The Simple Economics of Artificial Intelligence uit 2018 van Agrawal en Goldfarb aanbevelen. Zie ook Ford, M. (2016). *De opmars van robots. Hoe technologie veel banen zal doen verdwijnen*. (Vert. Blankestijn, M.) Amsterdam/Antwerpen: Uitgeverij Q.

12 Voor meer hierover, zie Gunkel, D. J. (2018). *Robot Rights*. Cambridge/London: The MIT Press.

13 Mijn antwoord op die vraag is gebaseerd op Adriaens, P. (2017). *Denken over leven. Wijsbegeerte voor bio-ingenieurs*. Kalmthout: Pelckmans Pro, pp. 1333.

14 Heidegger, M. (2014). *De vraag naar de techniek*. (Vert. Wildschut, M.) Nijmegen: Uitgeverij Vantilt, p. 7. (mijn cursivering)

15 Pitt, J.C. (2014). “Guns Don’t Kill, People Kill”; Values in and/or Around Technologies. In: Kroes, P. & Verbeek, P.-P. (eds.) (2014). *The Moral Status of Technical Artefacts*. Dordrecht: Springer, pp. 89-101.

16 Fry, H. (2018). *Algoritmes aan de macht. Hoe blijf je menselijk in een geautomatiseerde wereld?* (Vert. Jonkers, J.) Amsterdam: De Geus, p. 15.

17 Geciteerd in Veletsianos, G. (2014). On Noam Chomsky and technology’s neutrality. (January 23, 2014). https://www.veletsianos.com/2014/01/23/onnoam-chomsky-and-technologys-neutrality/.

18 Fry, H. (2018). *Algoritmes aan de macht. Hoe blijf je menselijk in een geautomatiseerde wereld?* (Vert. Jonkers, J.) Amsterdam: De Geus, pp. 8389.

19 Lin, L. Y., Sidani, J. E., Shensa, A., Radovic, A., Miller, E., Colditz, J. B., Primack, B. A. (2016). Association between social media use and depression among U.S. young adults. *Depression and Anxiety 33* (4), pp. 323-331.

20 Lambrecht, A. & Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science 65* (7), pp. 2966-2981.

21 Ik verwijs hier naar het boek *Weapons of Math Destruction* van Cathy O’Neil.

22 Verhoeven, B. (2020). Waarom zijn robots meestal wit? *Eos Wetenschap*. (18 augustus 2020). https://www.eoswetenschap.eu/technologie/waarom-zijnrobots-meestal-wit.

23 Gebaseerd op Whittaker, M., et al. (2019). Disability, Bias, and AI. https://ainowinstitute.org/disabilitybiasai-2019.pdf.

24 Criado Perez, C. (2019). *Onzichtbare vrouwen. Waarom we leven in een wereld voor en door mannen ontworpen*. (Vert. Heuvelmans, T., de Jong, S., Molegraaf, M. & Ridder, S.) Amsterdam: Prometheus, pp. 194-195.

25 De volgende bladzijden zijn gebaseerd op enkele inzichten uit van de Poel, I. & Kroes, P. (2014). Can technology embody values? In: Kroes, P. & Verbeek,

P.-P. (eds.) (2014). *The Moral Status of Technical Artefacts*. Dordrecht: Springer, pp. 103-124.

26 Garber, M. (2013). Funerals for Fallen Robots. *The Atlantic* (September 20, 2013). https://www.theatlantic.com/technology/archive/2013/09/funerals-forfallen-robots/279861/.

27 Srnicek, N. (2017). *Platform Capitalism*. Cambridge/Malden: Polity Press, pp. 9-13.

28 Dit voorbeeld ontleen ik aan Gabriels, K. (2019). *Regels voor robots. Ethiek in tijden van AI*. Brussel: Uitgeverij VUBPRESS, pp. 9-10.

29 Zie ook Miller, B. (2021). Is Technology Value-Neutral? *Science, Technology, & Human Values 46* (1), pp. 53-80.

30 Auteur onbekend (2019). Chinese politie houdt gegevens over Oeigoerse moslims bij via app. *Knack* (2 mei 2019). https://www.knack.be/nieuws/belgie/chinese-politie-houdt-gegevens-overoeigoerse-moslims-bij-via-app/article-belga-1459103.html.

31 Voor een uitgebreid verslag, zie Kaiser, B. (2019). *De datadictatuur. Hoe je wordt gemanipuleerd in wat je doet, denkt en stemt*. (Vert. Buesink, S. & Zwart, J.) Amsterdam: HarperCollins.

32 Geerts, T. (2018). *Digitalis. Hoe we onze wereld kunnen heruitvinden*. Tielt: Lannoo, p. 63.

33 Dit is gebaseerd op van de Poel, I. & Kroes, P. (2014). Can technology embody values? In: Kroes, P. & Verbeek, P.-P. (eds.) (2014). *The Moral Status of Technical Artefacts*. Dordrecht: Springer, p. 117.

34 Dit voorbeeld is gebaseerd op Warnier, M., Dechesne, F. & Brazier, F. (2015). Design for the Value of Privacy. In: van den Hoven, J., Vermaas, P. & van de Poel, I. (eds.) (2015). *Handbook of Ethics, Values, and Technological Design*. Dordrecht: Springer, pp. 1-14.

