De ontsnapping uit de donkere kamer

Sander Van de Cruys
Laboratory of Experimental Psychology, KU Leuven

Dit artikel verscheen eerder in het Karakter tijdschrift van wetenschap. Dit is deel 2 van een tweeluik, deel 1 vind je hier.

Volgens dichter en filosoof Paul Valéry (1871-1945) is het produceren van toekomst het doel van de hersenen. Het is precies dit idee dat zovele decennia later uitgewerkt en geformaliseerd wordt in de recente neurocognitieve theorie van de predictieve verwerking. Om de eigen toekomst te produceren moet het brein noodzakelijkerwijs een voorspellend model van die toekomst aanleren en toepassen. Zowel de binnenwereld (d.w.z. de noden van het eigen lichaam zoals ze evolueren doorheen de tijd) als de buitenwereld wordt voorspeld zodat prikkels die de het verdere voortbestaan van het organisme in gevaar brengen vaak op voorhand al geneutraliseerd kunnen worden. Als we ons voortbewegen in de wereld moeten we die wereld niet steeds van nul opbouwen puur op basis van onze zintuiglijke indrukken. Nee, we gebruiken voortdurend en onbewust onze mentale modellen om onze omgeving te construeren en om erop vooruit te lopen, terwijl de zintuiglijke informatie enkel ter correctie van die constructie dient, om te beletten dat onze constructies ontsporen in regelrechte hallucinaties. Dit wordt gevat in het centrale concept van de voorspellingsfout, het verschil tussen het verwachte, geconstrueerde inputpatroon en het werkelijke prikkelpatroon van onze zintuigen. Let wel, onze mentale modellen zijn in hiërarchische lagen opgebouwd, als een soort waterval van afhankelijke voorspellingen, waarbij de hogere lagen patronen voorspellen over grotere tijdschaal of ruimtespanne. In elke laag (elk gebied in de cortex) ontstaan voorspellingsfouten door confrontatie van de verwachte input (top-down; komende van een hogere regio) met de bottom-up neurale inputs van de regio. De theorie van de predictieve verwerking stelt dat het reduceren van voorspellingsfouten over al die lagen het fundamentele principe van de hersenwerking is. Zowel leren (het aanpassen van je modellen aan de wereld) als gedrag (het aanpassen van de wereld aan je modellen) kunnen ingezet worden om de gedetecteerde voorspellingsfouten te

verkleinen. Om een idee te krijgen van de wiskundige machinerie hierachter, stel je je best een berglandschap voor, waarbij de hoogte van de pieken de grootte van de voorspellingsfout voorstelt. Waar je je ook bevindt in het landschap, het doel is, net zoals een rollende bal, af te dalen naar de diepste dalen, de minima.

Critici van de theorie van de predictieve verwerking stellen echter dat het landschap dat de theorie tekent eerder een woestijnlandschap wordt, een verarmde voorstelling van ons mentale leven, waarin geen plaats is voor verlangens of emoties. Alles wordt gereduceerd tot voorspellingen met het reduceren van voorspellingsfouten als enige doel. De theorie mag het leren en onze waarneming dan wel goed verklaren (zie deel1), ze schiet schromelijk te kort in "zaken van het hart". Bovendien, zo gaan de critici verder, kan de theorie enkel gewoontedieren voortbrengen. Net zoals volgens velen de sociale media ons zouden opsluiten in echokamers die ons enkel dat wat we reeds verwachten voorschotelen zo zou ook de theorie van de predictieve verwerking een al te conservatieve reflex bestendigen. Ten dele klopt dit natuurlijk met de dagelijkse werkelijkheid voor mens en dier: het merendeel van de tijd zijn we inderdaad gewoontebeestjes, met ieder z'n eigen gedachte- en gedragsroutines. Ieder heeft daarbij z'n eigen selectieve blik op de werkelijkheid die wel min of meer werkt voor ons, maar ook regelmatig de onderlinge communicatie bemoeilijkt. Psychologen spreken hier over de confirmation bias: we voorspellen de wereld niet alleen, maar gaan ook selectief op zoek naar een omgeving die onze voorspellingen bevestigt (en daarmee de kans op voorspellingsfouten verkleint).

