Wat hersenonderzoek ons (niet) leert over wie wij zijn

Annemarie van Stee

'Wij zijn ons brein', 'hersenonderzoek bevestigt vooroordelen over mannen en vrouwen', 'ons brein is een open boek', 'de vrije wil bestaat niet' en 'liefde: wat onze hersenen onthullen over de klik, de kus en al het andere'. Hersenonderzoek wordt inmiddels al meer dan een decennium gepresenteerd als bron van zelfkennis. Deze boektitels en krantenkoppen uit de Volkskrant en het NRC zijn zomaar een aantal voorbeelden daarvan. Tegelijkertijd zijn er, zeker de laatste jaren, ook kritische stemmen opgestaan. In de Engelstalige wereld verschenen boeken met woorden als 'Neuromania' en 'Brainwashed' in de titels, die de invloed van neurowetenschap op ons denken over onszelf bekritiseerden. In de Duitstalige wereld verscheen in 2004 *Das Manifest*. Daarin voorspelde een groep neurowetenschappers dat er binnen tien jaar grote stappen gezet zouden worden in het begrijpen van het brein en het wetenschappelijk funderen van een heel nieuw mensbeeld. In 2014 concludeerde een andere groep Duitse neurowetenschappers en filosofen dat er nog bijzonder weinig van de grote claims uit 2004 was uitgekomen.

Wat draagt hersenonderzoek nou wel of niet bij aan zelfinzicht? Ik analyseer neurowetenschappelijk onderzoek naar liefde, naar de vrije wil en naar gedachtenlezen. Op basis daarvan beargumenteer ik dat hersenonderzoek wel degelijk inzicht in onszelf kan opleveren, met name wanneer het relevant is om onszelf als wezens-met-hersenen te beschouwen. Tegelijkertijd stel ik dat dat inzicht niet zo spectaculair is als boektitels en krantenkoppen soms willen doen geloven. Bovendien leg ik uit waar een belangrijk deel van de misverstanden hierover vandaan komen.

Liefde

'Ben jij tot over je oren verliefd? Doe dan mee aan ons experiment!' Deelnemers die zich aanmelden voor een cognitief neurowetenschappelijk onderzoek naar verliefdheid, moeten foto's meenemen van de persoon op wie ze verliefd zijn. Terwijl ze in een fMRI-scanner liggen, kijken ze naar die foto's. Zo wordt de hersenactiviteit gemeten die betrokken is bij de ervaring van verliefdheid. Maar ook de hersenactiviteit die betrokken is bij het kijken naar foto's, bij het herkennen van bekende gezichten en bij het focussen van aandacht. Tussen de foto's van de geliefde door, krijgen deelnemers foto's van collega's of goede vrienden te zien. Daarbij kijken, herkennen en focussen ze ook, maar ervaren ze (als het goed is) geen verliefdheid. Het verschil tussen de hersenactiviteit in de beide gevallen is dan de hersenactiviteit die betrokken is bij het ervaren van verliefdheid.

De resultaten van dit soort 'liefdesexperimenten' zijn niet eenduidig, maar onderzoekers vinden bijvoorbeeld activatie in de nucleus caudatus, het ventrale tegmentale gebied en de insula en deactivatie in de amygdala (Van Stee 2017). Dat zegt de gemiddelde Nederlander niet zo veel natuurlijk, en dat is relevant. Zulke resultaten zijn interessant voor neurowetenschappers, omdat die inherent geïnteresseerd zijn in het brein, en in de hersenactiviteit die ons in staat stelt te leven zoals we doen. Leken hebben echter behoorlijk wat basiskennis van de hersenen nodig om chocola te kunnen maken van dit soort beschrijvingen. En dan nog: is 'nucleus caudatus' nou werkelijk het antwoord op onze prangende vragen over onszelf? Kan het überhaupt bijdragen aan zelfinzicht? En wat zou het dan kunnen bijdragen?