35 Winner, L. (1980). Do Artifacts Have Politics? *Daedalus 109* (1), pp. 121-136.

36 Zie ook Miller, B. (2021). Is Technology Value-Neutral? *Science, Technology, & Human Values* 46 (1), pp. 53-80.

37 De volgende paragrafen zijn gebaseerd op mijn lectuur van Brey, P. (2015). Design for the Value of Human Well-Being. In: van den Hoven, J., Vermaas, P. & van de Poel, I. (eds.) (2015). *Handbook of Ethics, Values, and Technological Design*. Dordrecht: Springer, pp. 1-14.

38 Zie Kearns, M. and Roth, A. (2020). *The Ethical Algorithm. The Science of Socially Aware Algorithm Design*. Oxford: Oxford University Press, pp. 69-72.

39 Pitt, J.C . (2014). “Guns Don’t Kill, People Kill”; Values in and/or Around Technologies. In: Kroes, P. & Verbeek, P.-P. (eds.) (2014). *The Moral Status of*

Technical Artefacts. Dordrecht: Springer, p. 95. (mijn vertaling)

40 De volgende paragrafen zijn voornamelijk gebaseerd op Latour, B. (2016). *Wij zijn nooit modern geweest. Pleidooi voor een symmetrische antropologie*. (Vert. Van Dijk, J. & De Vries, G.) Amsterdam: Boom, pp. 11-28.

41 Zie Savulescu, J. and Maslen, H. (2015). Moral Enhancement and Artificial Intelligence: Moral AI? In: Romportl, J., et al. (eds.) (2015). *Beyond Artificial Intelligence*. Dordrecht: Springer, pp. 79-95.

42 Aral, S. (2020). *The Hype Machine. How Social Media Disrupts Our Elections, Our Economy, and Our Health – and How We Must Adapt*. London: HarperCollinsPublishers, pp. 36-39.

43 Fisman, R. (2013). Did eBay Just Prove That Paid Search Ads Don't Work? *Harvard Business Review* (March 11, 2013). https://hbr.org/2013/03/did-ebayjust-prove-that-paid.

44 Geerts, T. (2018). *Digitalis. Hoe we onze wereld kunnen heruitvinden*. Tielt: Lannoo, p. 51.

45 Zuboff, S. (2019). *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*. London: Profile Books, p. 50. (mijn vertaling)

46 Voor meer hiervoor, zie Smuha, N. (2019). Artificiële intelligentie bij de overheid. Opportuniteiten en uitdagingen vanuit ethisch-juridisch perspectief. *Vlaams Tijdschrift voor Overheidsmanagement* (4), pp. 43-61.

47 Voor uitvoerige beschrijvingen van casussen als die van Kala en Justin verwijs ik de geïnteresseerde lezer door naar de studie *Ghost Work* uit 2019 van Gray en Suri.

48 Ik ontleen de uitdrukking aan priesterfilosoof Ivan Illich, al moet ik er wel aan toevoegen dat ik er niet precies dezelfde invulling aan geef.

49 Frey, C. B. & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change 114* (issue C), pp. 254-280.

50 Een interessante studie over AI en werk is *Automation and Utopia. Human Flourishing in a World without Work* van John Danaher.

51 Voor een uitgebreide inleiding tot AI-ethiek, zie Coeckelbergh, M. (2020). *AI Ethics*. Cambridge/London: MIT Press; Amoore, L. (2020). *Cloud Ethics. Algorithms and the Attributes of Ourselves and Others*. Durham and London: Duke University Press; Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence*. Dordrecht: Springer; Wallach, W. & Collin, A. (2009). *Moral Machines. Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

52 Moody, O. (2020). German AI posts fake child-abuse videos online to catch abusers. *The Times* (August 13, 2020). https://www.thetimes.co.uk/article/german-ai-posts-fake-child-abuse-videosonline-to-catch-abusers-rq0hc0wxs.

53 Geciteerd in Clark, S. (2014). Artificial intelligence could spell end of human race – Stephen Hawking. *The Guardian* (December 2, 2014). https://www.theguardian.com/science/2014/dec/02/stephen-hawking-intelcommunication-system-astrophysicist-software-predictive-text-type. (mijn vertaling)

54 Bostrom, N. & Yudèkowsky, E. (2014). The Ethics of Artificial Intelligence. In Frankish, K. & Ramsey, W. (eds.) (2014). *Cambridge Handbook of Artificial Intelligence*. New York: Cambridge University Press, p. 330.

55 Gebaseerd op Véliz, C. (2021). *Privacy is Power. Why and How You Should Take Back Control of Your Data*. Brooklyn/London: Melvillehouse, pp. 1-5.

56 Blenner, S. R., et al. (2016). Privacy Policies of Android Diabetes Apps and Sharing of Health Information. *Journal of American Medicine 315* (10), pp. 1051-1052.

57 Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (October 11, 2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insightidUSKCN1MK08G.

