RADBOUD UNIVERSITEIT

Zelfkennis en Zelfbedrog

Een onderzoek naar de relatie tussen zelfkennis en zelfbedrog

Tom Broeder S4623452

28-12-2019

Begeleid door: L.C. de Bruin

Aantal woorden: 17.460

Masterscriptie van de Masteropleiding Analytische Filosofie aan de Faculteit Filosofie, Theologie en Religiewetenschappen van de Radboud Universiteit Nijmegen.

Zelfkennis is de kennis die wij hebben van onze mentale toestanden, zoals overtuigingen, verlangens, intenties, etc. Zelfbedrog is het fenomeen waarbij we ten onrechte een overtuiging vormen, terwijl er duidelijk bewijs is voor het tegenovergestelde. In deze scriptie onderzoek ik hoe deze onderwerpen aan elkaar gerelateerd zijn.

Tom Broeder
S4623452

Hierbij verklaar en verzeker ik, Tom Broeder, dat voorliggende eindwerkstuk getiteld Zelfkennis en Zelfbedrog, zelfstandig door mij is opgesteld, dat geen andere bronnen en hulpmiddelen dan die door mij zijn vermeld zijn gebruikt en dat de passages in het werk waarvan de woordelijke inhoud of betekenis uit andere werken – ook elektronische media – is genomen door bronvermelding als ontlening kenbaar gemaakt worden.

Nijmegen, 28-12-2019

Inhoud

Inleiding

In de analytische filosofie verstaan we onder zelfkennis de kennis die wij hebben van onze mentale toestanden, zoals overtuigingen, verlangens, intenties, etc. Men is het er doorgaans over eens dat deze zelfkennis een speciale vorm van kennis is, die ons rechtstreeks gegeven wordt en waar wij agentschap over hebben. Als het inderdaad zo is dat wij ons in deze unieke positie bevinden, waarbij we op het gebied van zelfkennis epistemisch bevoorrecht zijn, dan zou het natuurlijk handig zijn als deze kennis die we vergaren van onze mentale toestanden ook juist is. In de meeste gevallen weten we ongeveer wel of we ervan overtuigd zijn dat het regent, of we verlangen naar die lekkere chocoladereep, of dat we de intentie hebben om morgenochtend vroeg op te staan.

Soms komt het echter voor dat we, in het vergaren van zelfkennis, onszelf met opzet voor de gek houden. Dit fenomeen wordt zelfbedrog genoemd. Er is bijvoorbeeld sprake van zelfbedrog als we een onjuiste overtuiging vormen, terwijl het bewijs waarop wij die overtuiging vormen wijst naar iets tegenovergestelds. Dit proces wordt meestal geleid door een motivatie zoals bijvoorbeeld een verlangen. Zo kun je bijvoorbeeld een sterk vermoeden hebben dat je partner vreemdgaat, er ook nog eens bewijs voor hebben, maar vanwege je verlangen dat dit niet het geval is jezelf aanpraten dat het ook daadwerkelijk niet het geval is. Toch is er in het geval van zelfbedrog wel sprake van enig bewustzijn van de waarheid. We weten dat we iets niet zouden moeten doen, geloven of willen, maar we weten onszelf toch naar het tegenovergestelde te laten handelen.

Het lijkt er dus op dat zelfbedrog negatieve gevolgen heeft of zelfs een gevaar vormt voor de manier waarop wij zelfkennis vergaren. In deze scriptie onderzoek ik daarom de vraag hoe zelfkennis en zelfbedrog aan elkaar gerelateerd zijn.

Om deze vraag te beantwoorden zal ik in het eerste hoofdstuk uitweiden over het onderwerp zelfkennis. Ik zal hierbij ingaan op het onderscheid tussen triviale en substantiële zelfkennis, oftewel het onderscheid tussen kennis van onze mentale toestanden en kennis van ons karakter/persoonlijkheid. Ook zal ik het hebben over het onderscheid tussen zelfkennis en andere vormen van kennis. In het tweede hoofdstuk zal ik het fenomeen zelfbedrog in kaart brengen. Dit omvat een uitleg over hoe zelfbedrog op verschillende wijzen gedefinieerd wordt, een beschrijving van twee theorieën van zelfbedrog, namelijk een intentionele en een motivationalistische theorie, een uitweiding over de aard van zelfbedrog aan de hand van het fenomeen confabulatie, en de mogelijke verklaringen voor het bestaan van zelfbedrog. In de derde sectie zal ik enkele uitwerkingen van de relatie tussen zelfkennis en zelfbedrog analyseren aan de hand van Fernández en Scott-Kakures, die zelfbedrog verklaren als een fout die optreedt in het vergaren van zelfkennis. Tot slot zal

ik op basis van het voorgaande concluderen dat er aan zelfbedrog een falen van zelfkennis ten grondslag ligt.

1 Zelfkennis

1.1 Triviale en substantiële zelfkennis

Zoals in de inleiding aangegeven verstaan we zelfkennis in de analytische filosofie doorgaans als de kennis die wij hebben van onze mentale toestanden. Dit kunnen bijvoorbeeld overtuigingen, verlangens of intenties zijn. De soort mentale toestanden die worden behandeld in de analytische filosofie zijn meestal niet de meest belangrijke of ongrijpbare toestanden. Als ik de overtuiging heb dat ik sokken draag en ik weet dat ik die overtuiging heb, dan is het zo dat ik iets weet over de toestand van mijn *mind.* Dit kunnen we vervolgens zelfkennis noemen, maar het lijkt een triviale vorm van zelfkennis te zijn. De 'volks-' notie van zelfkennis omvat doorgaans namelijk ook kennis van je karakter, je vaardigheden, je emoties, of dat wat je gelukkig maakt. Dit lijkt veel belangrijker dan triviale zelfkennis.

Deze twee verschillende vormen van zelfkennis typeert Cassam als triviale en substantiële zelfkennis (Cassam 2015, preface). Het weten dat ik de overtuiging heb dat ik sokken draag, wordt zelfkennis genoemd. Toch is dit een triviaal voorbeeld van zelfkennis als we het vergelijken met substantiële zelfkennis. Als ik bijvoorbeeld weet dat ik een aanleg heb voor het omgaan met ongemakkelijke collega's of dat ik in discussies geneigd ben het voortouw te nemen, heb ik wat lijkt op substantiële zelfkennis; zelfkennis die het waard is om te hebben (Cassam 2015, 29).

Voorbeelden van substantiële zelfkennis zijn:

- Weten dat je gul bent (kennis van je karakter).
- Weten dat je niet racistisch bent (kennis van je waarden).
- Weten dat je Spaans kan spreken (kennis van je vaardigheden).
- Weten dat je verliefd bent (kennis van je emoties). (Cassam 2015, 30)

Het onderscheid tussen substantiële en triviale zelfkennis is een gradueel onderscheid, waarbij sommige van de bovengenoemde voorbeelden substantiëler zijn dan de ander. Als we willen weten waarom deze vormen van zelfkennis substantieel zijn, moeten we kijken naar de vraag wat een stuk zelfkennis substantieel maakt. Hiervoor heeft Cassam een lijst samengesteld met kenmerken van wat stukken zelfkennis substantieel maakt. Het idee is dat hoe meer van deze kenmerken een stuk zelfkennis heeft, hoe substantiëler die zelfkennis is. Gezien de grootte van de lijst zal ik slechts een paar kenmerken benoemen om een indruk te geven:

- De feilbaarheidsvoorwaarde: met substantiële zelfkennis is er altijd de mogelijkheid van een fout of vergissing.

- De obstakelvoorwaarde: de mogelijkheid van vergissing in zulke voorbeelden is een reflectie van het feit dat er voor mensen obstakels zijn in het vergaren van substantiële zelfkennis, zoals repressie, zelfbedrog, vooringenomenheid, schaamte, etc.
- De waardevoorwaarde: substantiële zelfkennis doet er toe in praktische, morele zin. Niet weten wat je blij maakt kan bijvoorbeeld resulteren in het maken van slechte keuzes. (Cassam 2015, 31-3)

In de filosofie is zelfkennis veel specifieker, aangezien het de kennis betreft van iemands specifieke mentale toestanden. Wanneer het aankomt op specifieke zelfkennis, kunnen we drie aspecten benoemen die deze vorm van zelfkennis kenmerken:

1. In de context van het uitleggen van kennis van onze eigen propositionele houdingen, is de focus meer gelegd op een smalle reikwijdte van propositionele houdingen, met een bijzonder sterke nadruk op kennis van onze overtuigingen en verlangens. De vraag hoe je je eigen overtuigingen kunt kennen wordt bijvoorbeeld uitvoeriger besproken dan de vraag hoe je een emotie als hoop kunt kennen.
2. Wanneer het gaat over kennis van onze overtuigingen, zijn de voorbeelden vaak triviaal. De aandacht binnen het filosofisch discours over zelfkennis ligt veeleer bij het uitleggen van de kennis van een persoon dat hij gelooft dat het regent, in plaats van dat hij gelooft dat man en vrouw gelijk zijn of dat God bestaat. De nadruk ligt bij het verklaren van zelfkennis van relatief triviale houdingen.
3. Kennis van iemands specifieke overtuigingen en verlangens kan kennis van wat iemand gelooft of verlangt impliceren, of kennis van waarom iemand gelooft wat diegene gelooft of wilt wat diegene wilt. Filosofen zijn voornamelijk geïnteresseerd in het ‘wat’ in plaats van het ‘waarom’ bij het verklaren van zelfkennis. (Cassam 2015, 39-40)

Waarom ligt de voorkeur voor filosofen bij deze, op het eerste gezicht gelimiteerde, vorm van zelfkennis? Een mogelijkheid is dat specifieke zelfkennis voor veel filosofen interessant is omdat het begrepen wordt als kennis die verschilt van kennis van de externe wereld. Maar waarom is dit belangrijk? Het is natuurlijk handig om te weten of het buiten regent als je een stukje gaat wandelen, maar waarom zou het voor jou interessant zijn om te weten dat je de overtuiging hebt dat het regent? Een mogelijke verklaring hiervoor is het zogeheten fundamentalisme in de epistemologie. Fundamentalisten prefereren deze triviale vorm van zelfkennis omdat zij het idee hebben dat onze overtuigingen als het ware een piramidestructuur hebben. Hierbij vormen basale overtuigingen de fundering en alle andere gerechtvaardigde overtuigingen ondersteund door rede ontstaat daaruit. Wat overtuigingen basaal maakt, is dat ze beschouwd worden als onfeilbaar en non-inferentieel

gerechtvaardigd. Specifieke zelfkennis is fundamenteel in vergelijking met de rest van onze kennis en dat is waarom deze vorm van zelfkennis belangrijk is (Cassam 2015, 40-1).

Een niet-fundamentalistische benadering van het belang van specifieke zelfkennis wordt gegeven door Burge (Burge 1998). Volgens Burge wordt kritische redenatie geleid door appreciatie, gebruik en toepassing van redenen en redenatie als zodanig. Om kritisch te redeneren moet iemand in staat zijn om zijn kennis van redenen te gebruiken om *commitments* (toewijdingen) ten opzichte van proposities te maken, bekritiseren, aan te passen en te bevestigen. Dit houdt in dat iemand die kritisch redeneert alleen veranderingen in zijn mentale toestanden aanbrengt die gebaseerd zijn op redenen. Volgens deze benadering vereist kritisch redeneren het nadenken over je eigen gedachten; je kan je eigen gedachten niet kritisch evalueren zonder na te denken over die gedachten. Sterker nog, je moet je gedachten kennen. Burge biedt hiermee een transcendentaal argument voor het belang van zelfkennis: kritische redenatie komt voor onder onze soort, specifieke zelfkennis is noodzakelijk voor kritische redenatie, dus we hebben specifieke zelfkennis. Zelfs de overtuiging dat je sokken draagt staat open voor kritiek, bevestiging en kan aangepast worden op basis van redenen. Dit is alleen mogelijk als je weet dat je die overtuiging hebt en daarom is deze vorm van zelfkennis van belang (Cassam 2015, 41-2).

Hoewel de filosofische literatuur over zelfkennis neigt naar de triviale variant, wordt er in de literatuur over zelfbedrog veel geschreven over zelfkennis in substantiële zin. In het geval van zelfbedrog is het namelijk veel interessanter om te weten of je jezelf niet voor de gek houdt met betrekking tot je waarden en vaardigheden dan met betrekking tot het feit of je sokken aan hebt. Daarom zal ik me in deze scriptie niet beperken tot een van de twee soorten zelfkennis.