Mogelijke medische doeleinden

De eerste bijdrage die hersenonderzoek potentieel kan leveren aan zelfinzicht, is inzicht in de oorzaken van

Annemarie van Stee is filosoof aan de Radboud Universiteit Nijmegen. Ze behaalde cum laude masters in de filosofie en de cognitieve neurowetenschap en promoveerde in Leiden op het proefschrift *Understanding existential self-understanding: philosophy meets cognitive neuroscience*. Van Stee spreekt en schrijft ook graag voor een breed publiek: www.avanstee.nl .

niet-functioneren. Wanneer we niet in staat zijn enige liefde te ervaren voor wie of wat dan ook, dan kan het zijn dat hersenfalen daaraan ten grondslag ligt. De vraag 'waarom ervaar ik geen liefde meer voor mijn man, voor mijn werk, voor mijn kinderen?' kan dan wel degelijk beantwoord worden door te verwijzen naar de gemankeerde nucleus caudatus. Voorlopig is dit een fictief voorbeeld, want we weten nog niet genoeg over de neurale voorwaarden van liefde. Desalniettemin ligt er zeker potentiële maatschappelijke meerwaarde van cognitieve neurowetenschap in de relevantie voor medische doeleinden.

Tegelijkertijd is het belangrijk te begrijpen wat dit wel en niet zegt. Het zijn namelijk niet alleen onze hersenen die ons in staat stellen liefde te ervaren, daarvoor moet ook aan andere voorwaarden zijn voldaan. Een jeugd waarin er van ons werd gehouden, bijvoorbeeld, of waarin er in ieder geval niet uitsluitend sprake was van onverschilligheid en mishandeling. Voorwaarden voor gezond gedrag bevinden zich niet alleen op het niveau van de hersenen, maar ook elders. En zelfs wanneer we kunnen zien dat er een probleem ligt in de hersenen, hoeft dat nog niet te betekenen dat direct ingrijpen in de hersenen door bijvoorbeeld medicatie het gepaste antwoord is. Misschien dat het gebrek aan liefdevolle gevoelens en de slecht functionerende nucleus caudatus beide baat zouden hebben bij acht uur slaap per nacht. Of bij therapie die helpt de depressie waaraan we lijden het hoofd te bieden. Wij zijn nooit alleen maar ons brein, maar ook onze geschiedenis, onze interacties met onze omgeving en nog heel veel meer.

De vrije wil

Claims over de bijdrage van neurowetenschap aan zelfinzicht gaan echter meestal niet over de situatie waarin onze hersenen het niet goed doen. 'De vrije wil bestaat niet' zou voor iedereen moeten opgaan, niet alleen voor diegenen met hersenkwalen. Er is een hele batterij aan experimenten die de hersenactiviteit onderzoekt die betrokken is bij vrijwillige handelingen. Het meest bekende experiment is een experiment dat al uit 1983 stamt, van Benjamin Libet en zijn collega's.

Libet liet de deelnemers aan zijn experiment plaatsnemen bij een scherm waarop een stip als de wijzer van een klok rap in het rond draaide. Hij instrueerde hen om een kleine vinger- of polsbeweging te maken wanneer ze de zin daartoe in zichzelf voelden opkomen (*when they first felt the urge to act*). Ze moesten rapporteren waar op dat moment de wijzer op de klok stond. Ondertussen registreerde een elektromyografie-apparaat (EMG-apparaat) wanneer de beweging precies plaatsvond. Een elektro-encefalografie-apparaat (EEG-apparaat) registreerde de hersenactiviteit. Door dit vaak te herhalen en het gemiddelde te nemen van de verschillende herhalingen, kon Libet een zogenaamd *readiness potential* destilleren uit de hersenactiviteit. De 'bereidheidspotentiaal' is een hersensignaal dat voorafgaat aan alle vrijwillige bewegingen die mensen maken; het wordt geïnterpreteerd als de voorbereiding van de beweging door het brein. Libet ging ervan uit dat de intentie om te bewegen eerst zou opkomen, dat dan de voorbereiding door de hersenen zou starten en dat daarna de daadwerkelijke beweging zou volgen.

Maar wat bleek? Het bereidheidssignaal startte eerst, gemiddeld 550 milliseconden voorafgaand aan de beweging. Pas daarna ervaarden deelnemers de zin om te bewegen, gemiddeld zo'n 200 milliseconden voorafgaand aan de beweging. Omdat oorzaken vooraf moeten gaan aan gevolgen, kan de ervaarde intentie om te bewegen onmogelijk de oorzaak zijn van de bereidheidspotentiaal. En dus is er eigenlijk ook geen grond om aan te nemen dat die intentie de oorzaak is van de beweging zelf, aldus Libet en collega's. Het lijkt erop dat het brein de oorzaak is van de beweging.