58 Chugh, D. (2018). *The Person You Mean to Be. How Good People Fight Bias*. HarperCollins: New York, pp. 44-58.

59 Asimov, I. (1969). *Ik, Robot*. (Vert. Leo Zelders) Utrecht/Antwerpen: Uitgeverij Het Spectrum, p. 6.

60 Algar, C. (2020). New 10-year low in global executions, but progress marred by spikes in a few countries. (April 21, 2020). https://www.amnesty.org/en/latest/news/2020/04/op-ed-new-10-year-low-inglobal-executions-but-progress-marred-by-spikes-in-a-few-countries/.

61 Royakkers, L., van de Poel, I. & Pieters, A. (2004). *Ethiek en techniek. Morele overwegingen in de ingenieurspraktijk*. Baarn: HBuitgevers, p. 200.

62 De Ketelaere, G. M. (2020). *Mens versus machine. Artificiële intelligentie ontrafeld*. Kalmthout: Pelckmans, pp. 93-94.

63 Crawford, K. (2021). *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven and London: Yale University Press, p. 42.

64 De Ketelaere, G. M. (2020). *Mens versus machine. Artificiële intelligentie ontrafeld*. Kalmthout: Pelckmans, pp. 93-94.

65 Belkhit, L. & Elmligi, A. (2018). Assessing ICT Global Emissions Footprint: Trends to 2040 and Recommendations. *Journal of Cleaner Production* 177, pp. 448-463.

66 Voor een goede inleiding op de thematiek van *killer robots*, zie Leveringhaus, A. (2016). *Ethics and Autonomous Weapons*. London: Palgrave Macmillan en Schwarz, E. (2018). *Death Machines. The Ethics of Violent Technologies*. Manchester: Manchester University Press.

67 Cramer, M. (2021). A.I. Drone May Have Acted on Its Own in Attacking Fighters, U.N. Says. *The New York Times* (June 3, 2021). https://www.nytimes.com/2021/06/03/world/africa/libya-drone.html.

68 Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6, pp. 175183.

69 De volgende paragrafen zijn gebaseerd op mijn lectuur van Fischer, J. M. & Ravizza, M. (1998). *Responsibility and Control. A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

70 Jonas, H. (1979). *Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation.* Frankfurt: Suhrkamp Taschenbuch.

71 Royakkers, L., van de Poel, I. & Pieters, A. (2004). *Ethiek en techniek. Morele overwegingen in de ingenieurspraktijk*. Baarn: HBuitgevers, pp. 166-168.

72 Dit experiment is een variant op het bekende gedachte-experiment over de Chinese kamer dat is bedacht door John Searle.

73 Voor een uitstekende filosofische blik op risico's, zie Hansson, S. O. (2013). *The Ethics Of Risk. Ethical Analysis in an Uncertain World*. New York: Palgrave Maxmillan.

74 Norden, E. (1969). The Playboy interview: Marshall McLuhan. A Candid Conversation with the High Priest of Popcult and Metaphysician of Media. *Playboy magazine 16* (3), p. 68. (mijn vertaling)

75 Heidegger, M. (2014). *De vraag naar de techniek*. (Vert. Wildschut, M.) Nijmegen: Uitgeverij Vantilt, p. 27.

76 Heilbroner, R. L. (1994). Do Machines Make History? In: Smith, M. R. & Marx, L. (eds.) (1994). *Does Technology Drive History? The Dilemma of Technological Determinism*. Cambridge/London: The MIT Press, p. 59. (mijn vertaling)

77 Zie Smith, M. R. (1994). Technological Determinism in American Culture. In: Smith, M. R. & Marx, L. (eds.) (1994). *Does Technology Drive History? The Dilemma of Technological Determinism*. Cambridge/ London: The MIT Press, pp. 2-35.

78 Geciteerd in Heilbroner, R. L. (1994). Do Machines Make History? In: Smith, M. R. & Marx, L. (eds.) (1994). *Does Technology Drive History? The Dilemma of Technological Determinism*. Cambridge/London: The MIT Press, p. 55. (mijn vertaling)

79 Geciteerd in Pitt, J. (1987). The Autonomy of Technology. In: Durbin, P.T. (ed.) (1987). *Technology and Responsibility. Philosophy and Technology*. Dordrecht: Springer, p. 99. (mijn vertaling)

80 Bijker, W. E. (1995). *Of bicycles, bakelites, and bulbs: toward a theory of sociotechnical change*, Cambridge: MIT Press.

81 Pitt, J. (1999). *Technological Determinism. Foundations of the Philosophy of Technology*. New York, Seven Bridges Press, pp. 84-87.

82 De komende paragrafen zijn uitsluitend gebaseerd op Heideggers tekst *Die Frage nach der Technik*. Ik laat buiten beschouwing wat Heidegger in andere teksten over deze thematiek heeft geschreven. Mijn samenvatting en analyse van Heideggers tekst is beïnvloed door de uitstekende inleiding op Heideggers tekst van de hand van Peter-Paul Verbeek in *De daadkracht der dingen* en Andrew Feenberg *Questioning Technology*.

83 Voor een hedendaagse variant van deze invalshoek, zie Couldry, N. & Mejias U. A. (2019). *The Costs of Connection. How Data IS Colonizing Human Life and Appropriating It for Capitalism*. Stanford: Stanford University Press.