1.2 Onderscheid zelfkennis met andere vormen van kennis

Zelfkennis wordt doorgaans beschouwd als een vorm van kennis dat verschilt met andere vormen van kennis, zoals bijvoorbeeld kennis van de buitenwereld. Zo maken we bij het vergaren van zelfkennis gebruik van een speciale methode in de zin dat die methode exclusief beschikbaar is voor het subject zelf, zou zelfkennis epistemisch zekere kennis zijn en is het kennis waar wij agentschap over hebben. In deze sectie zal ik een aantal theorieën uiteenzetten die deze speciale status van zelfkennis verklaren.

1.2.1 Een speciale methode

Een punt waarop zelfkennis verschilt van andere vormen van kennis, is dat we bij het vergaren van zelfkennis gebruik maken van een speciale methode die exclusief voor ons, het subject, beschikbaar is. Een van die methodes is introspectie, een methode die in ieder geval vanaf Descartes al zijn oorsprong kent.

Descartes meent dat we aan de hand van introspectie onfeilbaar zijn in de manier waarop we zelfkennis vergaren. We zijn ons altijd bewust van alles dat zich voordoet in onze *minds*. Het enige dat zich in onze *minds* voordoet zijn gedachten of dingen die afhankelijk zijn van gedachten. We kunnen daarom geen gedachte hebben waar we ons niet van bewust zijn op het moment dat het zich in ons bevindt (Descartes 1985b, 171). Er zou dus ook geen kloof bestaan tussen de introspectieve verschijning en de realiteit, omdat de realiteit gereduceerd wordt tot de bewuste ervaring van onze gedachten. Dit houdt in dat we aan de hand van introspectie geen fouten kunnen maken. Als onze *mind* identiek is met een bepaalde verzameling aan gedachten, dan is het idee van een medium tussen de *mind* en zijn gedachten onmogelijk. We moeten daarom ook niet zeggen dat de *mind* deze gedachten kent, maar juist dat de *mind* deze gedachten is (Vendler 1972, 191). Alle sensaties, emoties en verlangens kunnen we helder waarnemen, mits we erg voorzichtig zijn in onze oordelen erover. We moeten bij het waarnemen van deze dingen niks meer omvatten dan dat wat strikt genomen in onze perceptie aanwezig is. Dit houdt dus in dat we in ons oordeel niks meer moeten omvatten dan ons innerlijk bewustzijn (Descartes 1985a, 216).

De theorie van Descartes is echter gedateerd. Een meer hedendaagse theorie over introspectie kunnen we vinden bij Goldman. Goldman meent dat het één ding is om een mentale toestand te hebben of te ervaren, maar dat het iets anders is om zo een toestand aan jezelf toe te schrijven, oftewel: om er een overtuiging over te hebben (Goldman 2006, 223). Het lijkt alsof er een soort asymmetrie bestaat tussen eerstepersoons en derdepersoons situaties, in de zin dat eerstepersoons verslagen van mentale toestanden vaak een hogere betrouwbaarheid hebben dan derdepersoons verslagen. Hoe moeten we deze asymmetrie verklaren? Een antwoord dat Goldman hierop geeft is wat hij de ‘speciale methode-benadering’ noemt. Deze benadering komt in een zwakke en sterke vorm. Volgens de zwakke vorm hebben mensen een speciaal proces of methode om hun huidige mentale toestanden te detecteren of om er toegang tot te krijgen die niet gebruikt kan worden voor mentale toestanden van anderen. De sterke variant omvat dit ook, in combinatie met het idee dat deze speciale methode epistemologisch superieur is aan derdepersoons methodes met betrekking tot betrouwbaarheid en rechtvaardiging (Goldman 2006, 224).

Een manier waarop we eerstepersoons verslagen kunnen maken van onze mentale toestanden, is door middel van directe introspectie. De theorie luidt dat er een introspectiesysteem, of proces, is die huidige mentale toestanden kan identificeren door middel van innerlijke herkenning, in plaats van inferentie. Het introspectie ‘orgaan’ is volgens Goldman aandacht; de oriëntatie die een subject in een gepaste relatie stelt met een toestand waar het subject op richt (Goldman 2006, 244). Introspectie is echter een ambigue term. Het kan verwijzen naar een proces van onderzoek, gericht op mentale toestanden, of naar een proces van het beantwoorden van een dergelijk onderzoek.

Volgens de eerste definitie is introspectie binnenwaarts gerichte aandacht, die geselecteerde toestanden kiest om te analyseren. Volgens de laatste definitie is introspectie het proces van het analyseren of classificeren van geselecteerde toestanden (Goldman 2006, 246). Goldman richt zich op deze laatste definitie.

Goldman ziet introspectie als een vorm van perceptie. Beter gezegd: een deel van introspectie moeten we behandelen als perceptie-achtig. Als een deel van introspectie perceptie-achtig is, moet dat deel een transductieproces omvatten. Een transductieproces omvat inputs (gebeurtenissen of eigenschappen waar het proces causaal gevoelig voor is) en outputs (representaties die gegenereerd zijn als antwoord op deze inputs (Goldman 2006, 246). Wat zijn dan precies deze inputeigenschappen voor introspectie en wat zijn de outputs?

De outputs van introspectie zijn representaties van 'token' mentale toestanden die deze classificeren langs een of meer van de volgende dimensies:

1. De algemene categorie van de 'token' toestand (geloof, verlangen, pijn, boosheid, hitte, visuele representatie).
2. De inhoud van die toestand.
3. De sterkte of intensiteit van die toestand. (Goldman 2006, 247)

Zo zou een 'token' visuele toestand geclassificeerd kunnen worden als zijnde van het type *zien* en als hebbende de inhoud "er staan drie vazen op tafel". De visuele toestand zelf heeft natuurlijk de inhoud, of het nou wel of niet geïntrospecteerd is. Met introspectie kan de inhoud echter opnieuw gepresenteerd worden (zonder visueel format). In Goldmans benadering is het inhoud-aspect van introspectie niet perceptie-achtig, maar gaat het te werk via herimplementatie. Introspectie is wel perceptie-achtig op de manier hoe het de toestand waarop het gericht is herkent of classificeert in termen van zijn algemene categorie en in termen van kenmerken als sterkte of intensiteit (Goldman 2006, 247).

Het proces van visuele perceptie omvat meerdere componenten. Zo moet je denken aan het verwerken van vorm, kleur, textuur, oriëntatie, diepte, etc. Deze componenten zijn allemaal van elkaar te onderscheiden. Volgens Goldman zou introspectie ook zo werken, met twee of meer van elkaar te onderscheiden componenten. Op zijn minst een component voor mentale types en een (of twee) andere voor mentale inhouden. De representatie van mentale types wordt bereikt door een perceptie-achtig herkenningsproces, waarbij een gegeven huidige 'token' toestand in kaart wordt gebracht in een mentale categorie die geselecteerd is vanuit een relatief klein aantal types. De representatie van mentale inhouden kan niet op dezelfde manier bereikt worden, omdat er geen

klein aantal van inhoudstypes zijn waarin de inhoud van een token in kaart gebracht kan worden (Goldman 2006, 253).

De inhoud van een mentale toestand introspecteer ik door het te herimplementeren. De inhoud van bijvoorbeeld mijn hoop wordt nagemaakt in de meta-representerende toestand. De meta-representatie herimplementeert niet de hoop omdat het zelf geen hoop is; het is een overtuiging of een oordeel (Goldman 2006, 254).

In deze theorie van Goldman wordt er dus gesteld dat het toeschrijven van mentale toestanden aan jezelf plaatsvindt door middel van introspectie. Een andere theorie stelt dat de speciale methode waarop wij zelfkennis vergaren niet gebeurt via introspectie, maar juist door 'naar buiten' te kijken. Een voorstander van deze theorie is Evans. Hij meent dat in het maken van een zelf-toeschrijving van een overtuiging de ogen buitenwaarts gericht zijn; naar de wereld (Evans 1982, 225). Dit maakt hij duidelijk in de volgende passage:

> "If someone asks me 'Do you think there is going to be a third world war?', I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?'. I get myself in a position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p*." (Evans 1982, 225)

Als een subject deze procedure toepast, zal hij noodzakelijk kennis verkrijgen van een van zijn eigen mentale toestanden. Evans vat deze uitleg in een regel: wanneer je in de positie komt om te beweren dat *p*, ben je *ipso facto* in een positie om 'ik geloof dat *p*' te beweren (Evans 1982, 225-6).

Toch kan zo een procedure niet een volledig begrip geven van de inhoud van het oordeel 'ik geloof dat *p*'. Een volledig begrip van de inhoud van het oordeel moet het bezit van het psychologische concept dat uitgedrukt wordt door 'x gelooft dat *p*' bevatten. Het oordeel kan dit concept omvatten door een waardering van het feit dat de soorten bewijs waartoe het subject in staat is om te herkennen als relevant aan de toeschrijving van het predicaat aan anderen eveneens de waarheid van deze claim bevestigt, met daarbij een bereidwilligheid om bewijs te herkennen dat zij dezelfde procedure hebben uitgevoerd die ten grondslag ligt aan zijn eigen zelf-toeschrijving (Evans 1982, 226).

Je zou kunnen zeggen dat we zonder deze achtergrond geen oprechte 'ik denk (dat *p*)' kunnen veiligstellen die onze gedachte *p* begeleidt. De 'ik denk' welke al onze gedachten begeleidt is puur formeel, maar het toevoegen van de achtergrond maakt geen verschil in de methode van zelf-toeschrijving; we hebben de binnenwaartse blik dus niet nodig, meent Evans (Evans 1982, 226).

1.2.2 Epistemische zekerheid

Zoals we net gezien hebben is een manier waarop wij kennis van onze mentale toestanden verkrijgen 'introspectie'. Zo zouden we onze mentale toestanden kennen door geen aandacht te schenken aan de externe wereld zodat we kunnen waarnemen wat er intern gebeurt. Siewert is het niet eens met deze opvatting, maar gaat wel in op wat deze opvatting zo aantrekkelijk maakt. Volgens hem is eerstepersoons introspectieve kennis mogelijk door middel van bewustzijn en aandacht op ervaring (Siewert 2012, 129).

Volgens Siewert moet er een onderscheid gemaakt worden tussen aan de ene kant dat wat jou in jouw oordeel rechtvaardigt en aan de andere kant dat wat anderen rechtvaardigt in hun oordelen over jou. Dit houdt in dat jij een andere, onafhankelijke eerstepersoons rechtvaardiging hebt voor jouw oordelen over jouw ervaring (Siewert 2012, 130-1).

Ervaring in deze zin houdt hetzelfde in als fenomenaal bewustzijn, wat Siewert definieert als "iets wat het voor iemand is om te hebben". Als ik wil weten wat ik ervaar, hoef ik daar niet achter te komen door iets anders uit te leggen; ik neem mijn ervaring direct waar. Dit idee legt Siewert verder uit door een analogie te trekken met visie. Stel dat je in een doos kijkt waarvan jij alleen de inhoud kan zien. We zouden dan kunnen zeggen dat alleen jij nu kan voelen en waarnemen wat er in jouw *mind* omgaat. Vervolgens kan jij de doos aan mij geven en presenteren op een manier die dezelfde kijk oplevert als dat jij had in de doos, maar er is niks wat er voor zorgt dat ik in jouw *mind* kan kijken op de manier waarop jij dat doet (Siewert 2012, 131).

Ongeacht of het nu visie is of een andere zintuiglijke modaliteit waardoor we iets waarnemen, we kunnen wel een onderscheid aanbrengen tussen dat wat je waarneemt (bijvoorbeeld wat er in de doos zit) en dat wat je erover denkt (bijvoorbeeld of dat wat er in de doos ligt een potlood is). Daarnaast heeft waarnemen een epistemische functie: het voorziet jou met ofwel kennis van het object dat je waarneemt, dan wel een rechtvaardiging voor jouw denken wat je met die waarneming doet. Het belangrijkste punt is wellicht nog wel dat dit onderscheid tussen waarnemen en denken kan worden geïtereerd. Je kunt het uitleggen als iets waar niet alleen iets "zonder de *mind*" aan bod komt, maar ook als iets waarbij mentale toestanden aan bod komen, bijvoorbeeld iemand zijn eigen waarnemen. Dit houdt in dat er dus een eerste-orde waarnemen is (van bijvoorbeeld kleuren, vormen, geuren, etc.) en een waarneming van dit waarnemen (en wellicht van andere mentale toestanden). Dit waarnemen levert ons kennis op via de aandacht die we richten op de waargenomen objecten (Siewert 2012, 132).