De afgelopen vijfendertig jaar is dit onderzoek van boven tot onder aan kritische analyse onderworpen. Niet in het minst door Benjamin Libet zelf overigens, die flink geschrokken was van zijn resultaten. Allerlei potentiële problemen met de experimentele opstelling zijn ondervangen door vervolgexperimenten net iets anders op te zetten. Ik zet zelf nog altijd vraagtekens bij de interpretatie van het signaal dat aan vrijwillige handelingen voorafgaat. Dat het signaal opkomt is een feit, maar dat we het 'bereidheidspotentiaal' noemen is een conventie. Die naam is net zozeer een interpretatie als de aanname dat al bij de start van de potentiaal vastligt welke beweging iemand gaat maken. Recenter onderzoek suggereert dat het niet de beste interpretaties zijn en dat het moment waarop de beweging definitief vaststaat veel dichter bij de daadwerkelijke beweging ligt, zelfs rond het moment dat mensen ervaren dat ze beslissen (Schurger, Sitt en Dehaene 2012).

Hoe dan ook komt uit alle onderzoeken een overkoepelend plaatje naar voren dat tegen de intuïties van veel mensen lijkt in te druisen. Aan iedere vrijwillige handeling gaat een vergelijkbaar hersensignaal vooraf (die zogenaamde bereidheidspotentiaal) en dat signaal start voordat mensen de intentie om te handelen ervaren. De keuze die mensen uiteindelijk maken kan soms zelfs al voorspeld worden op basis van zulke hersenactiviteit.

Bovendien laat geen enkel experiment zien dat er intenties kunnen zijn die *niet* uit onbewuste hersenactiviteit voortkomen (Slors 2012).

Het belang van de hersenen en onbewuste processen onderkennen

Dat brengt ons bij een tweede bijdrage van neurowetenschap aan zelfinzicht: het kan corrigerend werken op ons zelfbeeld waar we impliciet al iets aannemen over het brein. Waar we nog ongemerkt dualistisch denken dat wij niet *ook* ons brein en ons lichaam-met-brein zijn, daar corrigeert neurowetenschap ons bijvoorbeeld. 'Vrij willen' kan niet zonder hersenactiviteit, laat staan dat het losstaand van hersenactiviteit toch hersenactiviteit zou aansturen. Het beeld dat de meeste van onze handelingen vrij zijn doordat ze veroorzaakt worden door een bewuste beslissing die direct aan de handeling vooraf gaat, is ook niet houdbaar. Hoe indrukwekkend je dit soort onderzoek vindt, hangt voor een groot deel af van hoe overtuigd je ervan bent dat mensen zo over 'vrij willen' denken. Benjamin Libet lijkt zelf wel zo'n idee over de vrije wil te hebben gehad.

Algemener gesteld laat dit soort neurowetenschappelijk onderzoek zien hoeveel van ons gedrag tot stand komt zonder dat we er bewust bij nadenken. Voor we ons er goed en wel van bewust zijn, zijn we al in beweging gekomen. We handelen meestal op de automatische piloot. En dat niet alleen, dat automatische handelen van ons kan worden beïnvloed door externe en interne factoren waar we ons niet van bewust zijn. In een ander onderzoek stimuleerden neurowetenschappers de hersenen van mensen wiens schedel was gelicht, omdat ze direct na het onderzoek een hersenoperatie moesten ondergaan. Bij stimulatie van gebieden in de pariëtale cortex, ervaarden deelnemers dat ze hun arm wilden bewegen, of hun lippen, of hun been. Bij krachtigere stimulatie van diezelfde gebieden hadden sommige deelnemers zelfs de indruk dat ze hun arm of lippen of been hadden bewogen. Maar dat was niet zo. Bij stimulatie van een heel ander gebied echter, de premotor cortex, bewogen er wel degelijk ledematen. Maar dat hadden de deelnemers dan weer niet in de gaten (Desmurget et al. 2009). Deelnemers deden ervaringen op, van willen en van bewegen en van niet-bewegen, doordat hun hersenen van buitenaf werden gestimuleerd. En die ervaringen kwamen niet overeen met de werkelijkheid.