84 Ik ontleen de uitdrukking ‘inforgs’ aan Luciano Floridi.

85 Voor een kritische studie over sociale media, zie Vaidhyanathan, S. (2018). *Antisocial Media. How Facebook Disconnects Us and Undermines Democracy*. Oxford: Oxford University Press.

86 Voor meer hierover, zie het goed gedocumenteerde werk *Big Tech. Hoe we onze privacy, vrije markt en democratie in de uitverkoop doen* van Foroohar.

87 Petzet, H. W. (1993). *Encounters and Dialogues with Martin Heidegger: 19291976.* (Trans. Emad, P. and Maly, K.) Chicago: University of Chicago Press, pp. 149-150, p. 210.

88 De volgende bladzijden vloeien voort uit mijn lectuur van het derde hoofdstuk uit *Beyond Humanity?* van Allen Buchanan.

89 Claussen, J., Peukert, C. & Sen, A. (2019). The Editor vs. the Algorithm: Returns to Data and Externalities in Online News. *CESifo Working Paper Series*.

90 Zie bijvoorbeeld Levy, R. (2021). Social Media, News Consumption, and Polarization: Evidence from a Field Experiment. *American Economic Review* 111 (3), pp. 831-870.

91 Zie ook Agar, N. (2015). *The Sceptical Optimist. Why Technology Isn’t the Answer to Everything*. Oxford: Oxford University Press, p. 44-45.

92 Heaven, W. D. (2020). Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review* (July 17, 2020). https://www.technologyreview.com/2020/07/17/1005396/predictive-policingalgorithms-racist-dismantled-machine-learning-bias-criminal-justice/.

93 Smit, W. A. & van Oost, E. C. J. (1999). *De wederzijdse beïnvloeding van technologie en maatschappij. Een Technology Assessment-benadering*. Bussum: uitgeverij Coutinho, pp. 25-53.

94 Smit, W. A. & van Oost, E. C. J. (1999). *De wederzijdse beïnvloeding van technologie en maatschappij. Een Technology Assessment-benadering*. Bussum: uitgeverij Coutinho, pp. 27-29.

95 Basalla, G. (1993). *Geschiedenis van de technologie*. (Vert. Kolthoff, B.) Utrecht: Uitgeverij Het Spectrum, pp. 69-76.

96 Zie ook Arthur, W. B. (2011). *The Nature of Technology. What It Is and How It Evolves*. New York/London/Toronta/Syndey: Free Press.

97 Basalla, G. (1993). *Geschiedenis van de technologie*. (Vert. Kolthoff, B.) Utrecht: Uitgeverij Het Spectrum, pp. 18-24.

98 Kelly, K. (2011). *De Wil van Technologie*. (Vert. Grootveld, M., Tuiter, P. & Scherpenisse, W.) Amsterdam: Maven Publishing, pp. 130-160.

99 Kelly, K. (2011). *De Wil van Technologie*. (Vert. Grootveld, M., Tuiter, P. & Scherpenisse, W.) Amsterdam: Maven Publishing, pp. 162-190.

100 Geciteerd in Zuboff, S. (2019). *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*. London: Profile Books, p. 222.

101 Smit, W. A. & van Oost, E. C. J. (1999). *De wederzijdse beïnvloeding van technologie en maatschappij. Een Technology Assessment-benadering*. Bussum: uitgeverij Coutinho, pp. 79-81.

102 Basalla, G. (1993). *Geschiedenis van de technologie*. (Vert. Kolthoff, B.) Utrecht: Uitgeverij Het Spectrum, pp. 53-58.

103 Basalla, G. (1993). *Geschiedenis van de technologie*. (Vert. Kolthoff, B.) Utrecht: Uitgeverij Het Spectrum, pp 36-37.

104 Basalla, G. (1993). *Geschiedenis van de technologie*. (Vert. Kolthoff, B.) Utrecht: Uitgeverij Het Spectrum, pp. 48-53.

105 Kelly, K. (2011). *De Wil van Technologie*. (Vert. Grootveld, M., Tuiter, P. & Scherpenisse, W.) Amsterdam: Maven Publishing, pp. 192-211.

106 Dit is een variatie op het bekende gedachte-experiment van Harry Frankfurt.

107 Dit deel is beïnvloed door mijn lectuur van het zesde hoofdstuk uit *Kernthema's in de hedendaagse technische wetenschap* van onder anderen Pieter Vermaas.

108 Kantor, J. (2014). Working Anything but 9 to 5. *The New York Times* (August 13, 2014). https://www.nytimes.com/interactive/2014/08/13/us/starbucksworkers-scheduling-hours.html?mtrref=www.google.be&gwh=0A6989B77130BE-779940400A811F5F43&gwt=pay&asset-Type=PAYWALL.

109 Basalla, G. (1993). *Geschiedenis van de technologie*. (Vert. Kolthoff, B.) Utrecht: Uitgeverij Het Spectrum, pp. 145-146.

110 Zie bijvoorbeeld Eynon, R. & Young, E. (2021). Methodology, Legend, and Thetoric: The Constructions of AI by Academia, and Policy Groups for Lifelong Learning, *Science, Technology, and Human Values 46* (1), pp. 166-191.