Horgan omarmt eveneens de notie van fenomenaal bewustzijn als "iets wat het voor iemand is om te hebben" (Horgan 2012, 405). Hij meent dat het fenomenale karakter zelf-presenterend is voor het

individu dat ervaart. Het hele "wat het is om te hebben" van fenomenaal bewustzijn is iets dat onmiddellijk gegeven wordt in de ervaring. Dit betekent dat er geen sprong wordt gemaakt tussen verschijning en realiteit, omdat de verschijning zelf de realiteit is; hoe het fenomenale karakter voor het individu overkomt, is hoe het is (Horgan 2012, 406). Het zijn volgens hem de intrinsieke fenomenale aspecten van iemands huidige ervaring die epistemisch speciaal zijn in de zin dat ze zelf-presenterend zijn en niet onderhevig zijn aan de kloof tussen verschijning en realiteit (Horgan 2012, 420). Op die wijze verschilt zelfkennis van andere vormen van kennis.

1.2.3 Agentschap

Als iemand aan zichzelf de vraag stelt "geloof ik dat *p*?", dan zal diegene deze vraag behandelen op dezelfde manier als een corresponderende vraag die niet naar hem zelf verwijst, namelijk "is *p* waar?". Dit hebben we zojuist al gezien in de passage van Evans, waarbij een individu op de vraag "geloof ik dat er een derde wereldoorlog aan komt?" antwoordt door de ogen als het ware naar buiten te richten en een antwoord te geven op de vraag "komt er een derde wereldoorlog?". Dit wordt de transparantie van iemands denken genoemd, oftewel de transparantiemethode (Moran 2001, 60).

Zeggen dat een vraag transparant is aan een ander, is niet hetzelfde als het zeggen dat de ene vraag de ander reduceert. De feiten die de vraag over de derde wereldoorlog beantwoorden verschillen van de feiten over iemands persoonlijke geloof. De juiste antwoorden op de twee vragen hoeven niet hetzelfde te zijn. Het is echter niet juist om te zeggen dat de twee vragen onafscheidelijk van elkaar zijn vanuit een eerstepersoons standpunt (Moran 2001, 61). Het zal onder iedereen met het concept van een geloof of overtuiging algemene kennis zijn dat, ook al gelooft iemand iets *als* waar, het feit dat geloofd wordt en het feit van iemands geloof twee verschillende zaken zijn. Vanuit het eerstepersoons standpunt erken ik de twee vragen als verschillend, ten gunste van de erkenning dat waar mijn overtuigingen op gericht zijn een onafhankelijke wereld is, los van mij. Daarom bestaat er ook de mogelijkheid dat mijn overtuigingen niet overeenkomen met die onafhankelijke wereld. Een eerstepersoons vraag over iemands geloof wordt beantwoord door te verwijzen naar dezelfde redenen die een antwoord op de corresponderende vraag over de wereld zouden rechtvaardigen (Moran 2001, 62).

Ten opzichte van overtuigingen, behandel ik de vraag over mijn overtuiging dat *p* dus hetzelfde als de vraag of *p* waar is. Het is volgens de transparantiemethode een vereiste dat ik de vraag over mijn *state of mind* op een deliberatieve manier benader en zelf beslis en verklaar wat ik doe, in plaats van de vraag puur psychologisch te benaderen (Moran 2001, 63).

Moran meent dat ons gevoel van agentschap en verantwoordelijkheid ten opzichte van onze overtuigingen essentieel gekoppeld is aan onze vaardigheid om ze te belijden in overeenstemming met de transparantiemethode: "… only if I can see my own belief as somehow "up to me" will it make sense for me to answer a question as to what I believe about something by reflecting exclusively on that very thing …" (Moran 2001, 66-7). Verantwoordelijkheid en agentschap aangaande onze eigen overtuigingen en andere houdingen hangen af van het vormen en het hebben ervan, op een manier die gevoelig is voor redenen voor en tegen hen. Een individu is een *agens* ten opzichte van zijn overtuigingen door te reflecteren op wat waar is, of door zich te oriënteren op de vraag over zijn verlangens door te reflecteren op wat de moeite waard is, wat bevredigend is, of wat juist niet. Er is sprake van een rol van een *agens* in zoverre we spreken over iemands verantwoordelijkheid voor zijn houdingen (Moran 2001, 64). Gevoeligheid en responsiviteit voor redenen zijn dus essentieel voor de soort agentschap en verantwoordelijkheid die betrokken zijn bij zelfkennis.

Moran meent dat we veeleer een deliberatieve houding hebben ten opzichte van onze *state of mind* dan een theoretische: "The phenomena of self-knowledge … are themselves based as much in asymmetries of responsibility and commitment as they are in differences in capacities, or in cognitive access" (Moran 2001, 64). Houdend aan het idee van de transparantie tussen zelfgericht en wereldgericht onderzoek lijkt minder een zaak te zijn van de logica van zelfverwijzing en juist meer een zaak van het aannemen van een bepaalde houding ten opzichte van jezelf en je mentale houdingen (Moran 2001, 64). We moeten onszelf dus zien als een auteur van gedachten, in plaats van iets waarin gedachten zich slechts voordoen. Het is precies deze agentschap dat ervoor zorgt dat we in staat zijn om zelfkennis te hebben.

2 Zelfbedrog

Na een beter begrip te hebben gegeven over zelfkennis en waarom het verschilt van andere vormen van kennis, zal ik het in dit hoofdstuk hebben over het fenomeen zelfbedrog: hoe het gedefinieerd wordt, hoe deze definities leiden tot verschillende benaderingen, over de aard en uiteindelijk over de oorzaken van zelfbedrog.

2.1 Twee paradoxen omtrent de traditionele opvatting

Zelfbedrog als het onszelf voor de gek houden, door onszelf iets te laten geloven waarvan we willen dat het waar is, klinkt voor de meeste mensen als een herkenbaar fenomeen. Toch bestaat er onenigheid over de manier waarop we zelfbedrog moeten definiëren. Mele benoemt drie manieren waarop zelfbedrog doorgaans gedefinieerd wordt.

Ten eerste kunnen we zelfbedrog lexicaal definiëren: hierbij start een theoreticus met een definitie van 'bedriegen' of 'bedrog' en gebruikt deze definitie vervolgens als een model voor het definiëren van zelfbedrog. Een tweede manier is zelfbedrog definiëren op basis van voorbeelden: hierbij onderzoekt een theoreticus representatieve voorbeelden van zelfbedrog om er vervolgens essentiële gemeenschappelijke kenmerken uit af te leiden. De derde manier is een theorie-geleide definitie van zelfbedrog: hierbij wordt de zoektocht naar een definitie geleid door een *commonsense* theorie over de etiologie en de aard van zelfbedrog (Mele 2001, 5).

Traditioneel gezien wordt in het definiëren van zelfbedrog de voorkeur gegeven aan de lexicale benadering, voornamelijk door theoretici die ontkennen dat zelfbedrog mogelijk is. Zelfbedrog zou namelijk onmogelijk zijn als we het werkwoord 'bedriegen' als uitgangspunt zouden nemen. Dat wordt duidelijk door de volgende aannames omtrent 'bedriegen':

1. Volgens de definitie van bedriegen bedriegt persoon A persoon B (waarbij B wel of niet dezelfde persoon als A is) door hem te laten geloven dat *p* alleen als A weet, of op zijn minst oprecht gelooft, dat ~*p* en ervoor zorgt dat B gelooft dat *p*.
2. Volgens de definitie van bedriegen is bedriegen een intentionele activiteit: niet-intentioneel bedriegen is conceptueel onmogelijk. (Mele 2001, 6)

Deze aannames brengen problemen met zich mee als we op basis hiervan zelfbedrog willen definiëren. Als aanname 1 klopt, dan vereist jezelf bedriegen in het geloven dat *p* dat je weet, of op zijn minst oprecht gelooft, dat ~*p* en dat je ervoor zorgt dat je gelooft dat *p*. Zo kun je dus stellen dat iemand begint met het geloven dat ~*p* en dat diegene het op een of andere manier voor elkaar krijgt om te geloven dat *p*. Sommige theoretici stellen dat, op een zeker moment, deze zelfbedrieger tegelijkertijd gelooft dat *p* en ~*p*. Dit zou, zo wordt geclaimd, niet mogelijk zijn: de aard van een

overtuiging verhindert dat iemand tegelijkertijd kan geloven dat *p* waar is en dat *p* niet waar is. Dit noemt Mele het 'statische' probleem van zelfbedrog: zelfbedrog vereist, volgens deze benadering, dat iemand in een onmogelijke *state of mind* verkeert (Mele 2001, 6-7).

Aanname 2 zou daarentegen een 'dynamisch' probleem met zich meebrengen. Aan de ene kant is het moeilijk voor te stellen hoe een persoon iemand anders kan bedriegen in het geloven dat *p* als de ander precies weet wat deze persoon van plan is. Ook is het moeilijk voor te stellen hoe dit makkelijker wordt als de bedrieger en de bedrogene een en dezelfde persoon zijn. Aan de andere kant wordt bedrog doorgaans gefaciliteerd doordat de bedrieger er een strategie voor heeft die hij intentioneel probeert uit te voeren. Hoe kan iemand slagen in zichzelf bedriegen als diegene weet dat hij zichzelf gaat bedriegen (Mele 2001, 8)?

De moeilijkheid zit 'm erin om te verklaren dat zelfbedrog in het algemeen een psychologisch mogelijk proces is. Als zelfbedriegers zichzelf intentioneel bedriegen, kun je je afvragen wat die intentie verhindert om zijn eigen functioneren te ondermijnen. Als zelfbedrog niet intentioneel is, wat motiveert en stuurt dan de processen van zelfbedrog (Mele 2001, 8)? Veel voorkomende voorbeelden van zelfbedrog bevatten mensen die onjuist geloven – in aanwezigheid van bewijs dat sterk wijst op het tegenovergestelde – dat bijvoorbeeld hun partners niet vreemdgaan, dat hun kinderen geen drugs gebruiken, of dat ze zelf niet ernstig ziek zijn. Het is nog maar de vraag of in deze gevallen zelfbedrog vereist dat iemand zichzelf intentioneel bedriegt; dat iemand zichzelf overtuigt van iets waarvan diegene op een eerder moment wist dat dat niet waar was (Mele 2001, 9).

Theoretici die de lexicale aannames accepteren kunnen op twee manieren te werk gaan. Ze kunnen aannemen dat veel gevallen van zelfbedrog niet tellen als zelfbedrog omdat deze er niet in slagen om aan een of beide aannames te voldoen. Daarnaast kunnen ze aannemen dat alle of de meeste gevallen die doorgaans worden gezien als situaties waarin zelfbedrog plaatsvindt wel aan de lexicale aannames voldoen, ook al lijkt dat op het eerste gezicht niet zo (Mele 2001, 9). Theorieën die hierop gebaseerd zijn worden intentionalistische theorieën genoemd. Een alternatieve manier om zelfbedrog te definiëren is om beide aannames te ondermijnen en te laten zien dat ze niet zorgen voor een goed begrip van zelfbedrog. Theorieën die op deze gedachte zijn gebaseerd worden motivationalistische genoemd. Dit is de kant die Mele op wil gaan (Mele 2001, 8).

2.2 Intentionalisme en Motivationalisme

2.2.1 Intentionalisme: het oplossen van de paradoxen

Wanneer een theoreticus de lexicale aannames gebruikt om zelfbedrog te definiëren, neemt die theoreticus ook de taak op zich de paradoxen die deze aannames met zich meebrengen op te lossen.

Een manier om dit te doen wordt gegeven door Sorensen. Hij wil de paradox van zelfbedrog oplossen door een analogie te trekken met de paradox van 'de moord', zoals hij dat noemt. Beide paradoxen kunnen worden opgelost door ons meer bekend te maken met het concept van een over tijd verspreide gebeurtenis. Wat deze paradox van 'de moord' precies inhoudt, legt Sorensen uit aan de hand van de volgende passage:

> "On September 9th, 1935, Carl Austin Weiss, a bright young surgeon, shot Senator Huey Long in the Louisiana State Capitol with a .35 caliber pistol. Long was to die from this wound thirty hours later on September 10th. Weiss, on the other hand, received between thirty-two and sixty .44 and .45 calibre hollow point bullets from Long's agitated bodyguards. Thus Long died later than his assassin. The philosophical question raised by this violent episode of mid-Depression Louisiana politics is 'When did Weiss kill Long?'". (Sorensen 1985, 64)

Het lijkt erop dat:

1. Weiss Long vermoordde wanneer hij op hem geschoten had, of toen Long overleed.

Er zijn echter bezwaren te bedenken bij beide mogelijkheden. Neem bijvoorbeeld deze proposities die het eerste alternatief tegenspreken:

2. Als Weiss Long vermoordde toen hij hem had beschoten, was er een periode waarin Long vermoord was terwijl hij nog dertig uur leefde.
3. Als iemand op tijd *t* is vermoord, is diegene meteen dood na *t*.