Dit sluit goed aan bij psychologisch onderzoek dat aantoont hoe onze keuzes gestuurd worden door zowel externe factoren als automatismen van binnenuit. En dat we ons dit geregeld niet realiseren, maar redenen verzinnen waarom we doen wat we doen. Om een klassiek voorbeeld te nemen: de Amerikaanse onderzoekers Richard Nisbett en Timothy Wilson legden in warenhuizen vier paar identieke sokken op een rij. Voorbijgangers werd gevraagd welk paar ze dachten dat de beste kwaliteit zou hebben. Veruit de meeste voorbijgangers kozen voor het rechterpaar. Het is niet precies duidelijk waarom; wellicht heeft het met rechtshandigheid te maken. Hoe dan ook gaven de voorbijgangers desgevraagd allerlei redenen voor hun keuze, maar nooit hadden ze het over de positionering van de sokken. Wanneer hen expliciet werd gevraagd: 'zou uw keuze ermee te maken kunnen hebben dat deze sokken helemaal rechts lagen?' leidde dat meestal tot een ongeruste blik richting de onderzoekers: 'alsof we niet helemaal goed bij ons hoofd waren', schrijven Nisbett en Wilson (1977). Minder onschuldige voorbeelden vinden we in hedendaags onderzoek naar onbewust racistisch en seksistisch gedrag. Zelfs wanneer mensen er geen expliciete racistische overtuigingen op na lijken te houden, beoordelen ze een ambigu object in de handen van een zwart persoon vaker als een pistool en in de handen van een wit persoon vaker als ongevaarlijk. Zelfs wanneer mensen er geen expliciete seksistische overtuigingen op na lijken te houden, en ook wanneer ze zelf vrouw zijn, beoordelen ze een CV met een mannennaam erboven als competenter en een hoger salaris waard dan wanneer er boven een identiek CV een vrouwennaam staat (Saul 2013).

Zo zien we dat hersenonderzoek, en psychologisch onderzoek, corrigerend kunnen werken op ideeën over onszelf die we al hadden. Alles wat we doen en ervaren gaat gepaard met hersenactiviteit. En geregeld is die hersenactiviteit al in gang gezet voordat we ons bewust realiseren wat we op het punt staan te doen. Een groot deel van ons gedrag ontstaat op de automatische piloot. Daarbij kan het gebeuren dat we beïnvloed worden door omgevingsfactoren en automatische gedragspatronen waar we eigenlijk niet achter staan. We zijn geneigd dit te ontkennen. We zijn geneigd redenen te geven voor onze keuzes die de invloeden van buitenaf en de (al dan niet bevooroordeelde) neigingen van binnenuit verdonkeremanen. Onderzoek naar hersenen en gedrag noopt ons tot bescheidenheid over ons inzicht in de oorzaken van ons gedrag.

Misvattingen corrigeren zonder juiste opvattingen te kunnen poneren

Maar betekent dit nu ook dat de vrije wil niet bestaat? Dat ligt er natuurlijk aan wat we verstaan onder 'de vrije wil'. Als we denken dat onze handelingen vrij zijn doordat ze veroorzaakt worden door bewuste beslissingen die er direct aan voorafgaan, dan klopt dat waarschijnlijk niet. Zelfs al zou de zogenaamde bereidheidspotentiaal anders geïnterpreteerd moeten worden, dan nog doen we ook veel op de automatische piloot. En daar komt geen bewuste beslissing bij kijken. Is zulk automatisch handelen onvrij?

In de filosofische reflectie op ons willen en handelen, draait vrijheid veelal niet om het bewust kiezen uit meerdere opties. Zelfs als we ons maar halfbewust zijn van wat we doen op het moment dat we het doen en zelfs als we niets anders hadden kunnen doen dan wat we doen, kunnen die handelingen nog steeds vrij zijn. Waar het om gaat, aldus een wijdverspreide opvatting, is of we ons onze handelingen kunnen toe-eigenen. Wanneer het *onze* handelingen zijn, zijn ze vrij en dragen we er morele verantwoordelijkheid voor.

Harry Frankfurt is een van de belangrijkste pleitbezorgers van dit idee. Hij werkt het uit door te stellen dat we vrij willen wanneer we ons kunnen identificeren met wat we willen: wanneer we verlangen naar een sigaret en dat verlangen ook beamen, roken we uit vrije wil. Wanneer we daarentegen zouden willen dat we niet rookten, maar toch de zin in een sigaret niet kunnen weerstaan, handelen we niet vrij. In bredere zin kun je stellen dat onze handelingen eigen zijn en daarmee vrij wanneer ze in lijn zijn met onze overkoepelende waarden en levensprojecten. Anderen dan Frankfurt benadrukken op verschillende manieren het vermogen om desgevraagd goede redenen te kunnen aanvoeren voor onze handelingen. Want dat vermogen staat onder druk bij veel mensen die onvrij, want bijvoorbeeld obsessief, gedrag vertonen.

In dit kader is het empirische onderzoek naar onze automatische neigingen ook interessant. Niet omdat dat het bestaan van de vrije wil zou ontkrachten, maar omdat het laat zien dat we lang niet altijd op basis van goede redenen handelen. Inzicht in onbewuste (en eventueel bevooroordeelde) invloeden op ons gedrag kan ons helpen om dat gedrag meer in lijn te laten verlopen met waar we achter kunnen staan. Het kan zo helpen om onze vrijheid te vergroten.