111 Hacking, I. (1999). *The Social Construction of What?* Cambridge/London: Harvard University Press, p. 1.

112 Smit, W. A. & van Oost, E. C. J. (1999). *De wederzijdse beïnvloeding van technologie en maatschappij. Een Technology Assessment-benadering*. Bussum: uitgeverij Coutinho, pp. 100-102.

Bibliografie

Adriaens, P. (2017). *Denken over leven. Wijsbegeerte voor bio-ingenieurs*. Kalmthout: Pelckmans Pro.

Agar, N. (2015). *The Sceptical Optimist. Why Technology Isn't the Answer to Everything*. Oxford: Oxford University Press.

Agrawal, A., Gans, J. & Goldfarb, A. (2018). *Prediction Machines. The Simple Economics of Artificial Intelligence*. Boston: Harvard Business Review Press.

Algar, C. (2020). New 10-year low in global executions, but progress marred by spikes in a few countries. (April 21, 2020). https://www.amnesty.org/en/latest/news/2020/04/op-ed-new-10-year-low-in-globalexecutions-but-progress-marred-by-spikes-in-a-few-countries/.

Amoore, L. (2020). *Cloud Ethics. Algorithms and the Attributes of Ourselves and Others*. Durham and London: Duke University Press.

Aral, S. (2020). *The Hype Machine. How Social Media Disrupts Our Elections, Our Economy, and Our Health – and How We Must Adapt*. London: HarperCollinsPublishers.

Arthur, W. B. (2011). *The Nature of Technology. What It Is and How It Evolves*. New York/ London/Toronto/Syndey: Free Press.

Asimov, I. (1969). *Ik, Robot*. (Vert. Leo Zelders) Utrecht/Antwerpen: Uitgeverij Het Spectrum.

Auteur onbekend (2021). Twee op de drie Vlamingen is nog onbekend met AI. *Knack* (18 juni 2021). https://datanews.knack.be/ict/nieuws/twee-op-de-drievlamingen-is-nog-onbekend-met-ai/article-news-1748229.html.

Auteur onbekend (2019). Chinese politie houdt gegevens over Oeigoerse moslims bij via app. *Knack* (2 mei 2019). https://www.knack.be/nieuws/belgie/chinesepolitie-houdt-gegevens-over-oeigoerse-moslims-bij-via-app/article-belga1459103.html.

Bartoletti, I. (2020). *An Artificial Revolution. On Power, Politics and AI*. London: The Indigo Press.

Basalla, G. (1993). *Geschiedenis van de technologie*. (Vert. Kolthoff, B.) Utrecht: Uitgeverij Het Spectrum.

Belkhit, L. & Elmligi, A. (2018). Assessing ICT Global Emissions Footprint: Trends to 2040 and Recommendations. *Journal of Cleaner Production* 177, pp. 448-463.

Belpaeme, T. (2019). *Artificial Intelligence. Op weg naar de artificiële mens?* Gent: Academia Press.

Bijker, Wiebe E. (1995). *Of bicycles, bakelites, and bulbs: toward a theory of sociotechnical change*, Cambridge: MIT Press.

Blenner, S. R., et al. (2016). Privacy Policies of Android Diabetes Apps and Sharing of Health Information. *Journal of American Medicine 315* (10), pp. 10511052.

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

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

Bostrom, N. & Yudèkowsky, E. (2014). The Ethics of Artificial Intelligence. In Frankish, K. & Ramsey, W. (eds.) (2014). *Cambridge Handbook of Artificial Intelligence*. New York: Cambridge University Press, pp. 316-334.

Brey, P. (2015). Design for the Value of Human Well-Being. In: van den Hoven, J., Vermaas, P. & van de Poel, I. (eds.) (2015). *Handbook of Ethics, Values, and Technological Design*. Dordrecht: Springer, pp. 1-14.

Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems 30* (1), pp. 107-117.

Buchanan, A. (2011). *Beyond Humanity?* Oxford: Oxford University Press.

Bunge, M. (1966). Technology As Applied Science, *Technology and Culture 7* (3), pp. 329-347.

Butler, S. (1863). Darwin among the Machines. In: Jones, J. H. and Bartholomew, A. T. (eds.) (1923). *A First Year in Canterbury Settlement and Other Early Essays. Shrewsbury Edition of the Works of Samuel Butler*. London: Cape, pp. 179-80.

Chugh, D. (2018). *The Person You Mean to Be. How Good People Fight Bias*. HarperCollins: New York.

Clark, S. (2014). Artificial intelligence could spell end of human race – Stephen Hawking. *The Guardian* (December 2, 2014).

https://www.theguardian.com/science/2014/dec/02/stephen-hawking-intelcommunication-system-astrophysicist-software-predictive-text-type.

Claussen, J., Peukert, C. & Sen, A. (2019). The Editor vs. the Algorithm: Returns to Data and Externalities in Online News. *CESifo Working Paper Series*.

Coeckelbergh, M. (2020). *AI Ethics*. Cambridge/ London: MIT Press.

Couldry, N. & Mejias U. A. (2019). *The Costs of Connection. How Data IS Colonizing Human Life and Appropriating It for Capitalism*. Stanford: Stanford University Press.