Ten tweede moeten we op basis van de volgende proposities de andere mogelijkheid afwijzen:

4. Als Weiss Long had vermoord toen Long overleed, is Long vermoord door een dode man.
5. Dode mensen kunnen geen handelingen uitvoeren.

Propositie 1 t/m 5 is een set plausibele maar tezamen inconsistente proposities. Vandaar dat het een paradox is, welke Sorensen de paradox van de moord noemt. Om de paradox op te lossen moeten we laten zien dat een van deze proposities niet klopt (Sorensen 1985, 64).

De paradox van zelfbedrog ontstaat uit een vereiste dat iemand een slachtoffer is en een vereiste dat iemand een bedrieger is, zoals we in de vorige sectie al hebben kunnen zien. Als de bedrieger en de bedrogene een en dezelfde persoon zijn, zou die persoon twee conflicterende overtuigingen moeten hebben, wat onmogelijk wordt geacht. De notie van 'bedrog' in zelfbedrog wordt begrepen zoals het begrepen wordt in gevallen van 'een-ander-bedriegen' – oftewel, gevallen waarin een persoon een ander bedriegt. Sorensen meent dat we 'jezelf bedriegen' moeten begrijpen op dezelfde manier als

we iets ‘jezelf aanleren’ of ‘jezelf uitnodigen’ begrijpen. Als we de logica aanhouden dat de bedrieger en de bedrogene dezelfde persoon zijn, zouden we onszelf ook nooit iets aan kunnen leren of onszelf uit kunnen nodigen (Sorensen 1985, 66).

De paradox van zelfbedrog is in deze aspecten gelijk aan de paradox van de moord aangezien beide symptomen zijn van onze onbekendheid met het concept van over tijd verspreide gebeurtenissen. Het vermoorden van Long is een verspreide gebeurtenis bestaande uit de subgebeurtenissen van het schieten van Weiss en Longs dood. Hoewel de gebeurtenis bestaat uit delen die plaatsvinden op verschillende tijdstippen, betekent dat niet dat het een veranderende gebeurtenis is (Sorensen 1985, 67).

Sorensen stelt voor dat veel van onze handelingen verspreide gebeurtenissen zijn. Denk bijvoorbeeld aan typen, koken, eten, stofzuigen, etc. Dit zijn allemaal over tijd verspreide gebeurtenissen. Denk daarnaast bijvoorbeeld aan het lekken van een kraan. Als er geen tijd tussen de lekkende druppels zit, zou dat betekenen dat de kraan aan staat en niet lekt. Je kunt het hier natuurlijk mee eens zijn, maar je tegelijkertijd nog steeds afvragen wanneer Long was vermoord (Sorensen 1985, 68). Eveneens kun je je bijvoorbeeld afvragen waar het Drielandenpunt ligt, terwijl je wel kan erkennen dat het een ruimtelijk verspreid object is. Hiervoor stelt Sorensen een principe op voor het lokaliseren van objecten:

(S) Een object bevindt zich op plaats *s* alleen als *s* de plaatsen van zijn delen overlapt.

Volgens dit principe zou het antwoord ‘het Drielandenpunt ligt in Nederland’ fout zijn, terwijl het wel antwoorden toestaat zoals ‘het Drielandenpunt ligt in het Europa’. Dit principe is analoog aan het principe met betrekking tot het dateren van gebeurtenissen:

(T) Een gebeurtenis vindt plaats op tijdstip *t* alleen als *t* de tijden van zijn delen overlapt.

Principe (T) verklaart waarom de vraag ‘wanneer vermoordde Weiss Long?’ onproblematische antwoorden zoals ‘in 1935’ of ‘tijdens Longs ambtsperiode als senator’ kent. Daarentegen zijn antwoorden zoals ‘Weiss vermoordde Long toen hij hem beschoot’ en ‘Weiss vermoordde Long toen Long overleed’ niet compatibel met dit principe (Sorensen 1985, 68).

Dezelfde redenatie biedt ons een antwoord op de paradox van zelfbedrog. Bedrog is een complexe gebeurtenis bestaande uit de initiatie van bedrog en het voor het slachtoffer verkrijgen van de overtuiging waarop gedoeld wordt. Op basis hiervan kunnen we de eerste lexicale aanname, waarbij de zelfbedrieger twee contradictoire overtuigingen heeft, begrijpen als volgt:

6. Als persoon A persoon B bedriegt op tijdstip *t* door te laten geloven dat *p*, dan gelooft B dat *p* op het moment dat het bedriegen is voltooid.
7. Alf persoon A persoon B bedriegt op tijdstip *t* door te laten geloven dat *p*, dan is het op het moment dat het bedriegen start niet het geval dat A gelooft dat *p*.

Volgens deze uitleg is de bedrieger iemand die ervoor zorgt dat zijn slachtoffer gelooft dat *p* ondanks het feit dat de bedrieger zelf niet gelooft dat *p* wanneer hij begint met het bedriegen (Sorensen 1985, 68-9).

Wat we uit dit voorgaande kunnen opmaken is het inzicht dat achter het probleem van zelfbedrog ligt, namelijk het inzicht dat zelfbedrog niet onmiddellijk kan zijn. Zelfbedrog is veeleer een gradueel proces. Er moet voldoende tijd zitten tussen het begin en het einde van het bedriegen om in staat te zijn een overtuiging te verkrijgen. Dit impliceert niet een ontkenning van het feit dat de bedrieger tegelijkertijd de bedrieger en de bedrogene is. Bedrog is een complexe en meestal verspreide gebeurtenis, waarbij we vaak een deel van de gebeurtenis door elkaar halen met de volledige gebeurtenis. Nu we meer weten over verspreide gebeurtenissen zijn we in staat onderscheidingen aan te brengen tussen de gebeurtenis en zijn tijdelijke delen. Wanneer deze onderscheidingen gemaakt zijn, worden de fouten in het dateren verholpen en zijn de paradoxen opgelost (Sorensen 1985, 69). Het hebben van contradictoire overtuigingen kan volgens deze uitleg dus worden verklaard door te stellen dat een persoon op tijdstip *t1* de overtuiging heeft dat *~p* en op tijdstip *t2* de overtuiging heeft dat *p*.

2.2.2 Motivationalisme: het verwerpen van de aannames

Mele stelt dat de theorie van over tijd verspreide gebeurtenissen niet een onsamenhangende theorie is aangezien het een oplossing biedt voor de paradox van zelfbedrog, maar wel dat het een ongerechtvaardigde theorie is omdat het slechts uitzonderlijke gevallen van zelfbedrog verklaart (Mele 2001, 17). Mele verwerpt de lexicale aannames van zelfbedrog en stelt daarvoor in de plaats een ander model van zelfbedrog voor.

Een belangrijk punt van Mele is dat zelfbedrog doorgaans begrepen wordt als een gemotiveerd fenomeen. Ons verlangen naar *p* kan bijvoorbeeld op verschillende manieren bijdragen aan ons geloven dat *p* in gevallen van rechtstreekse zelfbedrog. Hier geeft hij vier voorbeelden voor:

1. *Negatieve misinterpretatie*. Ons verlangen naar *p* kan ertoe leiden dat we bewijs ten nadele van *p* niet meewegen, terwijl we die normaal gesproken wel zouden tellen in de afwezigheid van dit verlangen.

2. *Positieve misinterpretatie*. Ons verlangen naar *p* kan ertoe leiden dat we bewijs interpreteren ten voordele van *p*, terwijl we die we normaal gesproken niet zouden tellen in de afwezigheid van dit verlangen.
3. *Selectieve aandacht*. Ons verlangen naar *p* kan ertoe leiden dat het ons niet lukt de aandacht te houden bij bewijs ten nadele van *p* en dat we de aandacht juist richten op bewijs dat ten voordele van *p* is.
4. *Selectieve bewijsverzameling*. Ons verlangen naar *p* kan ertoe leiden dat we makkelijk te verkrijgen bewijs voor *~p* over het hoofd zien en dat we bewijs vinden voor *p* dat minder toegankelijk is. (Mele 2001, 26-7)

Vervolgens is het de vraag hoe een verlangen naar *p* de fenomenen in deze voorbeelden triggeren en uiteindelijk tot vooringenomen overtuigingen dat *p* leiden. Overtuigingen die worden gekarakteriseerd door zelfbedrog zijn namelijk een soort vooringenomen overtuigingen, die tevens gemotiveerd zijn. Mele noemt drie bronnen voor deze motivatie:

1. *Helderheid van informatie*. De helderheid van een stuk bewijs is voor een individu vaak een functie van individuele interesses, de concreetheid van het stuk bewijs of de zintuiglijke, tijdelijke of ruimtelijke nabijheid ervan. Hoe helderder het bewijs, hoe groter de kans is dat het individu het zal gebruiken.
2. *De heuristiek van beschikbaarheid*. Wanneer we overtuigingen vormen over de frequentie, kans, of oorzaken van een gebeurtenis, worden we vaak beïnvloed door de relatieve beschikbaarheid van de objecten of gebeurtenissen, oftewel, hun beschikbaarheid in het proces van perceptie, geheugen of verbeelding.
3. *De vooringenomenheid van bevestiging*. Mensen die een hypothese testen zijn vaker geneigd te zoeken naar bewijs dat hun hypothese bevestigd dan bewijs te vinden dat hun hypothese ontkracht en zijn eveneens geneigd dit eerste sneller te herkennen. (Mele 2001, 28-9)

Het kan natuurlijk zo zijn dat het meest heldere of beschikbare bewijs soms ook de grootste bewijskracht heeft; de invloed van zulk bewijs heeft niet altijd een vooringenomen invloed. Het punt zit ‘m erin dat, hoewel bronnen van een vooringenomen overtuiging onafhankelijk van elkaar kunnen functioneren, ze ook getriggerd en behouden kunnen worden door motivatie in het produceren van individuele gemotiveerde vooringenomen overtuigingen (Mele 2001, 29).

Het model dat Mele wil bieden is een model dat volgens hem een betere uitleg biedt van zelfbedrog dan de aannames omtrent interpersoonlijk bedrog. Dit model noemt hij het FTL model; een model dat een combinatie is van ideeën van Friedrich, Trope en Liberman. Hij combineert het idee van Friedrich dat zelfbedrog zich voordoet bij het genereren van hypotheses (Mele 2001, 31-4) met het

idee van Trope en Liberman dat zelfbedrog zich voordoet bij de evaluatie van hypotheses (Mele 2001, 34-5). Met andere woorden: zelfbedrog doet zich voor wanneer verlangens en andere motivaties bepalen *welke* hypotheses ik test en *hoe (grondig)* ik ze test.

Een centraal element in het model van Trope en Liberman is de notie van een 'drempel'. Hoe lager de drempel, hoe minder sterk het bewijs dat vereist is om die drempel te halen. Twee drempels zijn relevant voor elke hypothese. De acceptatiedrempel is de minimale hoeveelheid zekerheid die vereist is voor de waarheid van een hypothese om die hypothese te accepteren, in plaats van door te gaan met testen. De afwijzingsdrempel is de minimale hoeveelheid zekerheid die vereist is voor de onwaarheid van een hypothese om die hypothese af te wijzen en te beëindigen (Mele 2001, 34).