Algemeen gesteld kan hersenonderzoek niethoudbare ideeën over onszelf ontkrachten. Maar hoe we onszelf dan wel moeten begrijpen, blijft een open vraag. Alle visies op onze vrijheid en morele verantwoordelijkheid die ik net aanstipte, blijven ook in het licht van neurowetenschappelijk onderzoek houdbaar. In geen van die visies speelt het brein een cruciale rol en dus draagt neurowetenschap verder niets bij aan het debat hierover.

De bron van heel veel misverstanden. Nogmaals liefde

Cognitief neurowetenschappelijk onderzoek geeft inzicht in de hersenprocessen die ten grondslag liggen aan ons gedrag en onze ervaringen, bijvoorbeeld bij liefde. Krantenkoppen en boektitels beloven echter meer, namelijk dat onze hersenen ook dingen zullen onthullen over liefde zelf: over 'de klik, de kus en al het andere'. Hoe zit dat?

Stel, cognitieve neurowetenschappers vinden uit dat het kijken naar foto's van geliefden, in vergelijking met het kijken naar foto's van collega's, leidt tot activatie in een netwerk van dopamine-rijke gebieden in het midden van het brein. Ze vormen dan hypotheses over waarom ze precies daar activiteit hebben gevonden en rapporteren die aan het einde van hun onderzoek. Eerder onderzoek vond activatie in dat dopamine-netwerk wanneer mensen onder invloed waren van cocaïne, bijvoorbeeld. Wellicht werkt liefde verslavend op het brein, opperen de onderzoekers dan.

Strikt genomen is die gevolgtrekking ongeldig. Het is een gevolgtrekking in de omgekeerde richting namelijk, een *reverse inference*. Dat bij het gebruiken van cocaïne activatie in het dopamine-netwerk optreedt, wil omgekeerd nog niet zeggen dat iedere keer dat het dopaminenetwerk actief is, er ook sprake is van cocaïnegebruik, of van verslavingsprocessen in bredere zin. Het dopaminenetwerk wordt ook actief wanneer mensen iets nieuws zien, iets onverwachts. En het wordt actief wanneer ze worden geconfronteerd met iets dat heel belangrijk voor hen is. Maar ook wanneer iets hen veel stress oplevert (Lammel, Lin en Malenka 2014). En natuurlijk wanneer mensen kijken naar foto's van hun geliefde. Het is daarom volstrekt niet duidelijk wat we uit activatie van het dopamine-netwerk kunnen afleiden over ons gedrag of onze ervaringen.

Onderzoekers presenteren de omgekeerde gevolgtrekking doorgaans als een hypothese ('het zou kunnen zijn dat…'). Dat lijkt onschuldig, ware het niet dat er een paar grote adders onder het gras zitten. Allereerst is lang niet voor iedereen die onderzoeksartikelen leest duidelijk dat het hypotheses zijn en geen

onderzoeksresultaten. Persberichten van universiteiten en de krantenartikelen die daar weer uit voort komen, vergeten het geregeld te vermelden. Helaas willen ook neurowetenschappers zelf nog weleens deze hypotheses presenteren alsof het onderzoeksresultaten zijn, vooral wanneer ze spreken met een publiek dat niet inherent geïnteresseerd is in dopaminenetwerken en het ventrale tegmentale gebied. Bij het brede publiek, maar ook bij fondsenverstrekkers trekt 'liefde is verslavend' nou eenmaal meer aandacht.

Ten tweede is het opvallend welke van de mogelijke omgekeerde gevolgtrekkingen de onderzoekers kiezen als hypothese. 'Kijken naar je geliefde levert stress op' zou op basis van de neurowetenschappelijke onderzoeksliteratuur net zo (on)geldig kunnen zijn als de interpretatie dat liefde verslavend is. Maar dat idee druist natuurlijk in tegen wat onderzoekers zelf al dachten over liefde. Ze kiezen eerder voor een interpretatie die aansluit bij volkswijsheden over liefde ('liefde maakt blind'), of bij hun eigen theorie over liefde ('liefde is een motivatie, meer dan een gevoel') en/of bij ideeën die een hoge nieuwswaarde opleveren voor het onderzoek ('liefde is verslavend') (Van Stee 2017). En ten derde: wanneer interpretaties eenmaal in omloop zijn, worden ze vaak herhaald. Sommige interpretaties worden zo vaak herhaald dat iedereen vergeet dat het interpretaties zijn. De 'bereidheidspotentiaal' is hier een voorbeeld van. Maar ook het dopamine-netwerk waar ik het over heb, want dat wordt doorgaans 'beloningsnetwerk' genoemd. En veel mensen, inclusief neurowetenschappers zelf, realiseren zich überhaupt niet dat dat een interpretatie is, die bovendien lastig te rijmen valt met het feit dat het netwerk ook actief kan worden bij stress.