Cramer, M. (2021). A.I. Drone May Have Acted on Its Own in Attacking Fighters, U.N. Says. *The New York Times* (June 3, 2021). https://www.nytimes.com/2021/06/03/world/africa/libya-drone.html.

Crawford, K. (2021). *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven and London: Yale University Press.

Criado Perez, C. (2019). *Onzichtbare vrouwen. Waarom we leven in een wereld voor en door mannen ontworpen*. (Vert. Heuvelmans, T., de Jong, S., Molegraaf, M. & Ridder, S.) Amsterdam: Prometheus.

Danaher, J. (2019). *Automation and Utopia. Human Flourishing in a World without Work*. Cambridge/London: Harvard University Press.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (October 11, 2018). https://www.reuters.com/article/us-amazoncom-jobs-automation-insight-idUSKCN1MK08G.

De Ketelaere, G. M. (2020). *Mens versus machine. Artificiële intelligentie ontrafeld*. Kalmthout: Pelckmans.

Dignum, V. (2019). *Responsible Artificial Intelligence. How to Develop and Use AI in a Responsible Way*. Dordrecht: Springer.

Eynon, R. & Young, E. (2021). Methodology, Legend, and Thetoric: The Constructions of AI by Academia, and Policy Groups for Lifelong Learning, *Science, Technology, and Human Values 46* (1), pp. 166-191.

Feenberg, A. (1999). *Questioning Technology*. London/New York: Routledge.

Fischer, J. M. & Ravizza, M. (1998). *Responsibility and Control. A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Fisman, R. (2013). Did eBay Just Prove That Paid Search Ads Don't Work? *Harvard Business Review* (March 11, 2013). https://hbr.org/2013/03/did-ebay-justprove-that-paid.

Ford, M. (2016). *De opmars van robots. Hoe technologie veel banen zal doen verdwijnen*. (Vert. Blankestijn, M.) Amsterdam/Antwerpen: Uitgeverij Q.

Foroohar, R. (2020). *Big Tech. Hoe we onze privacy, vrije markt en democratie in de uitverkoop doen*. (Vert. Matthews, S.) Rotterdam: Lemniscaat.

Frey, C. B. & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change 114* (issue C), pp. 254-280.

Fry, H. (2018). *Algoritmes aan de macht. Hoe blijf je menselijk in een geautomatiseerde wereld?* (Vert. Jonkers, J.) Amsterdam: De Geus.

Gabriels, K. (2019). *Regels voor robots. Ethiek in tijden van AI*. Brussel: Uitgeverij VUBPRESS.

Garber, M. (2013). Funerals for Fallen Robots. *The Atlantic* (September 20, 2013). https://www.theatlantic.com/technology/archive/2013/09/funerals-for-fallenrobots/279861/.

Geerts, T. (2018). *Digitalis. Hoe we onze wereld kunnen heruitvinden*. Tielt: Lannoo.

Gray, M. L. & Suri, S. (2019). *Ghost Work. How to Stop Silicon Valley from Building a New Global Underclass*. Boson/New York: Houghton Mifflin Harcourt.

Gunkel, D. J. (2018). *Robot Rights*. Cambridge/ London: The MIT Press.

Hacking, I. (1999). *The Social Construction of What?* Cambridge/London: Harvard University Press.

Hansson, S. O. (2013). *The Ethics Of Risk. Ethical Analysis in an Uncertain World*. New York: Palgrave Maxmillan.

Heaven, W. D. (2020). Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review* (July 17, 2020). https://www.technologyreview.com/2020/07/17/1005396/predictive-policingalgorithms-racist-dismantled-machine-learning-bias-criminal-justice/.

Heidegger, M. (2014). *De vraag naar de techniek*. (Vert. Wildschut, M.) Nijmegen: Uitgeverij Vantilt.

Heilbroner, R. L. (1994). Do Machines Make History? In: Smith, M. R. & Marx, L. (eds.) (1994). *Does Technology Drive History? The Dilemma of Technological Determinism*. Cambridge/London: The MIT Press.

Jonas, H. (1979). *Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation.* Frankfurt: Suhrkamp Taschenbuch.

Kaiser, B. (2019). *De datadictatuur. Hoe je wordt gemanipuleerd in wat je doet, denkt en stemt*. (Vert. Buesink, S. & Zwart, J.) Amsterdam: HarperCollins.

Kantor, J. (2014). Working Anything but 9 to 5. *The New York Times* (August 13, 2014). https://www.nytimes.com/interactive/2014/08/13/us/starbucks-workersscheduling-hours.html?mtrref=www.google.be&gwh=0A6989B77130BE779940400A811F5F43&gwt=pay&as-setType=PAYWALL.

Kapp, E. (1877). *Grundlinien einer Philosophie der Technik. Zur Entstehungsgeschichte der Cultur aus neuen Gesichtspunkten*. Georg Westermann: Braunschweig.

Kearns, M. & Roth, A. (2020). *The Ethical Algorithm. The Science of Socially Aware Algorithm Design*. Oxford: Oxford University Press.

Kelly, K. (2011). *De Wil van Technologie*. (Vert. Grootveld, M., Tuiter, P. & Scherpenisse, W.) Amsterdam: Maven Publishing.

Lambert, D. (2019). *La robotique et l'intelligence artificielle*. Namur: Éditions jésuites.

