

“Know how to rank beliefs not according to their plausibility but by the harm they may cause”.

Jelle over The Black Swan – 01
Feb 2008

Ik lees nu The Black Swan van Taleb. Dat gaat over de onevenredig grote invloed van extreme gevallen. Ik ga dat hier niet helemaal bespreken, voor meer info moet je even Googlen of Wiki-en. Het gaat mij om de toepassing van de discussie in het boek van Taleb op de psychologie. In eerste instantie dacht ik: ojee, als extreme scores een onevenredig grote invloed hebben op het resultaat, dan heeft de psychologie een groot probleem. Dat heeft te maken met dat de psychologie gebaseerd is op de aanname van een ‘normale verdeling’, ofwel wat men wel noemt “The Bell Curve”. En Taleb heeft een enorme hekel aan de Bell Curve. We passen hem toe overal waar hij niet van toepassing is. Dus ik voelde direct nattigheid. Maar, zegt Taleb, gelukkig heb je daar in de psychologie – op zich - geen last van. Psychologisch onderzoek meet altijd of er een verschil is tussen twee groepen mensen, twee ‘condities’. En dan gelden de problemen die Taleb bespreekt niet. Waar gaat het allemaal om? Bijvoorbeeld: je wilt weten of er een verschil is in de manier waarop mensen een willekeurige andere persoon beoordelen (laten we hem Johan noemen), waarbij je twee groepen vergelijkt: de eerste groep beoordelers, groep A, krijgt van te voren een beker warme choco in de handen geduwd en de tweede groep beoordelers, groep B, krijgt van te voren een beker koude cola in de handen geduwd. Dat gebeurde in de lift, waar de proefleider schijnbaar toevallig iets uit zijn tas moest halen en vroeg: wil je mijn beker even vasthouden? Daarna hadden ze een kort gesprekje met Johan en daarna, wanneer Johan de kamer heeft verlaten, werd hen gevraagd: zou jij Johan aannemen als je een vacature had, je baserende op deze korte ervaring met hem? Wat blijkt is dat mensen die eerst een korte tijd een beker koffie hebben vastgehouden, groep A, vervolgens een positiever oordeel velt over de capaciteiten van Johan. Positiever dan het oordeel van groep B, die even tevoren een tijdje een beker koude cola in de handen hebben gehad. Een mysterieus effect van ‘warmte’, maar het effect is echt. (Afgelopen dinsdag werd het overigens besproken in een aflevering van de documentaireserie Horizon op BBC2). Het effect is statistisch significant. De reden dat ik het hier noem is dat Taleb nu uitlegt, dat statistisch significante effecten in dit soort gevallen wel degelijk betekenis hebben, maar in heel veel andere situaties in onze werkelijkheid niet. De vraagstelling is zwart/wit: je vraagt je af *of* er een verschil is tussen beide groepen. Het antwoord op die vraag is ja, of nee. Niet: een beetje, of half-om-half. En de psychologie doet al helemaal geen uitspraken over de precieze grootte van het effect. Het vergelijken van groepen op een dichotome variabele is veilig, omdat losse scores van individuele personen in beide groepen ieder voor zich maar een kleine invloed hebben op het beantwoorden van die ja/nee vraag. In dit geval is het meetniveau van het oordeel van de proefpersonen in beide groepen zelf ook zwart/wit (de technische term hiervoor is: een categorische variabele). Dat wil zeggen: mensen moesten zeggen of zij Johan WEL of NIET zouden aannemen. Dat betekent dat er slechts twee meetwaarden zijn: ja/ nee, 1 of 0. Zelfs bij een groep van, zeg, zes proefpersonen maakt het oordeel van elk van die personen al niet meer uit: je kunt willekeurig een persoon uitkiezen, zijn oordeel omdraaien (van ja naar nee of van nee naar ja) en dan nog zou de groep als geheel tot hetzelfde oordeel komen als ervoor. Individuele scores hebben weinig invloed op het totaal. Maar zelfs als je een ‘rijkere’ score bedenkt, bijv je mag Johan indelen op een schaal van 0 tot 100, waarbij 0 een