Sommige critici gaan nog verder: het centrale dictum van de reductie van voorspellingsfouten zou noodzakelijkerwijs ontaarden in een dictatuur van de foutenreductie. Die zou er in extremis toe leiden dat we onszelf moeten opsluiten in een donkere kamer, waar alle voorspellingsfouten sowieso zouden verdwijnen (als je gewoon duisternis voorspelt). De echte wereld zit vol ruis, verandering en onzekerheid, allemaal bronnen van voorspellingsfouten, dus is het gewoon gemakkelijker om je terug te trekken in een donkere kamer en daar de rest van je dagen te slijten. Dit staat bekend als het donkerekamerprobleem en is natuurlijk een metafoor voor het blijven hangen in een bepaald niveau of patroon van prikkeling. Men kan zich evenzeer een muziekkamer voorstellen waarin tot in de eeuwigheid een Bachconcerto wordt afgespeeld (wat, alle complexiteit ten spijt, na verloop van tijd natuurlijk ook extreem voorspelbaar wordt).

In zekere zin vinden we de figuurlijke donkere kamer wel terug in een psychiatrische stoornis als autisme. Mensen met dat syndroom sluiten zich geregeld af van de wereld. Ze onttrekken zich uit onvoorspelbare, veranderlijke sociale situaties, ten voordele van een

omgeving die vertrouwd en dus wel voorspelbaar is. In recent onderzoek (onder andere uit ons eigen labo) wordt de theorie van de predictieve verwerking dan ook met succes gebruikt om de manier waarop het brein van mensen met autisme informatie verwerkt beter te begrijpen. Als we ervan uitgaan dat mensen met autisme voorspellingsfouten een hoger gewicht geven, dan moeten we de karakteristieke repetitieve en rigide gedragingen (het 'wiegen' van het lichaam, het fladderen met de handen, het ordenen van speelgoed) misschien zien als een poging om terug te keren naar een aanvaardbaar, lager niveau van voorspellingsfouten. Het zich opsluiten in een extreem voorspelbare 'donkere kamer', inclusief het zelf creëren van voorspelbaarheid met cyclische bewegingen, wordt dan een logische compensatiestrategie in situaties met hoge onzekerheid.

Los van dit soort eerder atypisch gedrag, klopt het beeld van de donkere kamer uiteraard patent niet met onze ervaring. Wij (en andere dieren) zijn vaak uit op het uitbreiden van onze actieradius. We banen nieuwe paden in een onzekere wereld, verkennen met veel plezier nieuwe omgevingen, en laten ons graag verrassen in de kunst, muziek en de humor. Hoe valt dit te rijmen met een wezen dat voortdurend gericht is op het verlagen van voorspellingsfouten? Toch staat de theorie hierin niet machteloos.

De letterlijke donkere kamer valt het gemakkelijkst af te serveren: zo'n kamer blijft in de praktijk niet lang voorspelbaar voor een organisme dat ook evolutionair bepaalde, homeostatische verwachtingen heeft (bijvoorbeeld over het gevoed zijn, of een bepaald verwacht glucosepeil in het bloed; zie deel1). Er bestaan echter wel 'wezens van de duisternis': eencelligen die op zoek gaan naar duisternis (skototroop), omdat dit in hun leefomgeving nu eenmaal samengaat met de aanwezigheid van vocht en voedingsstoffen. Het gaat dan opnieuw om een organisme dat een relevante regelmaat in zijn omgeving internaliseerde, een voorspelling dus. Onze leefomgeving en dus onze evolutionair meegegeven modellen zijn heel anders. Onze capaciteit voor het leren van nieuwe, hiërarchische voorspellingen is ook zoveel groter, doordat het vervullen van homeostatische verwachtingen (lees: overleving) erg afhankelijk is van tijdens het leven aangeleerde voorspellingen. Die voorspellingen situeren zich op een abstracter niveau (van objecten, gebeurtenissen, soortgenoten en hun intenties) in plaats van op het laagste niveau (van bijvoorbeeld lichtcontrasten). Onze modellen moeten dus inhaken op die gedragsrelevante (bijvoorbeeld sociale) regelmatigheden in onze omgeving, in plaats van voortdurend gegijzeld te worden door de chaotische, ruizige inputs van het hier en nu. Het vermijden van voorspellingsfouten op de laagste trede, met name door middel van duisternis, biedt geen antwoord op de uitdagingen van een organisme, met name het vervullen van homeostatische verwachtingen.