Het genereren van een hypothese die iemand zal testen vereist alleen dat die hypothese in diegene opkomt. De evaluatie van een hypothese omvat een begrip van zijn implicaties, informatieverzameling, interpretatie en categorisatie van het verzamelde bewijs, en gevolgtrekkingen maken over de (on)waarheid van de hypothese op basis van het geïnterpreteerde bewijs (Mele 2001, 37). Mele geeft een voorbeeld van een gemotiveerd vooringenomen evaluatie:

> "Bob wants it to be true that he is the best third baseman in his league. Owing partly to that desire, he has a lower threshold for believing that he is than for believing that he is not. Bob examines the statistics on the competition and decides, correctly, that his main competitor is Carl. Bob and Carl have the same fielding percentage, but Carl has a few more home runs … , several more runs batted in … , and a higher batting average … . However, Carl's team is much better than Bob's, and, as Bob knows, players on better teams tend to have more opportunities to bat in runs … and to hit home runs … . Bob takes all this and more into account and comes to believe that he is a better player than Carl. As it turns out, however, a panel of experts properly decides that Carl is the better player … . They too take account of the fact that Carl's team was far superior to Bob's, but they also notice that Carl batted many fewer times than Bob … . And they are impressed that, given this statistic, Carl still outperformed Bob in home runs and runs batted in." (Mele 2001, 37-8)

Volgens het FTL model kunnen we verklaren dat het feit dat Bob tot deze conclusie komt voor een groot deel te danken is aan het feit dat hij een lagere drempel heeft voor het geloven dat hij beter is dan Carl, dan andersom. Dit laatste is grotendeels te verklaren doordat Bob het verlangen heeft dat hij de betere speler is. Gegeven het verschil in drempels kunnen we constateren dat voor Bob het verkrijgen van de overtuiging dat hij een betere speler is mindere bewijskracht vereist dan het verkrijgen van de overtuiging dat dit niet het geval is. Het feit dat Bob herkent dat zijn statistieken bijna net zo goed zijn als die van Carl, in combinatie met zijn observatie dat spelers in betere teams

geneigd zijn meer kansen te hebben, is voor hem genoeg om de drempel te halen voor de overtuiging dat hij de betere speler is. Door dit gedaan te hebben, neemt Bob niet eens meer de mogelijkheid in acht dat Carl minder speelminuten heeft gemaakt (Mele 2001, 38).

Bobs overtuiging dat hij een betere speler is dan Carl lijkt haast zeker een gemotiveerd vooringenomen overtuiging. Er is eveneens geen noodzaak om aan te nemen dat Bob het voor zichzelf makkelijker heeft gemaakt om tot deze overtuiging te komen. Het feit dat hij tot deze overtuiging is gekomen kan verklaard worden door het FTL model (Mele 2001, 38). Hiervoor hoeven we dus niet uit te gaan van de lexicale aannames omtrent interpersoonlijk bedrog.

Een mogelijk probleem met Mele's theorie is echter dat het niet uitlegt hoe een persoon selectief bewijs kan vermijden voor de overtuiging dat *p* zonder op een manier te geloven dat *p*. Hoe weet deze persoon anders welk bewijs hij moet vermijden? Een mogelijk antwoord op dit probleem geef ik in de volgende sectie over zelfbedrog en confabulatie.

2.3 Zelfbedrog en confabulatie

Om een beter begrip te krijgen van de aard van zelfbedrog, ga ik in deze sectie in op een fenomeen dat gerelateerd is aan zelfbedrog: confabulatie. Confabulatie is een verschijnsel waarbij een persoon onjuiste verklaringen geeft (over bijvoorbeeld een gebeurtenis die heeft plaatsgevonden) zonder de intentie te hebben om te liegen. Dit verschijnsel komt voornamelijk voor onder patiënten met het gespleten-hersenen syndroom, anosognosie en Korsakoff's syndroom (Hirstein 2000, 420).

Gespleten-hersenen syndroom houdt in dat de verbinding tussen twee hersenhelften afgesneden is. Dit vindt meestal plaats door middel van een chirurgische operatie om te voorkomen dat epileptische aanvallen van de ene naar de andere hersenhelft kunnen verspreiden. Uit onderzoeken van Gazzaniga (Gazzaniga 1995) en Sperry (Sperry 1985) naar patiënten met het gespleten-hersenen syndroom, bleek dat wanneer één van de hersenhelften gestimuleerd werd, het leek alsof er twee personen in één lichaam zaten. De linker hersenhelft is alleen in staat om verbale reacties te geven, terwijl de rechter hersenhelft talige input kan begrijpen en kan reageren door naar afbeeldingen te wijzen met de linkerhand (hersenhelften hebben controle over de arm die aan de tegenovergestelde zijde zit). In het onderzoek bleek dat wanneer een patiënt gevraagd werd over de activiteit van zijn linkerhand, de linker hersenhelft antwoordde alsof het controle had over de linkerhand, terwijl het door de operatie het geval was dat de linker hersenhelft geen idee had waarom de linkerhand deed wat het deed (Gazzaniga 1995, Sperry 1985).

Anosognosie houdt een onbewustzijn of ontkenning van ziekte in. Het wordt meestal veroorzaakt door schade in een bepaald deel van de rechter hersenhelft, wat kan resulteren in een verlamming

van de linkerarm of de gehele linkerhelft van je lichaam. Deze verlamming kan samengaan met verwaarlozing: de patiënt negeert de linkerhelft van zijn lichaam en de nabije ruimte. Dit uit zich doordat de patiënt bijvoorbeeld de linkerhelft van zijn bord niet opeet of de linkerhelft van zijn lichaam niet wast (Hirstein 2000, 420). Wanneer een dergelijke patiënt gevraagd wordt om met zijn linkerhand zijn neus aan te raken, zal hij dit tevergeefs proberen. Vaak komt het echter voor dat de patiënt begint te confabuleren door dingen zoals bijvoorbeeld “ik heb daar nu even geen zin in” of “ik zou het doen als ik geen last had van artritis” te uiten (Ramachandran 1995). Een groot gedeelte van de patiënten meent zelfs oprecht dat het hen gelukt is om met hun linkerhand hun neus aan te raken en stellen dat ze het zichzelf hebben zien doen (Hirstein 2000, 421).

Korsakoff’s syndroom is een vorm van geheugenverlies, meestal veroorzaakt door overmatig drankgebruik. Het geheugenverlies heeft betrekking op episodisch geheugen (hoe bepaalde gebeurtenissen hebben plaatsgevonden), maar niet op semantisch geheugen (zoals de kennis van concepten en betekenissen) (Tulving 1983). Het geheugenverlies is zodanig ernstig dat de patiënt meestal geen herinnering meer heeft van de dingen die hij gisteren heeft gedaan. Wanneer daar echter naar gevraagd wordt, produceert de patiënt vaak een gedetailleerde omschrijving van plausibel klinkende gebeurtenissen die hij in feite allemaal ter plekke verzint (Hirstein 2000, 421).

Geen van de patiënten geven een teken dat ze zich bewust zijn van wat ze aan het doen zijn: ze zijn blijkbaar niet aan het liegen en geloven hun verzonnen verhalen oprecht, waarbij ze ook nog eens een enorme zekerheid uitstralen (Gazzaniga 1995, Sperry 1985, Ramachandran 1995).

Waar de linker hersenhelft talige vaardigheden bevat, heeft de rechter hersenhelft de vaardigheid om informatie op te kunnen pikken. Dit verwerken van informatie gebeurt via twee routes of stromen, namelijk de ‘wat’-stroom en de ‘waar’-stroom. De ‘wat’-stroom heeft de functie om objecten te kunnen identificeren en de ‘waar’-stroom heeft de functie om de nabije visuele ruimte te kunnen representeren, zodat het subject er zich in kan navigeren, kan reiken naar objecten, etc. (Hirstein 2000, 422).

Uit onderzoek is gebleken dat de linker hersenhelft meer geneigd is naar de ‘wat’-stroom zodat het snel talige labels kan plakken op binnenkomende informatie. Dit houdt een soort digitalisering van informatie in wat in eerste instantie alleen nog in analoge vorm aanwezig is (namelijk visuele informatie). Aan de andere kant lijkt het verwerken van informatie in de rechter hersenhelft te neigen naar de ‘waar’-stroom. De binnenkomende informatie wordt opgeslagen als analoge informatie, in plaats van dat de informatie meteen vertaald wordt naar een conceptuele of talige vorm (Hirstein 2000, 423).

Dit verklaart waarom patiënten met Korsakoff's syndroom moeite hebben om eerder plaatsgevonden gebeurtenissen te herproduceren. Het autobiografisch geheugen heeft de vorm van een soort videoherhaling: ik herinner mij de gebeurtenissen van gisteren vanuit mijn standpunt, hoe ik ze heb ervaren. Dit is een analoge vorm van representatie. Confabulatie lijkt in dit geval voor te komen wanneer het analoge representatiesysteem verstoord wordt of geïsoleerd is van het conceptuele representatiesysteem in de linker hersenhelft (Hirstein 2000, 423). Daarnaast kunnen analoge representatiesystemen informatie belichamen over iets dat nog niet gedefinieerd is. Het conceptuele systeem kan dit niet, want deze krijgt informatie die geprepareerd is door het analoge systeem. Op het niveau van het conceptuele systeem worden gaten in de analoge representatie gevuld om deze representaties compleet te maken (Hirstein 2000, 424).

Nu gaan we terug naar zelfbedrog: hoe kan het dat ik mijzelf wijs kan maken dat *p* terwijl ik weet dat niet-*p*? Een moeilijkheid kan zitten in het feit dat het lastig is om voor te stellen dat informatie in het analoge representatiesysteem bestaat uit overtuigingen. Neem dit voorbeeld:

> "Suppose I were to enter my office one day and find a huge chasm in the floor. I would be surprised now, does this mean that I believe that my office floor is solid, or worse, that I have a belief that there is no chasm in the middle of my office floor?" (Hirstein 2000, 427)

Intuïtief gezien lijkt er iets mis te zijn met het toeschrijven van dit soort overtuigingen. Hirstein stelt voor dat dit is omdat, terwijl ik mijn kantoorvloer representeer als vast, ik het in analoge vorm representeer, in plaats van in de vorm van een uitgesproken overtuiging. Zelfbedrog kan in dit geval plaatsvinden wanneer het conceptuele systeem een overtuiging dat niet *p* bevat, terwijl het analoge systeem trouw blijft representeren dat *p* (Hirstein 2000, 427). Dit kan een oplossing zijn op het in de vorige sectie gestelde probleem met Mele's theorie.

Een manier om dit op te lossen is als volgt: de informatie dat *p* is reeds gerepresenteerd in het analoge systeem. Er is een soort conflict in iemands *mind*, maar de conflicterende informatie wordt gerepresenteerd in twee verschillende vormen: conceptueel en analoog. Dit is anders dan het hebben van twee contradictoire overtuigingen die beide conceptueel zijn. Wat er dus gebeurt is dat het brein op een bepaalde manier voorkomt dat bepaalde soorten van analoge informatie gerepresenteerd worden in conceptuele vorm. Zelfbedrog is te herleiden tot dit fenomeen (Hirstein 2000, 428).

In Mele's theorie hebben we kunnen zien dat een manier waarop mensen een ongewenste overtuiging weghouden mogelijk is door selectieve aandacht. Een verschil tussen het 'wat'-georiënteerde conceptuele systeem in de linker hersenhelft en het 'waar'-georiënteerde analoge

systeem in de rechter hersenhelft is dat het 'wat'-systeem voornamelijk input krijgt uit focaal/centraal zicht, terwijl het 'waar'-systeem input krijgt uit perifeer zicht. Het confabulatoire conceptuele systeem kan selectief informatie negeren door de ogen en de aandacht op iets anders te richten (Hirstein 2000, 428).

Over het algemeen is het verschil tussen iemand die confabuleert en iemand die zichzelf bedriegt dat degene die zichzelf bedriegt een conflict ervaart vanwege de conflicterende informatie in beide systemen. Degene die confabuleert heeft dit echter niet door, omdat het analoge systeem die deze conflicterende informatie normaal gesproken zou representeren beschadigd is. Iemand die confabuleert is daarnaast erg zelfverzekerd in zijn antwoorden, terwijl iemand die zichzelf bedriegt dit minder zal zijn (Hirstein 2000, 428).

Deze benadering van Hirstein is gegrond in experimentele bevindingen en verklaart het idee dat conflicterende informatie in de *mind* van iemand die zichzelf bedriegt op een bepaalde manier gescheiden is. Het conflict is dus niet tussen twee overtuigingen, maar veeleer tussen twee soorten representaties, namelijk een conceptuele en een analoge.

2.4 Mogelijke oorzaken zelfbedrog

In deze sectie zal ik verschillende mogelijke oorzaken van zelfbedrog beschouwen. Zo zijn mogelijke oorzaken van zelfbedrog dat het goed zou zijn voor je mentale gezondheid, dat het ons in staat stelt beter te worden in het bedriegen van anderen, of dat zelfbedrog voorkomt onder onze soort omdat dat nou eenmaal is hoe onze hersenen werken.