Omgekeerde gevolgtrekkingen vormen een belangrijke bron van misverstanden over de bijdragen die cognitieve neurowetenschap te leveren heeft aan ons inzicht in mensen. Het lijkt alsof neurowetenschappers bewijs vinden in de hersenen dat liefde verslavend is. Maar het is andersom: ze vonden iets uit over de hersenen en interpreteerden dat resultaat vervolgens aan de hand van een idee dat hen toch al aansprak. Op deze manier kan neurowetenschappelijk onderzoek ten onrechte bestaande ideeën bekrachtigen (Van Stee 2017). Dit misverstand kan kwalijke gevolgen hebben, waar bijvoorbeeld interpretaties worden gekozen die aansluiten bij bestaande vooroordelen. Denk aan 'Hersenonderzoek bevestigt vooroordelen over mannen en vrouwen'.

Zo zien we telkens weer dat hersenonderzoek ons inzicht biedt in de rol die onze hersenen spelen bij ons gedrag en onze ervaringen. Ook zien we telkens weer dat hersenonderzoek geen ander type inzicht in dat gedrag en die ervaringen oplevert.

Maar stel nou dat we de menselijke factor van het kiezen van interpretaties eruit zouden kunnen filteren. Stel dat we een algoritme zouden kunnen laten berekenen hoe groot de kans is dat bepaalde hersenactiviteit een indicatie is van de aanwezigheid van liefde. Of verslaving. Of willen. Of wat dan ook. Zou hersenonderzoek dan meer kunnen zeggen over wie wij zijn?

Gedachtenlezen

'Gedachtenlezen was lang sciencefiction. Maar anno 2018 kunnen hersenonderzoekers zien aan welke zinnen we denken [...] en of we suïcidaal zijn.' Zo begint een NRC-artikel dat onder de kop 'Ons brein is een open boek' zogenaamd 'gedachtenleesonderzoek' bespreekt. Hierbij worden statistische, zelflerende machines getraind (*machine learning*) om patronen in hersenactiviteit van elkaar te onderscheiden. In het oorspronkelijke voorbeeld: je laat mensen in een fMRI-scanner liggen en laat hen foto's zien van huizen en gezichten. De data over hersenactiviteit die je dan vindt, voer je aan zo'n machine. Je vertelt de machine er telkens bij: tijdens deze hersenscan zag de persoon een huis, tijdens deze scan een gezicht. De statistische machine leert dan zichzelf aan om op basis van de hersenscans te voorspellen of het een huis of een gezicht was dat de persoon met die hersenactiviteit zag. Na een tijdje kan de machine dat goed, ook wanneer het een scan gevoerd krijgt die het niet eerder in handen had. Bij het classificeren van huizen versus gezichten kan de machine het uiteindelijk zelfs foutloos (Haxby 2001).

Sinds dit pionierswerk is het type informatie dat statistische machines kunnen afleiden uit hersenactiviteit steeds complexer geworden. In 2017 publiceerden Marcel Just en zijn collega's twee onderzoeken waarbij een zelflerende machine 91% correcte classificaties kon maken. In het ene geval ging het erom aan welke van 36 mogelijke zinnen iemand dacht. In het andere onderzoek had de machine geleerd uit hersenactiviteit af te leiden of iemand suïcidaal was of niet.

Wanneer je de onderzoeken erop naslaat, wordt echter duidelijk hoe moeilijk het was voor de onderzoekers om tot zulke resultaten te komen en hoe kunstmatig de omstandigheden waren waaronder het uiteindelijk lukte. Om die 36 mogelijke zinnen van drie woorden te maken, hebben de onderzoekers in totaal maar 27 woorden gebruikt bijvoorbeeld. En die woor-

den zijn precies zo gekozen dat ze uit domeinen komen – onderdak, gereedschappen en voedsel – die met duidelijk van elkaar te onderscheiden patronen in hersenactiviteit gepaard gaan. De woorden werden een voor een op het scherm gepresenteerd tot ze een zin vormden en deelnemers werden geïnstrueerd om iedere keer dat hetzelfde woord op het scherm verscheen aan dezelfde eigenschappen van dat woord te denken. De 36 mogelijke zinnen kwamen ieder vier keer langs en daarna werden de 27 woorden ook nog eens ieder vijf keer los gepresenteerd. Tussen de zinnen door waren er lange pauzes om de hersenactiviteit weer tot rust te laten komen. Zo lagen de deelnemers een dikke 50 minuten doodstil en in opperste concentratie bij telkens dezelfde woorden dezelfde dingen te denken en verder niets.