Lambrecht, A. & Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science 65* (7), pp. 2966-2981.

Latour, B. (2016). *Wij zijn nooit modern geweest. Pleidooi voor een symmetrische antropologie*. (Vert. Van Dijk, J. & De Vries, G.) Amsterdam: Boom.

Leveringhaus, A. (2016). *Ethics and Autonomous Weapons*. London: Palgrave Macmillan.

Levy, R. (2021). Social Media, News Consumption, and Polarization: Evidence from a Field Experiment. *American Economic Review* 111 (3), pp. 831-870.

Lin, L. Y., Sidani, J. E., Shensa, A., Radovic, A., Miller, E., Colditz, J. B., Primack, B. A. (2016). Association between social media use and depression among U.S. young adults. *Depression and Anxiety 33* (4), pp. 323-331.

Marx, K. (2010). *Het kapitaal. Kritiek van de politieke economie*. (Vert. Lipschits, I.) Amsterdam: Boom.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6, pp. 175-183.

Miller, B. (2021). Is Technology Value-Neutral? *Science, Technology, & Human Values 46* (1), pp. 53-80.

Moody, O. (2020). German AI posts fake child-abuse videos online to catch abusers. *The Times* (August 13, 2020). https://www.thetimes.co.uk/article/germanai-posts-fake-child-abuse-videos-online-to-catch-abusers-rq0hc0wxs.

Moore, G. E. (1965). Cramming More Components onto Integrated Circuits. *Electronics 34* (8), pp. 114-117.

Norden, E. (1969). The Playboy interview: Marshall McLuhan. A Candid Conversation with the High Priest of Popcult and Metaphysician of Media. *Playboy magazine 16* (3), pp. 53-74.

O'Neil, C. (2016). *Weapons of Math Destruction. How Big Data Increased Inequality and Threatens Democracy*. New York: Crown.

Petzet, H. W. (1993). *Encounters and Dialogues with Martin Heidegger: 19291976*. (Trans. Emad, P. and Maly, K.) Chicago: University of Chicago Press.

Pitt, J. (1987). The Autonomy of Technology. In: Durbin, P.T. (ed.) (1987). *Technology and Responsibility. Philosophy and Technology*. Dordrecht: Springer, pp. 99-114.

Pitt, J. (1999). *Technological Determinism. Foundations of the Philosophy of Technology*. New York, Seven Bridges Press.

Pitt, J.C. (2014). "Guns Don't Kill, People Kill"; Values in and/or Around Technologies. In: Kroes, P. & Verbeek, P.-P. (eds.) (2014). *The Moral Status of Technical Artefacts*. Dordrecht: Springer, pp. 89-101.

Rosen, L D., Cheever N. A., and Carrier, L. M. (2015). *The Wiley Handbook of Psychology, Technology, and Society*. Malden/Oxford: Wiley Blackwell.

Royakkers, L., van de Poel, I. & Pieters, A. (2004). *Ethiek en techniek. Morele overwegingen in de ingenieurspraktijk*. Baarn: HBuitgevers.

Russell, S. (2019). *Human Compatible. Artificial Intelligence and the Problem of Control*. New York: Viking.

Savulescu, J. and Maslen, H. (2015). Moral Enhancement and Artificial Intelligence: Moral AI? In: Romportl, J., et al. (eds.) (2015). *Beyond Artificial Intelligence*. Dordrecht: Springer, pp. 79-95.

Schor, J. B. (2020). *After the Gig. How the Sharing Economy Got Hijacked and How to Win It Back*. Oakland: University of California Press.

Schwarz, E. (2018). *Death Machines. The Ethics of Violent Technologies*. Manchester: Manchester University Press.

Smit, W. A. & van Oost, E. C. J. (1999). *De wederzijdse beïnvloeding van technologie en maatschappij. Een Technology Assessment-benadering*. Bussum: uitgeverij Coutinho.

Smith, M. R. (1994). Technological Determinism in American Culture. In: Smith, M. R. & Marx, L. (eds.) (1994). *Does Technology Drive History? The Dilemma of Technological Determinism*. Cambridge/London: The MIT Press, pp. 2-35.

Smuha, N. (2019). Artificiële intelligentie bij de overheid. Opportuniteiten en uitdagingen vanuit ethisch-juridisch perspectief. *Vlaams Tijdschrift voor Overheidsmanagement* (4), pp. 43-61.

Soper, S. (2021). 'Fired by Bot at Amazon: "It's You Against the Machine"'. *Bloomberg* (June 28, 2021). https://www.bloomberg.com/news/features/2021-0628/fired-by-bot-amazon-turns-to-machine-managers-and-workers-are-losing-out.

Srnicek, N. (2017). *Platform Capitalism*. Cambridge/Malden: Polity Press.

Stiegler, B. (2019). *The Age of Disruption. Technology and Madness in Computational Capitalism*. London: Polity Press.

Tegmark, M. (2017). *Life 3.0. Mens zijn in het tijdperk van kunstmatige intelligentie*. (Vert. Paalman, W. & van der Waa, F.) Amsterdam: Maven Publishing.