2.4.1 Mentale gezondheid

Binnen de psychologie is het lang een traditie geweest om te denken dat een goede grip op de realiteit belangrijk is voor je mentale gezondheid. Taylor en Brown menen echter dat bepaalde illusies juist goed zijn voor je mentale gezondheid en je algemene welzijn. Specifiek zijn er drie soorten illusies te benoemen, namelijk: een onrealistisch positief zelfbeeld, de illusie van controle en onrealistisch optimisme (Taylor & Brown 1988, 193)

Uit onderzoek blijkt dat de meeste mensen een positief beeld van zichzelf hebben. Wanneer deelnemers van het onderzoek werd gevraagd hoe accuraat positieve en negatieve persoonlijke bijvoeglijke naamwoorden hun zelf beschreef, meenden de meesten dat de positieve persoonlijke bijvoeglijke naamwoorden hen als persoon beter beschreven dan negatieve. Voor de meeste mensen is het ook nog eens zo dat positieve persoonlijke informatie efficiënt te verwerken is en makkelijk te onthouden is, terwijl negatieve persoonlijke informatie dat juist niet is. Het feit dat de meeste

mensen positief over zichzelf denken hoeft overigens nog niet te betekenen dat dit onrealistisch of illusionair is (Taylor & Brown 1988, 195).

Toch is er bewijs beschikbaar dat het zelfbeeld van deze mensen wel een illusie is. Ten eerste zijn mensen geneigd zichzelf als beter te beschouwen dan anderen. Zo oordelen ze dat positieve persoonlijke kenmerken hen beter beschrijven dan de gemiddelde persoon en menen ze dat negatieve persoonlijke kenmerken hen slechter beschrijven dan de gemiddelde persoon. Uit meerdere onderzoeken blijkt dat dit geldt voor een breed scala aan kenmerken en vaardigheden. Omdat het logisch gezien onmogelijk is dat de meeste mensen beter zijn dat het gemiddelde, kan dit worden beschouwd als bewijs voor de onrealistische en illusionaire aard van deze percepties (Taylor & Brown 1988, 195).

Een andere bron van bewijs komt uit onderzoeken waarin de beoordelingen van mensen over zichzelf werden vergeleken met externe waarnemers. Deze waarnemers kregen de taak om een groep studenten te zien werken aan een groepsopdracht, waarna ze hen moesten beoordelen aan de hand van verschillende persoonlijkheidskenmerken (zoals vriendelijk, assertief, warm, etc.). De studenten kregen deze taak eveneens, maar dan voor zichzelf. De uitkomst liet zien dat de studenten zichzelf significant positiever hadden ingeschat dan dat de waarnemers dat hadden gedaan. Waar het dus op neer komt, is dat de meeste mensen zichzelf beschouwen als beter dan anderen en als beter dan dat anderen hen beschouwen. Vanwege deze redenen kunnen we hun denkbeelden als illusionair beschouwen (Taylor & Brown 1988, 195-6).

Een tweede domein waarin de meeste mensen onrealistische denkbeelden hebben is in hun controle over bepaalde gebeurtenissen. Uit meerdere onderzoeken blijkt dat mensen vaak denken dat ze controle hebben over een situatie terwijl die in feite worden bepaald door kans. De meeste mensen denken bijvoorbeeld dat ze meer controle als ze zelf een dobbelsteen gooien dan als iemand anders dat voor ze doet. Wanneer mensen een bepaalde uitkomst verwachten en de uitkomst vindt plaats, overschatten mensen vaak de mate waarin ze daadwerkelijk hebben meegespeeld in het realiseren van die uitkomst (Taylor & Brown 1988, 196)

Een derde domein is onrealistisch optimisme over de toekomst. Onderzoek wijst uit dat de meeste mensen geloven dat het heden beter is dan het verleden en dat de toekomst zelfs beter zal zijn. Toen er aan studenten werd gevraagd wat zij dachten dat mogelijk zou zijn voor hen in de toekomst, was de uitkomst dat zij vier keer zo veel positieve dan negatieve mogelijkheden noemden (Taylor & Brown 1988, 196). Ander bewijs bevestigt eveneens dat mensen onrealistisch positieve denkbeelden hebben over de toekomst. In een ander onderzoek bleek dat de voorspellingen van de meeste

mensen overeenkwam met wat ze zelf graag zouden zien gebeuren of wat sociaal wenselijk is, in plaats van wat objectief gezien waarschijnlijk zou zijn (Taylor & Brown 1988, 197).

Taylor en Brown menen dat wanneer deze illusies aanwezig zijn, mensen zich goed voelen of zelfverzekerder zijn. Wanneer deze illusies absent zijn, kan dit negatieve gevolgen hebben voor je mentale gezondheid. Zelfbedrog heeft dus positieve implicaties voor je mentale gezondheid. Aan de andere kant kunnen positieve illusies fysiek gezien wel slecht voor je zijn (Taylor & Brown 1988, 198). Stel je bijvoorbeeld een skater voor die denkt dat hij een bepaalde truc kan doen, maar vervolgens faalt omdat hij het toch niet bleek te kunnen.

2.4.2 Zelfbedrog in evolutionair opzicht

Zelfbedrog heeft naast voordelen voor je mentale gezondheid ook voordelen in biologisch aspect. Voordeel in deze zin moeten we opvatten als de positieve effecten op overleving en voortplanting. Trivers stelt de vraag hoe het kan zijn dat natuurlijke selectie voorkeur biedt aan zelfbedrogsmechanismen (Trivers 2011, 3).

De algemene stelling die Trivers poneert is dat zelfbedrog is geëvolueerd om ons beter te maken in het bedriegen van anderen. Zo redeneert hij dat we door onszelf te bedriegen beter in staat zijn anderen te bedriegen, in de zin dat de ander minder snel zal doorhebben wanneer we diegene bedriegen. Mocht de ander doorhebben dat we hem aan het bedriegen zijn en ons er vervolgens van beschuldigd, stelt zelfbedrog ons ook in staat ons beter te verdedigen tegen zulke claims. We zouden dankzij zelfbedrog een oprechtere indruk kunnen maken tegenover degene die we bedriegen (Trivers 2011, 4).

Vervolgens is het natuurlijk de vraag dat als zelfbedrog ervoor zorgt dat bedrog moeilijker te ontdekken is, hoe mensen doorgaans bedrog kunnen ontdekken. De belangrijkste signalen zijn volgens Trivers nervositeit, controle en cognitieve belasting. Omdat ontdekt worden in het bedriegen vaak negatieve gevolgen met zich meebrengt, zoals bijvoorbeeld een schuldgevoel, zijn mensen vaak nerveus wanneer ze liegen. Omdat we niet willen laten blijken dat we nerveus zijn, proberen we de situatie onder controle te houden. Dit kan echter detecteerbare bijwerkingen opleveren zoals te veel controle willen uitoefenen, overdrijving, een ingestudeerde indruk achterlaten, etc. Daarnaast kan liegen cognitief gezien veeleisend zijn. Je moet namelijk tegelijkertijd de waarheid onderdrukken en een leugen in stand houden die op het eerste gezicht plausibel klinkt, waarbij je er ook nog eens voor moet zorgen dat er geen contradicties in je verhaal zitten (Trivers 2011, 10). Omdat we onszelf kunnen bedriegen, en daarmee beter zijn in het bedriegen van anderen, zijn we ook beter in het onderdrukken van deze signalen.

Zelfbedrog heeft dus, samengevat, in biologisch opzicht meerdere voordelen:

> " ... the hallmark of self-deception in the service of deceit is the denial of deception, the unconscious running of selfish and deceitful ploys, the creation of a public persona as an altruist and a person "beneffective" in the lives of others, the creation of self-serving social theories and biased internal narratives of ongoing behavior, as well as false historical narratives of past behavior that hide true intention and causality". (Trivers 2011, 27)

2.4.3 Cognitieve architectuur

Kurzban meent echter niet dat zelfbedrog een functionele oorzaak heeft. Hij meent dat we onszelf bedriegen omdat onze hersenen nu eenmaal op een bepaalde manier werken. Zelfbedrog is volgens hem een gevolg van de "modular architecture of the mind" (Kurzban 2010, 133). Onze hersenen zijn in hoge mate modulair. Een manier om dit te begrijpen is als we onze hersenen zien als een doos waar iets in kan, er vervolgens iets in die doos gebeurt, en dat er vervolgens weer iets uit de doos komt.

Kurzban denkt dus niet dat we onszelf bedriegen omdat we een motivatie hebben die ons leidt. Het zijn juist onze hersenen die uit vele modules bestaan die gewoonweg op een bepaalde manier werken. Een thermostaat is bijvoorbeeld niet 'gemotiveerd' om de kamertemperatuur op 19 graden te houden. Het gedraagt zich misschien wel op deze manier, maar een thermostaat is in de eerste plaats natuurlijk ontworpen om de temperatuur op een bepaald punt te houden. Zo meent Kurzban ook dat de mechanismen in onze hersenen ontworpen zijn om bepaalde standen van zaken tot stand te brengen. Gemotiveerd moeten we in deze zin opvatten als een ontwerp dat er voor is om bepaalde doelen te realiseren (Kurzban 2010, 135).

Vanwege deze reden zou Kurzban het ook niet eens zijn met een evolutionaire uitleg die stelt dat we ons door zelfbedrog beter zouden voelen (Kurzban 2010, 136). Natuurlijke selectie werkt alleen vanwege de uitkomsten voor voortplanting. Modules zijn ontworpen om uitkomsten te realiseren die bijdragen aan voortplanting. Het is niet zo dat modules ontworpen zijn om een goed gevoel tot stand te brengen. Nu kan het natuurlijk wel zo zijn dat de uitkomsten van modules je een goed gevoel geven, maar het 'goed voelen' is op zichzelf niet de uitkomst waar de module voor ontworpen is om tot stand te brengen (Kurzban 2010, 137).

Evolutie stelt ons veeleer in staat om nuttige dingen te doen en ons constant aan te passen aan de omgeving:

> "My interpretation ... is that it is evolution's way of keeping the carrot just out of reach, motivating you to continue to do more useful and adaptive things. If you imagine an

> evolutionary history with two different mind designs – one complacent, in which once a goal is reached people rest on their laurels and whistle a happy tune all day – and one mind design that is never quite satisfied, in which each victory motivates further achievement – it's easy to see that the second one, while being less fun, would do more useful things". (Kurzban 2010, 142)

Kurzban meent dat wanneer we in het geval van zelfbedrog twee contradictoire overtuigingen hebben, het veeleer een geval is van verschillende overtuigingen in verschillende modules. Elke module heeft een eigen taak, verwerken informatie en zijn soms bewust (Kurzban 2010, 148). Zo is het in het geval van zelfbedrog zo dat onze hersenen nou eenmaal op deze manier in elkaar zitten, waarbij er zich in de ene module een verkeerde representatie bevindt en in een andere een meer nauwkeurige representatie (Kurzban 2010, 149). Vervolgens troeft de ene module als het ware de ander af. Je kunt echter niet stellen dat modules elkaar bedriegen, want modules hebben geen intenties.

3 Zelfbedrog en zelfkennis

Na uitvoerig de onderwerpen zelfbedrog en zelfkennis besproken te hebben, ga ik in dit hoofdstuk in op de relatie tussen de twee. Dit zal ik doen aan de hand van twee papers van Jordi Fernández en Dion Scott-Kakures. Hoewel ze niet allebei precies hetzelfde zeggen, zijn ze wel beide van mening dat zelfbedrog ontstaat doordat we in onze zelfkennis een fout maken. Ik zal eerst beide papers uiteenzetten, waarna ik vervolgens de argumentatie van beide papers zal analyseren.