In het onderzoek naar suïcidaliteit hield meer dan de helft van de deelnemers zulk soort omstandigheden niet vol. Ze bewogen hun hoofd een beetje, of raakten toch wat afgeleid. De zelflerende machine kon niet genoeg leren op basis van hun hersenactiviteit. Zo werd de machine uiteindelijk getraind met gegevens van 34 van de oorspronkelijke 79 deelnemers. Of beter gezegd: de machine werd getraind met gegevens van 33 deelnemers, waarna gekeken werd hoe hij de 34e deelnemer classificeerde, en of dat klopte. Dat werd herhaald met telkens een andere deelnemer als 34e. Gemiddeld classificeerde de zelflerende machine 91% van deze deelnemers correct. Wanneer de onderzoekers 17 van de 34 deelnemers in de trainingsgroep plaatste, classificeerde de machine de andere helft van deze groep in 76% van de gevallen correct.

Altijd waarschijnlijkheid

Ik vind dit soort inzet van *machine learning* een van de meest veelbelovende ontwikkelingen binnen de neurowetenschap. Toch zou ik willen stellen dat de term 'gedachtenlezen' hier volslagen misplaatst is. *Gedachten* gaan alle kanten op, in plaats van uitsluitend in vooraf bepaalde categorieën waar duidelijk van elkaar te onderscheiden hersenactiviteitspatronen bij horen. Gedachten zijn ook complexer dan die categorieën; ze betreffen niet alleen maar een plaatje van een huis, of het woord 'huis', maar ook de struik voor dat huis, de zon erboven en het gezicht dat door het raam kijkt. Gedachten komen bovendien in een constantere stroom dan in gefocuste brokken, met tussenpozen van 4 tot 7 seconden. Denken verloopt 'in het wild' heel anders dan in deze experimenten.

Door te spreken over het *lezen* van gedachten, suggereren journalisten verder dat neurowetenschappers met zekerheid iets kunnen zeggen. En ook dat scanners zinsbetekenissen of suïcidale gedachten aflezen in het brein, terwijl scanners alleen hersenactiviteit registreren. De zelflerende machines die op basis daarvan voorspellingen doen, kunnen dat alleen met behulp van een hele lading input door de onderzoekers tijdens de trainingsfase; en dan alleen met betrekking tot de categorieën waarop ze zijn getraind; en dat weer alleen met een bepaalde mate van waarschijnlijkheid.

Al met al is het dus ook geen gedachtenlezen – laat staan 'hacken' zoals het krantenartikel het ergens anders noemde – omdat het niet gaat zonder onze volledige medewerking. Je moet ervoor in een fMRI-scanner liggen of een EEG-kap op je hoofd zetten. Bovendien moet je doodstil blijven liggen en je volledig en uitsluitend focussen op de woorden of plaatjes op het scherm. Doe je dat niet, maar beweeg je je hoofd een millimeter of denk je toch af en toe aan iets anders, dan wordt het voor de machine erg moeilijk om je hersenscans te classificeren.

Wanneer in de toekomst de kwaliteit van de statistische machines toeneemt, zullen ook de classificaties met een hogere mate van zekerheid gemaakt kunnen worden. Tegelijkertijd zal het altijd een kwestie van waarschijnlijkheid blijven. En daarbij zal altijd gelden: hoe dichter de onderzoeksmethode in het lab de complexe werkelijkheid buiten het lab benadert (en hoe belangwekkender de classificatie dus is), hoe lager de waarschijnlijkheid waarmee de zelflerende machines voorspellingen kunnen doen.