heel negatief oordeel is en 100 heel positief, dan nog hebben individuele scores weinig invloed. Als je maar genoeg proefpersonen hebt, dan zullen de 'outliers' het gemiddelde niet zodanig beïnvloeden dat daarmee de groep als geheel ineens van oordeel switcht. Dat komt, uiteindelijk, omdat je altijd een grens trekt. Je wilt immers uiteindelijk alleen weten of er *een* verschil is tussen de twee groepen. En de totale spreiding van alle waarderingen voor Johan bevindt zich altijd tussen de 0 en de 100. Je kunt nu met een gerust hart twee gemiddelden berekenen: de gemiddelde score voor groep A (koffie) en de gemiddelde score voor groep B (cola). En de variatie rondom die scores, zo kun je gerust aannemen, is normaal verdeeld. Zelfs als het niet normaal verdeeld is, wordt de belangrijkste voorwaarde voldaan: elke individuele score bevindt zich tussen de 0 en de 100 en zal geen grote invloed hebben op de totale beslissing, namelijk, is de (gemiddelde) beoordeling van (de mensen in) groep A verschillend dan die van groep B?

Overigens is er dan wel een ander probleem, namelijk dat als je alleen maar wilt voorspellen of er een verschil is tussen groep A en B, dat je dan eigenlijk maar heel weinig 'informatie' hebt op basis waarvan je een theorie gaat maken. Dus op basis van bovenstaande experiment zijn er wel 100, ach nee wel 1000000 mogelijke theorieën die dit effect verklaren. Terwijl als je zou kunnen bewijzen dat het verschil tussen koffie-vasthouders en cola-vasthouders in hun oordeel over Johan gemiddeld de waarde 13,4759603 benadert, op een schaal van 1 tot 100, *dan* moet je wel een heel speciale theorie formuleren om te verklaren waarom dat verschil *precies* 13,4759603 is en niet 13,4759604. Ineens moet je nu diep het mechanisme induiken, terwijl je daarvoor nog een beetje verhaaltjes kon verzinnen die ons mensen 'logisch in de oren klinken'. (De onderzoekers van bovenstaand experiment verklaren het effect als een erfenis van het gevoel van warmte/koude in de vroege jeugd, aan de moederborst zagezegd, en menen dat dit gevoel nog steeds sterke emotionele gevoelens oproept. Dat kan waar zijn. Ik kan ook nog een ander verhaal verzinnen...).

Maar laten we nu eens aannemen dat je zelf mag weten wat voor getal je neemt. En we nemen uiteindelijk gewoon weer het gemiddelde voor de hele groep. Er is geen limiet aan het getal, dat jij mag uitkiezen om je waardering voor Johan tot uitdrukking te brengen. Dat zou betekenen dat er maar 1 persoon hoeft te zijn die als score 100.000 kiest, die ervoor kan zorgen dat het oordeel van een hele groep omzwaait van 'lager', naar 'hoger' dan de andere groep. Maar er hoeft vervolgens ook maar 1 toevallige grapjas in de andere groep te zijn die het oordeel 1000.000 noemt, om het weer om te laten zwaaien. En in de andere groep hoeft er maar eentje te zijn die 1000000000 kiest etc... Met andere woorden, in zo'n situatie kan een losse score onevenredig veel invloed hebben op het eindresultaat. [ik weet nog niet zeker of dat nu wel of niet invloed heeft op een ja/nee beslissing als in dit experiment, maar in de voorbeelden die Taleb noemt gaat het niet om een ja/nee beslissing maar om een precieze voorspelling van de waarde van een variabele, en dan is de invloed van 'outliers' enorm groot].