3.1 Fernández

3.1.1 Conflict en normativiteit

De soort zelfbedrog waar Fernández het over heeft is zelfbedrog gekarakteriseerd door wat hij noemt het 'conflict' en de 'normativiteit' van zelfbedrog. Om duidelijk te maken wat dit inhoudt geeft hij een paar voorbeelden:

> "*Jack's health*
> Lately Jack has been avoiding reading any magazine or newspaper article on medical issues. If they appear on a TV program that he is watching, he immediately switches channels. If they come up in a conversation to which he is a party, he changes the topic. He has been scheduled to have a regular check-up with his doctor several times, but it is proving difficult for him to get this done. Each time the appointment is scheduled, Jack forgets about it and misses the appointment. Eventually, Jack's relatives have asked him whether he believes that he is sick, but Jack sincerely denies believing that." (Fernández 2013, 381)

> "*Tom's marriage*
> Tom has been trying to read his wife's e-mail correspondence for a few weeks. He has also attempted to overhear her conversations on the phone. He has checked her text messages on her mobile. He has sometimes followed her from a distance when she goes out. And he often asks her to give him a detailed account of her daily activities while she has not been in the house. Noticing some of this behavior, Tom's friends have asked him whether he believes that his wife is hiding something from him, but Tom honestly claims not to believe that." (Fernández 2013, 381)

Het conflict van zelfbedrog houdt in dat, in dit geval, Jack en Tom twee conflicterende overtuigingen hebben. Ze zeggen namelijk het een, maar doen het ander. Daarom is het moeilijk een overtuiging aan ze toe te schrijven en als we dat wel doen, zouden het twee conflicterende overtuigingen zijn. De normativiteit van zelfbedrog houdt in dat als we een conflict herkennen tussen de claims en de

handelingen die gemaakt worden, we dit verwerpelijk vinden. We hebben het idee dat we Jack en Tom kunnen bekritiseren op hun gedrag (Fernández 2013, 382).

Dit is dus de soort zelfbedrog die Fernández behandelt. We kunnen hierbij de toestand van zelfbedrog karakteriseren als hebbende een eigenschap X, zodanig dat het subject dat X heeft het voor ons moeilijk maakt om aan hem een overtuiging toe te schrijven, en dat we tegelijkertijd wel vinden dat het subject blaam treft. Om zelfbedrog vervolgens uit te leggen, moeten we weten wat die eigenschap X precies inhoudt (Fernández 2013, 383).

Aan de hand van intentionalisme kunnen we vaststellen dat zelfbedrog niet iets is wat het subject zomaar overkomt, aangezien het subject zichzelf bedriegt met de intentie om zichzelf te bedriegen. Het is dus veeleer iets dat het subject zichzelf heeft aangedaan, het resultaat van zijn eigen handelingen. Deze handelingen worden uitgevoerd met een misleidend doel. Het doel van het subject in het proces van zelfbedrog is om oneerlijk of onoprecht te zijn met zichzelf. Daarom is het ook niet vreemd dat we zelfbedrog verwerpelijk vinden. Toch stuit intentionalisme wel op twee paradoxen, en dat is waarom Fernández een alternatieve benadering wil proberen m.b.t. zelfbedrog: motivationalisme (Fernández 2013, 384-5).

3.1.2 Eerste- en tweede-orde motivationalisme

Zoals we weten houdt de motivationalistische benadering van Mele in dat we door een motivatie zoals bijvoorbeeld verlangen worden geleid in het vormen van een overtuiging. We hebben bijvoorbeeld een verlangen dat *p*, vervolgens behandelen we het beschikbare bewijs op een vooringenomen manier ten gunste van *p* en komen tenslotte tot de overtuiging dat *p*. Fernández noemt deze benadering “first-order motivationalism” (Fernández 2013, 385).

Een andere benadering voor motivationalisme is niet eerste-orde, maar tweede-orde motivationalisme. In plaats van dat het subject een verlangen heeft dat dingen op een bepaalde manier het geval zijn, heeft het subject een verlangen om de overtuiging te hebben dat dingen op een bepaalde manier het geval zijn. Als het subject bijvoorbeeld gelooft dat *p*, dan zou het volgens deze benadering zo zijn omdat hij het verlangen heeft om te geloven dat *p*, en dit verlangen heeft hem ertoe geleid om het bewijs dat betrekking heeft op *p* te behandelen op een vooringenomen wijze. De noodzakelijke en voldoende voorwaarden voor het subject om zichzelf te bedriegen zijn dan de volgende:

1. *P* is niet waar.

2. S heeft het verlangen om te geloven dat *p* waar is, en dit verlangen leidt S ertoe om het bewijs dat betrekking heeft op de waarheidswaarde van *p* op een vooringenomen wijze te behandelen.
3. S' vooringenomen behandeling van dit bewijs zorgt ervoor dat S gelooft dat *p* waar is.
4. De totaliteit van het bewijs dat S op dat moment bezit lijkt meer te duiden op niet-*p* dan *p*, en als dat niet zo is, is de verklaring voor dat feit dat S selectief informatie heeft verzameld. (Fernández 2013, 386)

Het verschil tussen eerste- en tweede-orde motivationalisme, is dat tweede-orde motivationalisme twee dingen probeert uit te leggen aan de hand van één verlangen. Deze dingen zijn (1) zaken die eerste-orde motivationalisme probeert uit te leggen door beroep te doen op het verlangen dat *p*, en (2) zaken die uitgelegd proberen te worden door beroep te doen op het verlangen dat niet-*p*. Beide zaken legt tweede-orde motivationalisme uit aan de hand van het verlangen om te geloven dat *p* (Fernández 2013, 386).

Naast het feit dat beide vormen van motivationalisme niet te maken krijgen met de paradoxen waar intentionalisme tegen aan loopt, zijn er nog twee andere voordelen aan deze motivationalistische benaderingen. Ten eerste is er onafhankelijke steun voor het motivationalistische idee dat de mechanismen die werkzaam zijn in het vormen van overtuigingen soms ook daadwerkelijk werkzaam zijn in ons. Ten tweede maken motivationalistische benaderingen weinig aannames. Deze benaderingen stellen namelijk slechts dat de inhoud van onze motiverende toestanden van invloed kunnen zijn op de manier waarop we met bewijs omgaan (Fernández 2013, 387).

Toch is er sprake van een dilemma. Als iemand een verlangen uitspreekt zoals eerste-orde motivationalisme voorstelt, is het moeilijk om het conflict van zelfbedrog te verklaren. Als iemand een verlangen uitspreekt zoals tweede-orde motivationalisme voorstelt, is het moeilijk om de normativiteit van zelfbedrog te verklaren. Om dit duidelijk te maken kunnen we dit toepassen op het voorbeeld van Jack. Jack wil niet ziek zijn en als gevolg wijst hij het bewijs voor zijn kanker af, of hij interpreteert het bewijs verkeerd. Dit maakt dat Jack de overtuiging heeft dat hij niet ziek is, maar wat kunnen we zeggen over zijn ontwijkende gedrag met betrekking tot de dokter? Wat verklaart het feit dat hij zijn dokter niet wilt zien en blootgesteld zou worden aan de informatie over zijn ziekte? Zeker niet zijn verlangen; je zou namelijk verwachten dat iemand met het verlangen niet ziek te zijn geïnteresseerd is in het vaststellen of hij ziek is of niet. Eerste-orde motivationalisme kan dit moeilijk verklaren (Fernández 2013, 387).

Met tweede-orde motivationalisme komen we echter eveneens in de moeilijkheden. Jack wilt geloven dat hij niet ziek is en als gevolg wijst hij tegengesteld bewijs af, wat ervoor zorgt dat hij de

overtuiging heeft dat hij niet ziek is. Jacks ontwijkende gedrag kan in dit geval wel verklaard worden, namelijk als een uiting van zijn verlangen. Het lijkt erop dat Jack rationeel gedrag vertoont: hij had het verlangen om een bepaald doel te bereiken, dacht waarschijnlijk na over de stappen die hij moest nemen om daar te komen, en vervolgens heeft hij deze stappen genomen. Waarom hebben we dan toch het gevoel dat Jack schuldig is? Dat gevoel legt tweede-orde motivationalisme niet uit (Fernández 2013, 388).

3.1.3 Het bypass-model

Zelfbedrog moeten we daarom zien als het falen van zelfkennis, dat plaatsvindt wanneer we onze capaciteit ervoor op een verkeerde manier gebruiken. Het model van zelfbedrog dat Fernández op deze gedachte bouwt zal dus uitgaan van een model van hoe we kennis verkrijgen van onze eigen mentale toestanden, en hoe we erachter komen wat onze overtuigingen zijn (Fernández 2013, 389).

Centraal in deze benadering van Fernández is de notie van de 'fundering' van een overtuiging, oftewel dat waar we een overtuiging op baseren. Voor elke overtuiging van een subject is er een aantal mentale toestanden dat, mocht het subject in een van die toestanden verkeren, hem ertoe leidt om die overtuiging te hebben. De mentale toestanden die ervoor zorgen dat ik zo'n overtuiging verkrijg, noemt Fernández de 'fundering' van die overtuiging. Daarnaast gaat Fernández uit van de volgende aanname: een subject is gerechtvaardigd in het vormen van een overtuiging als hij die overtuiging vormt op basis van een mentale toestand die voldoende steun biedt voor die overtuiging. Deze mentale toestanden kunnen bijvoorbeeld de visuele ervaringen zijn van een subject, of geheugen (Fernández 2013, 389).

Een mentale toestand biedt voldoende steun voor een overtuiging als die toestand correleert, in dat subject, met de stand van zaken die de overtuiging waar maakt. De correlatie tussen een toestand van een subject en de stand van zaken die een van zijn overtuigingen waar maakt genereert epistemische rechtvaardiging voor die overtuiging wanneer het subject die overtuiging vormt op basis van het zijn in die toestand. Voor het plaatsvinden van deze relatie, stelt Fernández twee noodzakelijke voorwaarden voor:

1. Afhankelijkheidsvoorwaarde: als een subject een overtuiging vormt op basis van zijn zijn in een bepaalde toestand, dan heeft hij die overtuiging omdat hij in die toestand verkeert.
2. Beschikbaarheidsvoorwaarde: als een subject een overtuiging vormt op basis van zijn zijn in een bepaalde toestand, is hij geneigd om te geloven dat hij zich in die toestand verkeert. (Fernández 2013, 390)

We kunnen een onderscheid maken in de manier waarop we vaststellen wat iemands overtuigingen zijn, namelijk een eerste- en een derdepersoons standpunt. Vanuit een derdepersoons standpunt vormen we overtuiging over onze overtuigingen door ons gedrag te observeren en daar dingen uit af te leiden. In het geval van Jack denken we niet dat hij overtuigingen vormt op basis van zijn gedrag. In tegenstelling tot dit gaan we ervan uit dat hij een eerstepersoons standpunt inneemt ten opzichte van zijn eigen overtuigingen (Fernández 2013, 390).

Als we een eerstepersoons standpunt innemen, kijken we als het ware langs onze overtuigingen om ze aan onszelf toe te schrijven, meent Fernández. We vormen overtuigingen over onze overtuigingen die we hebben gebaseerd op de 'fundering' voor die overtuigingen. Dus als ik, vanuit een eerstepersoons standpunt, de overtuiging vorm dat ik de overtuiging heb dat ik ziek ben, heb ik die overtuiging gevormd omdat ik in een toestand verkeer die mij doorgaans doet geloven dat ik ziek ben, bijvoorbeeld op basis van mijn waarnemingen van de symptomen. Wanneer ik vervolgens de overtuiging vorm dat ik een bepaalde overtuiging heb, constitueert de toestand op basis waarvan ik deze meta-overtuiging vorm de gronden voor de eerste-orde overtuiging in mij (Fernández 2013, 391). Dit model formuleert Fernández als volgt:

> Voor elke propositie *p* en elk subject S geldt:
> Normaal gesproken, als S gelooft dat hij gelooft dat *p*, dan is er een toestand E zodanig dat:
>
> 1. S' meta-overtuiging gevormd is op basis van zijn zijn in E.
> 2. E de gronden constitueert voor de overtuiging dat *p* in S. (Fernández 2013, 391)

Dit model noemt Fernández het 'bypass model'. Het verklaart enerzijds drie karakteristieke epistemische kenmerken van het eerstepersoons standpunt, en anderzijds biedt het een verklaring voor de transparantie van geloof (Fernández 2013, 391). Wat dit precies is legt Fernández als volgt uit.

Het toeschrijven van overtuigingen aan jezelf verschilt in twee opzichten van het toeschrijven van overtuigingen aan jou door anderen, namelijk: mijn rechtvaardiging voor het toeschrijven van overtuigingen aan mijzelf hangt niet altijd af van bewijs dat ik kan vinden uit mijn gedrag en ook niet van een redenering, terwijl dit bij anderen allebei wel altijd geldt. Fernández stelt dat het epistemisch gerechtvaardigd is om overtuigingen aan jezelf toe te schrijven via het bypass model: overtuigingen over onze overtuigingen die gevormd worden via het bypass model zijn gerechtvaardigd, en hangen niet af van gedragsobservatie of redenatie (Fernández 2013, 391).