Conclusie

Cognitieve neurowetenschap biedt inzicht in de hersenprocessen die betrokken zijn bij onze ervaringen met liefde, bij onze vrijwillige handelingen, bij de gedachten die we vormen en eigenlijk bij al ons doen en laten. Dat vertaalt zich in zelfinzicht waar het relevant is om onszelf als biologische-wezens-met-brein te beschouwen. Zo belooft neurowetenschap inzicht te geven in de problemen die we ervaren die veroorzaakt worden door hersenfalen. Ook kan het ideeën die we hebben over onszelf corrigeren, juist waar die ideeën al betrekking hadden op het brein. Verder onderstreept hersenonderzoek tal van psychologische onderzoeksresultaten die erop wijzen hoe vaak we op de automatische piloot handelen. En onder sterk gecontroleerde en daarmee behoorlijk kunstmatige omstandigheden kunnen we, altijd met

een mate van waarschijnlijkheid, een zelflerende machine laten voorspellen wat iemand ziet of waar iemand aan denkt.

Het jeugdige enthousiasme waarmee cognitieve neurowetenschap van start ging, ging geregeld gepaard met een dosis jeugdige overmoed over hoeveel het zou kunnen bewerkstelligen in een korte tijd. We hebben de afgelopen dertig jaar veel gewonnen qua inzicht in de rol van het brein bij ons doen en laten. Ook hebben we nu veel beter in de gaten wat de grenzen zijn van die kennis. Om op de titels uit de inleiding terug te komen: ons brein is verre van 'een open boek'. Neurowetenschappers doen onderzoek naar de hersenactiviteit die betrokken is bij ervaringen van verliefdheid, maar die hersenactiviteit vertelt ons niet direct iets over 'de klik, de kus en al het andere'. Het kan lijken alsof dat wel het geval is wanneer mensen ongeldige omgekeerde gevolgtrekkingen maken. In het ergste geval krijgen op die manier bestaande vooroordelen, bijvoorbeeld over mannen en vrouwen, extra retorische kracht.

Hersenonderzoek kan onhoudbare ideeën over de relatie tussen onze hersenen en onszelf ontkrachten. Wie dacht dat ons zelf of onze vrije wil niet met hersenactiviteit te maken had, bijvoorbeeld, die komt bedrogen uit. Tegelijkertijd kan neurowetenschap niet zeggen hoe we onszelf dan wel moeten begrijpen. Dat we altijd *ook* ons brein zijn, betekent namelijk niet dat we *alleen maar* ons brein zijn. En de meeste filosofische visies op de vrije wil staan nog recht overeind. Neurowetenschap helpt niet om daartussen te beslissen. Al met al geldt: inzicht in hersenprocessen vertaalt zich maar in beperkte mate in zelfinzicht.

Literatuur

Desmurget, M., Reilly, K. T., Richard, N., Szathmari, A., Mottolese, C., & Sirigu, A. (2009). Movement Intention After Parietal Cortex Stimulation in Humans. *Science*, *324*(5928), 811–813. https://doi.org/10.1126/science.1169896

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Just, M. A., Pan, L., Cherkassky, V. L., McMakin, D. L., Cha, C., Nock, M. K., & Brent, D. (2017). Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature Human Behaviour*, *1*(12), 911. https://doi.org/10.1038/s41562-017-0234-y

Just, M. A., Wang, J., & Cherkassky, V. L. (2017). Neural representations of the concepts in simple sentences: Concept activation prediction and context effects. *NeuroImage*, *157*, 511–520. https://doi.org/10.1016/j.neuroimage.2017.06.033

Lammel, S., Lim, B. K., & Malenka, R. C. (2014). Reward and aversion in a heterogeneous midbrain dopamine system. *Neuropharmacology*, *76*, *Part B*, 351–359. https://doi.org/10.1016/j.neuropharm.2013.03.019

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential). The Unconscious Initiation of a Freely Voluntary Act. *Brain*, *106*(3), 623–642. https://doi.org/10.1093/brain/106.3.623

Nisbett, R. E., & Wilson, T. D. (1977). Telling More Than We Can Know – Verbal Reports on Mental Processes. *Psychological Review*, *84*(3), 231–259.

Saul, J. (2013). Implicit bias, stereotype threat, and women in philosophy. In K. Hutchison & F. Jenkins (Eds.), *Women in Philosophy: What Needs to Change?* (pp. 39–60). Oxford University Press.

Schurger, A., Sitt, J. D., & Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences*, *109*(42), E2904-E2913. https://doi.org/10.1073/pnas.1210467109

Slors, M. V. P. (2012). *Dat had je gedacht! Brein, bewustzijn en vrije wil in filosofisch perspectief*. Amsterdam: Boom.

van Stee, A. (2017). *Understanding Existential Self-Understanding. Philosophy Meets Cognitive Neuroscience*. Leiden University.