Taleb legt uit dat dit in de economie en in andere sociaal-maatschappelijke en historische processen voortdurend het geval is. Jarenlang stijgt de waarde van huizen, dan ineens klappt de hele boel in elkaar. Tijdreeksen zijn mysterieus en onvoorspelbaar. Je kunt op basis van het verleden wel een patroon herkennen, maar niets zegt je dat dat patroon in stand blijft, omdat er geen limiet is aan de mogelijke waarden van de variabele in kwestie (bijv: het inkomen van individuele mensen) en er

kan dus altijd ineens een nieuwe observatie zijn die *zo enorm veel afwijkt van de waardes die er tot nu toe zijn waargenomen*, dat die nieuwe waarde het hele patroon, het hele model, weer teniet doet: tot die tijd voldeed de variabele netjes aan een bepaalde trend, een bepaald model, een bepaald patroon, en daarna: ineens niet meer! Behaalde resultaten in het verleden bieden GEEN garantie voor de toekomst. Als je gelooft dat dat soort onverwachte klappers zelden voorkomen, heb je gelijk. Maar het vervelende is dat de zeldzaamheid er in dit geval helemaal niet toe doet. Het feit dat het een ‘zeldzame’ observatie is doet er niet toe, omdat het gewicht dat die ene observatie in de schaal ligt zo groot is dat het alle ‘veel voorkomende observaties’ in zijn eentje van tafel veegt. Neem de verdeling van inkomsten die schrijvers hebben van hun boeken. J.K Rowling verdient in haar eentje meer dan de helft van wat alle andere nu levende schrijvers in de wereld bij elkaar verdienen! Dat is enorm veel geld. Buitenproportioneel veel meer dan de anderen. Het verschil tussen Rowling en de eerste die na haar komt is enorm veel groter dan het verschil tussen al die ‘gemiddelde; schrijvers. Of stel, je observeert honderden jaren lang dat het aantal burgerslachtoffers bij kleine oorlogen nooit boven de 1000 komt. Dan ineens wordt de atoombom uitgevonden. Een klein conflict, een gekke dictator, en boem: de halve wereldbevolking gaat de pijp uit. Van max 1000 gaan we nu ineens naar 3 miljard! Het is een zelfdzaamheid, maar als het gebeurt, dan is je verklarende model ineens waardeloos geworden. Het probleem is, zegt Taleb, dat er wel enorm veel processen zijn die op deze manier in elkaar steken, dus dat er uiteindelijk enorm veel processen zijn waar we maar hoeven te wachten op die ene grote klapper die al onze verwachtingen doet ineenstorten.

Vanaf hier wordt ik wellicht wat vaag/associatief, dit moet nog worden uitgewerkt. Ik heb het boek nog niet uit, maar ik heb sterk het idee dat er in de psychologie wel degelijk vergelijkbare processen spelen. Niet zozeer mbt de methodologie van het psychologisch experiment, maar wel mbt de daadwerkelijke, onderliggende processen die deze experimenten proberen te verklaren. Bijv in de neurale netwerk-literatuur leert een netwerk van knopen, (in een computer) stap voor stap van nieuwe waarnemingen die hij doet. Bijv het computerprogramma ziet een hele reeks foto's en moet daar dan een patroon uit destilleren. Die patroonherkenning is volgens mij zonder uitzondering hetzelfde soort van modelvorming en ‘voorspellen op basis van historische gegevens’. Maar ook in die situatie kun je je voorstellen dat je vervolgens een waarneming doet die zover buiten de kaders van het voorgaande ligt dat jouw model ineens waardeloos is geworden. Met andere woorden ik heb het vermoeden dat het principe van The Black Swan ook bij theorieën en met name bij concrete computationele modellen van kennis en leren, relevant is. Het heeft allemaal te maken met inductie, en met de vraag wanneer het zinnig is om op basis van een reeks waarnemingen een algemeen patroon te ‘induceren’. Taleb zegt “Know how to rank beliefs not according to their plausibility but by the harm they may cause”. Hier ga ik nu verder over nadenken...