Maar de echte angel van het donkerekamerprobleem zit 'm natuurlijk in het verklaren van onze zucht naar nieuwigheid, onze zin voor creativiteit, onze drang naar volstrekte verrassing. Toch zit ook dit vervat in de theorie van de predictieve verwerking. Goed voorspellen betekent ook voorspellen wanneer je meer informatie nodig hebt om goede voorspellingen te kunnen maken (zie deel 1). Dat wil zeggen dat, in een omgeving die je nog niet goed kent, een verkennende actie erg belangrijk kan zijn, omdat ze als doel heeft meer van de voorspelbare structuur van de omgeving te ontsluiten, en dus kleinere voorspellingsfouten zal opwekken in de toekomst. Die zogenaamde 'epistemische' acties (exploratieve acties om meer te weten te komen over onze omgeving) kunnen gaan van kleine oogbewegingen tot het verkennen van een vreemd land. We zien dat we hier wel doelbewust op zoek gaan naar voorspellingsfouten, hoewel we die fouten niet lukraak kiezen: we zoeken die fouten die het meest informatief zijn voor de modellen waarover we op dit moment beschikken. Het doel blijft om op langere termijn (op alle niveaus van de hiërarchie) toekomstige voorspellingsfouten te vermijden. We gaan dus niet zomaar nieuwigheid om de nieuwigheid nastreven, in dat geval kunnen we evengoed eeuwig naar het sneeuwbeeld op de televisie staren. Daar vinden we niets dan voorspellingsfouten (ruis) maar ook geen structuur die geleerd kan worden. Wat we in de psychologie vaak zien is dat mensen een voorkeur hebben voor visuele prikkels of activiteiten met middelmatige complexiteit: niet te hoog, niet te laag (het zogenaamde Goudlokjeprincipe). Ze vermijden mentale tijd vruchteloos te spenderen aan enerzijds prikkels die ze door en door kennen en waarvan ze dus alle voorspelbare structuur reeds gevat hebben in hun modellen (weinig resterende voorspellingsfout) en anderzijds aan prikkels met erg hoge voorspellingsfouten omdat ze de gepaste voorspellende modellen ervoor nog niet geleerd hebben (te complex) of omdat er gewoon geen structuur uit af te leiden valt (ruisbeeld). De mens zoekt daardoor activiteiten en situaties op waar hij of zij de grootste voorspellingsvooruitgang kan boeken, waar er dus nog het meest te leren valt, gegeven zijn of haar capaciteiten (lees: mentale modellen). Dat we hier zo gevoelig voor zijn, betekent dat we ook (meta-)voorspellingen vormen over toekomstige voorspellingsfouten. Afgaande op onze ervaring schatten we voor de mogelijke handelingen in een bepaalde situatie de verwachte fouten in en hoe reduceerbaar die zullen zijn. Op basis hiervan kiezen we onze daden en omgeving.

Daarmee komen we terecht in het tegenfeitelijk denken dat zo karakteristiek is voor de menselijke cognitieve souplesse ('wat als…?') en ons vermogen om onze modellen offline te nemen en als simulaties te gebruiken. Hier opent zich de weg naar typisch menselijk beredeneerd en flexibel gedrag. En hier bereikt het denken zijn hoogtepunt wat betreft het loskomen van het hier en nu van onze voelsprieten. Het staat buiten kijf dat we erin slagen