Trafton, A. (2020). Artificial intelligence yields new antibiotic. *MIT News* (February 20, 2020). https://news.mit.edu/2020/artificial-intelligence-identifies-new-antibiotic0220.

Vaidhyanathan, S. (2018). *Antisocial Media. How Facebook Disconnects Us and Undermines Democracy*. Oxford: Oxford University Press.

Van de Poel, I., Fahlquist, J. N., Doorn, N., Zwart, S. & Royakker, L. (2012). 'The Problem of Many Hands: Climate Change as an Example'. *Science and Engineering Ethics* 18, 49-67.

Van de Poel, I. & Kroes, P. (2014). Can technology embody values? In: Kroes, P. & Verbeek, P.-P. (eds.) (2014). *The Moral Status of Technical Artefacts*. Dordrecht: Springer, pp. 103-124.

Van de Poel, I. (2015). 'Values in Engineering and Technology'. In: Gonzalez, W. J. (ed.) (2015). New Perspectives on Technology, Values, and Ethics. Dordrecht: Springer, pp. 29-45.

Veletsianos, G. (2014). On Noam Chomsky and technology's neutrality. (January 23, 2014). https://www.veletsianos.com/2014/01/23/on-noam-chomsky-andtechnologys-neutrality/.

Véliz, C. (2021). *Privacy is Power. Why and How You Should Take Back Control of Your Data*. Brooklyn/London: Melvillehouse.

Verbeek, P.-P. (2000). *De daadkracht der dingen. Over techniek, filosofie en vormgeving*. Amsterdam: Boom.

Verhoeven, B. (2020). Waarom zijn robots meestal wit? *Eos Wetenschap*. (18 augustus 2020). https://www.eoswetenschap.eu/technologie/waarom-zijn-robotsmeestal-wit.

Vermaas, P., Kroes, P., van de Poel, I., Franssen, M. & Houkes, W. (2009). *Kernthema's in de technische wetenschap*. Amsterdam: Boom.

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

Warnier, M., Dechesne, F. & Brazier, F. (2015). Design for the Value of Privacy. In: van den Hoven, J., Vermaas, P. & van de Poel, I. (eds.) (2015). *Handbook of Ethics, Values, and Technological Design*. Dordrecht: Springer, pp. 1-14.

Whittaker, M., et al. (2019). Disability, Bias, and AI. *https://ainowinstitute.org/disabilitybiasai-2019.pdf*.

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus 109* (1), pp. 121-136.

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

Zuidema, T. (2020). Computer strijdt mee tegen corona. *Eos Wetenschap Special*, pp. 66-69.

Dankwoord

In de eerste plaats veel dank aan Mauritz Kelchtermans die alle teksten van commentaar heeft voorzien. Verder dank ik graag Hilde Bonte, Tim Christiaens, Mark Coeckelbergh, Paul Moyaert, Benjamin De Mesel, Andreas De Block, Paulus Van Bortel, Frank Maet, AnnKatrien Oimann, Marc Van Aken, Martin Meganck, Pieter Adriaens, Olivier Lemeire, Stefan Ramaekers, Laurens Naudts, Massimiliano Simons en Pol Coudeville voor hun feedback op delen van het manuscript, de discussies, het delen van losse gedachten, het aanbevelen van teksten. Charles Derre en Niels Janssens van LannooCampus overzagen het geheel en waren erg geduldig. Mijn grootste dank gaat uit naar Jasmien Hoffelinck. Zonder haar (humor) had ik dit boek niet kunnen schrijven.

D/2021/45/571 – ISBN 978 94 014 8431 2 – NUR 730

Vormgeving omslag: Chloé D’hauwe
Vormgeving binnenwerk: Peer De Maeyer

Uitgeverij LannooCampus maakt deel uit van Lannoo Uitgeverij, de boeken- en multimediadivisie van Uitgeverij Lannoo nv.

Uitgeverij LannooCampus

Vaartkom 41 bus 01.02
3000 Leuven
België
www.lannoocampus.be

Postbus 23202
1100 DS Amsterdam
Nederland
www.lannoocampus.nl

Veel leesplezier!

Een filosofische blik op technologie en artificiële intelligentie

'Lauwaert ontleedt aannames over technologie en AI, geeft voorbeelden en stuurt de sentimenten van de lezer bij: zowel pessimisten als fans van technologie moeten hun standpunten bijschaven. Een boek dat tot alertheid noopt.'

ANNELIES VERBEKE, schrijfster bekroond met de F. Bordewijkprijs en de J.M.A. Biesheuvelprijs

'*Wij, robots* gaat in tegen de verafgoding van technologie, maar ook tegen zij die denken dat technologie alleen onheil brengt. Een oefening in helder redeneren.'

PETER HINSSEN, technologie-ondernemer en docent aan het Massachusetts Institute of Technology (MIT)

'Van hamers tot AI, Lauwaerts boek bespreekt belangrijke thema's die allemaal rond dezelfde kern draaien: techniek is niet neutraal. Dat is een boodschap zowel voor ingenieurs als voor ons allemaal. Want technologie, dat zijn wij.'

MARK COECKELBERGH, hoogleraar filosofie aan de Universiteit van Wenen en lid van de High Level Expert Group on Artificial Intelligence van de Europese Commissie

www.lannoocampus.com