om onze modellen offline, los van actuele zintuiglijke input, te gebruiken en zelfs te verfijnen in ons denken en onze dromen. Wat we daar doen, komt technisch gezien neer op het verminderen van de interne complexiteit van onze modellen. Als dat niet gebeurt, gaan onze modellen de zintuiglijke data overfitten, dat wil zeggen: ze gaan te zeer toegespitst zijn op de training data, de data waarmee ze geleerd zijn. Die modellen zijn verkwistend omdat ze regelmatigheden in de omgeving veronderstellen die er niet zijn. Daardoor gaan ze het ook slechter doen (lees: meer voorspellingsfouten genereren) op nieuwe data, in nieuwe situaties. Met andere woorden: hun toepasbaarheid is beperkt. Modellen die minder complex zijn en die de omgeving compacter samenvatten, kunnen het op lange termijn toch beter doen in termen van voorspellingsfouten. We ervaren het effect van dat proces van complexiteitsreductie en reductie van voorspellingsfouten vaak als een positieve aha-erlebnis, wat aangeeft dat ook onze emotionele ervaring sterk vervlochten is met het succes en falen van onze predictieve machinerie doorheen de tijd. Dat is goed te zien in de opbouw van moppen. Moppen starten vaak met het zorgvuldig opwekken van een sterke verwachting, soms door expliciete herhaling, waarna die verwachting geschonden kan worden (voorspellingsfout), wat op zijn beurt de gelegenheid geeft om de voorspellingsfout te reduceren (tenminste, als je de clou hebt). Ervaar het zelf in dit voorbeeld, vertaald uit Inside Jokes (Hurley, Dennett, & Adams, 2013):

Een man en een vrouw die elkaar nog nooit eerder hebben ontmoet, delen een slaapwagon in een nachttrein. Na de aanvankelijke schaamte gaan ze allebei slapen in hun compartiment. Maar midden in de nacht leunt de vrouw voorover en zegt tegen de man: 'Het spijt me, maar ik heb het een beetje koud. Zou jij nog een deken voor me willen halen?' 'Ik heb een beter idee', antwoordt de man met een glimp in zijn ogen. 'Laten we voor één nachtje net doen of we getrouwd zijn.' 'Oké, waarom niet', giechelt de vrouw. 'Geweldig', zegt de man. 'Haal je verdomde deken zelf!'

Het onverwacht oplossen van de voorspellingsfout komt meestal na een switch in het gehanteerde model voor de situatie, en geeft de karakteristieke positieve emotie (de lach). Hoewel de mop het mentaal model (de verwachting) van een pasgetrouwd stel opwekt, kan je de verrassende wending op het einde pas wegverklaren met een model van een echt (verkild) huwelijk.

Tot slot wordt de theorie van de predictieve verwerking recent ook volop gebruikt om onze appreciatie van kunst en muziek systematisch terug te voeren tot dynamieken in (on)voorspelbaarheid. Kunstenaars en musici gebruiken intuïtief onze sterke verwachtingen (al dan niet door ze eerst te bevestigen) om ze bruusk te kunnen schenden in hun

kunstwerken (zoals Picasso die een aangezicht, een patroon dat we erg goed kennen, compleet 'verknipt'). Soms leidt dit ertoe dat we compleet afhaken en zo verdere 'fouten' vermijden —het uitzetten van een moeilijk stuk jazz—, maar vaak laten net die opzettelijke voorspellingsfouten ons toe nieuwe betekenis (voorspelbare structuur) te vinden in het kunstwerk (het is bijvoorbeeld heel passend dat een droef gezicht 'gebroken' is). Het plezier zit hier voor een groot stuk in de predictieve vooruitgang die we boeken, het ontdekken van nieuwe patronen in de werkelijkheid (soms onbedoeld). Onze 'voorspellingslust' verschijnt hier als bron van in plaats van als rem op wetenschap en kunst. Van kunstenaars wordt terecht gezegd dat ze ons de werkelijkheid laten ervaren zoals we dat de allereerste keer deden, als verwonderde kinderen. Dat is niet toevallig de periode waarin we ook de grootste leervooruitgang boekten. Door een 'kunstgreep' geven kunstenaars ons die ervaring terug.

Tot daar mijn onbesuisde poging om aan te tonen dat ook wetenschap zulks soms kan betrachten.

Meer lezen?

Sun, Z., & Firestone, C. (2020). The Dark Room Problem. Trends in Cognitive Sciences. https://doi.org/10.1016/j.tics.2020.02.006.

Van de Cruys, S., Friston, K.J., & Clark, A. (2020) Controlled optimism. Trends in Cognitive Sciences. (in press)