Als ik bijvoorbeeld vaststel dat ik de overtuiging heb dat er een appel voor me ligt vanuit het eerstepersoons standpunt, dan vorm ik de overtuiging dat ik de overtuiging heb dat er een appel

voor mij ligt via het bypass model. E is de toestand op basis waarvan ik mijn meta-overtuiging vorm, want E is een toestand dat in mij de overtuiging veroorzaakt dat er een appel voor me ligt. Doordat ik in toestand E verkeer, is die toestand geneigd om te correleren met de stand van zaken die mijn meta-overtuiging waar maakt. Dus, E constitueert voldoende steun voor mijn overtuiging dat ik de overtuiging heb dat er een appel voor me ligt. Daarom is mijn overtuiging dat ik de overtuiging heb dat er een appel voor me ligt gerechtvaardigd. Hier komt dus geen gedragsobservatie of redenatie aan te pas (Fernández 2013, 392-3).

Dit model biedt een theorie van zelfkennis over overtuigingen. Het beeld van zelfkennis dat opkomt uit het bypass model is het volgende. Onze visuele ervaringen, herinneringen en andere toestanden op basis waarvan we eerste-orde overtuigingen vormen, hebben, vanuit epistemologisch standpunt, als het ware twee taken. Ze rechtvaardigen overtuigingen van twee verschillende ordes door betrokken te zijn bij twee verschillende soorten regelmatigheden. Onze visuele ervaringen bijvoorbeeld, constitueren voldoende steun voor onze visuele overtuigingen op grond van de correlatie met de stand van zaken. Dezelfde visuele ervaringen constitueren eveneens voldoende steun voor onze overtuigingen over onze visuele overtuigingen op grond van de correlatie met die visuele overtuigingen. Dezelfde toestanden kunnen daarom enerzijds onze gedachten over welke overtuigingen we hebben rechtvaardigen, en anderzijds die overtuigingen zelf rechtvaardigen. Wanneer dit echter gebeurt, oefenen ze deze twee rollen uit op basis van verschillende feiten (Fernández 2013, 393).

Daarnaast legt het bypass model de transparantie van geloof uit. Zoals we al eerder hebben kunnen zien bij Evans, richten we ons als het ware buitenwaarts als ons gevraagd wordt of wij geloven dat *p*. Vervolgens stellen we zelf de vraag “is het zo dat *p*?”. Als mij dus gevraagd wordt of ik geloof dat er een appel voor me ligt, ga ik niet bij wijze van spreken mijn hersenen scannen op zoek naar een toestand die ik kan identificeren als de overtuiging dat er een appel voor me ligt. Wat er gebeurt is dat ik kijk naar wat er voor me ligt. Ik hou me bezig met de relatie tussen mij en de buitenwereld en richt mijn aandacht op het intentionele object van de overtuiging (Fernández 2013, 394).

We hebben gezien dat het bypass model gebruik maakt van drie conceptuele middelen, namelijk (1) een minimale notie van de relatie waarop we overtuigingen vormen, (2) een notie van epistemische rechtvaardiging en (3) de aanname dat onze overtuigingen gebaseerd zijn op een bepaalde ‘ground’ (Fernández 2013, 394). Vervolgens moeten we kijken hoe dit model zelfbedrog uitlegt.

3.1.4 Bypass en zelfbedrog

Het type zelfbedrog waar Fernández zich op richt moet gezien worden als het falen van zelfkennis van een specifieke soort. Het is het soort falen dat gebeurt wanneer iemand niet alleen een fout

begaat, maar ook dat die fout het resultaat is van iemands nalatigheid. Deze fout is de misvatting over welke overtuigingen iemand heeft. De nalatigheid heeft betrekking op het schenden van een bepaalde norm die betrekking heeft op de formatie van (meta-)overtuigingen. Fernández' voorstel is dat het eerste aspect van dit falen het conflict van zelfbedrog verklaart terwijl het tweede aspect de normativiteit verklaart (Fernández 2013, 395).

Laten we nu weer het voorbeeld van Tom erbij halen. Enerzijds gedraagt hij zich alsof hij gelooft dat zijn vrouw iets voor hem verbergt, anderzijds beweert hij stellig dat hij dat niet gelooft. Dit conflict kan worden opgelost als we ervan uitgaan dat Tom een fout maakt over zijn eigen overtuigingen. Zijn acties kunnen uitgelegd worden als uitingen van een eerste-orde overtuiging, namelijk, de overtuiging dat zijn vrouw iets voor hem verbergt. Zijn bewering dat hij die overtuiging niet heeft kan worden uitgelegd als een uiting van een meta-overtuiging, namelijk, de overtuiging dat hij niet gelooft dat zijn vrouw iets voor hem verbergt. Aangezien de overtuigingen hier van verschillende ordes zijn, hoeven we niet twee tegengestelde overtuigingen te postuleren om het conflict van zelfbedrog te verklaren. Hiermee ontwijken we de 'statische' paradox (Fernández 2013, 395).

Het zelfbedrogen subject bevindt zich in een staat van vergissing. Intentionalisme en motivationalisme incorporeren beide dit idee. Het is echter wel zo dat in beide benaderingen de aanname wordt gedaan dat de relevante soort fout waar naar gezocht moet worden de eerste-orde overtuigingen van het subject betreft. Fernández' voorstel plaatst de fout juist in de meta-overtuigingen van het subject zelf. Dit maakt het zelf het object van zelfbedrog. Zelfbedrog gaat, in dit opzicht, over iemands zelf (Fernández 2013, 395).

Fernández stelt voor dat de reden waarom we zelfbedrog verwerpelijk vinden ligt aan het idee dat het zelfbedrogen subject iets doet wat hij niet zou moeten doen. Het gaat hier niet om een 'zou moeten' in morele zin van het woord, maar juist in epistemische zin. We hebben namelijk het idee dat het subject een bepaalde epistemische norm schendt. Stel je de volgende norm over het vorm van overtuigingen voor: voor elke propositie *p* geldt dat iemand niet moet geloven dat *p* als diegene voldoende steun heeft voor de overtuiging dat *p* niet het geval is, tenzij er belangrijkere overwegingen zijn ter steun van de overtuiging dat *p*. Voor het vormen van meta-overtuigingen geldt: voor elke propositie *q*, moet het subject niet geloven dat dat hij niet gelooft dat *q* als hij voldoende steun heeft voor het geloven dat hij gelooft dat *q*, tenzij er belangrijkere overwegingen meespelen ter steun van het geloven dat hij de overtuigingen *q* niet heeft. Fernández stelt voor dat het schenden van deze norm onze intuïtie dat zelfbedrog verwerpelijk is verklaart (Fernández 2013, 396).

Als dit allemaal klopt, gelooft Tom dat zijn vrouw iets voor hem verbergt en gelooft Jack dat hij ziek is. Het lijkt in eerste instantie aannemelijk om te stellen dat Tom en Jack voldoende grond hebben voor hun overtuigingen. Deze 'funderingen' hoeven echter niet voldoende steun te constitueren voor beide overtuigingen. Ze hoeven alleen toestanden van die aard te zijn dat ze, wanneer Tom en Jack ze bezitten, geneigd zijn om deze overtuigingen te vormen. Als het bypass model klopt, dan is het zo dat welke gronden dan ook Tom heeft voor zijn overtuiging dat zijn vrouw iets voor hem verbergt, voldoende steun constitueren in Tom voor de overtuiging dat hij de overtuiging heeft dat zijn vrouw iets voor hem verbergt. We nemen hier aan dat wanneer we zeggen dat Tom voldoende grond heeft voor zijn eerste-orde overtuigingen, hij voldoende steun heeft voor het geloven dat hij deze eerste-orde overtuigingen heeft. Maar, als de diagnose van het conflict van zelfbedrog klopt, gelooft Tom eigenlijk dat hij de overtuiging dat zijn vrouw iets voor hem verbergt niet heeft. Het zal voor ons dan ook geen verassing zijn dat we de intuïtie hebben dat hun omstandigheden verwerpelijk zijn. Bovendien zijn Tom en Jack blind voor het feit dat ze voldoende steun hebben voor meta-overtuigingen die tegengesteld zijn aan de meta-overtuigingen die ze daadwerkelijk hebben gevormd. Het feit dat ze beide vasthouden aan hun meta-overtuigingen in zo'n situatie bewijst dat ze zich gedragen op een epistemisch nalatige manier. Dit is volgens Fernández de reden waarom we het idee hebben dat Tom en Jack blaam treffen (Fernández 2013, 396-7).

3.2 Scott-Kakures

3.2.1 Tussen intentionalisme en motivationalisme

In zijn paper verdedigt Scott-Kakures de volgende twee stellingen:

1. Reflectieve, kritische redenatie is essentieel voor het proces van zelfbedrog.
2. Het proces van zelfbedrog impliceert een bepaald karakteristiek falen van zelfkennis. (Scott-Kakures 2002, 576)

Redeneren zou ons verleiden om proposities te omarmen die beantwoorden aan onze interesses. Wat dit reflectieve kritische redeneren inhoudt hebben we al eerder kunnen lezen bij Burge. De verleiding door middel van redeneren vereist een bepaald karakteristiek falen van zelfkennis. Met zelfbedrog vorm ik mijn mentale houdingen in overeenstemming met mijn epistemische evaluaties. Sterker nog, ik ga iets geloven door iets in aanvulling op mijn epistemische evaluaties – iets dat deze evaluaties zelf (ver)vormt. Op deze manier is zelfbedrog een probleem en een falen van zelfkennis (Scott-Kakures 2002, 576).

Scott-Kakures meent dat motivationalisme grotendeels gelijk heeft met betrekking tot de aard van het ontstaan van zelfbedrogen overtuigingen. Daarnaast meent hij dat intentionalisme gelijk heeft in

het stellen dat motivationalisme iets achterwege laat, namelijk reflectieve kritische redenatie. Dit is karakteristiek voor de activiteit van het zelf bedriegen (Scott-Kakures 2002, 577).

Dieren kunnen zichzelf niet bedriegen. Een kat kan niet zelfbedrogen zijn omdat ze simpelweg processen ondergaat die buiten haar controle liggen. Ze is als het ware een hulpeloze omstander. Een normaal persoon is niet een hulpeloze omstander; een normaal persoon kan meta-cognitieve controle over zijn geloofsvormingsproces hebben. Deze controle vereist het bezit van conceptuele complexiteit. Een mens zou kunnen vragen: "is *dit* voldoende reden om *dat* te concluderen?", of "zal ik meer bewijs verzamelen?", of zelfs "is mijn evaluatie van het bewijs bevooroordeeld door verlangen(s)?". Zulke vragen vereisen een begrip van het concept van overtuigingen en de relatie daarvan met rechtvaardiging en rede (Scott-Kakures 2002, 583).

Scott-Kakures probeert aan te tonen dat zelfbedrog ontstaat vanuit kenmerkende menselijke activiteiten en mislukkingen. Hoewel de intentionalist terecht stelt dat zelfbedrog reflectieve activiteit vereist, eist hij onterecht dat die activiteit een intentionele activiteit moet zijn die gericht is op de bewerkstelliging van de begunstigde overtuiging, en is hij eveneens fout in het aandringen dat het relevante bewustzijn de vorm aanneemt van een overtuiging dat de contradictoire propositie beter wordt ondersteund door zijn bewijs (Scott-Kakures 2002, 585-6). In zelfbedrog is cognitie gevormd door interesses en verlangens, zelfs wanneer het subject kritisch en reflectief redeneert, waarbij het subject zijn mentale houdingen vormt in overeenstemming met zijn epistemische evaluaties (Scott-Kakures 2002, 586).

3.2.2 Reflectieve kritische redenatie

Iemand die zichzelf bedriegt redeneert reflectief terwijl hij zich begeeft in het zelfbewust testen van hypotheses. Dit idee van het testen van hypotheses neemt Scott-Kakures over van Mele's FTL-model. Zo zal diegene bijvoorbeeld gedachten hebben in de vorm van "*p*, omdat *r* en *s*". Door dit te concluderen gelooft hij dat hij is geleid door en heeft hij zijn houdingen gevormd in overeenstemming met zijn epistemische evaluaties (Scott-Kakures 2002, 586).

Scott-Kakures meent dat voorbeelden van zelfbedrog waarin kritische redenatie afwezig is, gevallen van "wishful thinking" zijn, oftewel verlangend denken. Deze voorbeelden kunnen volgens hem gemakkelijk voorbeelden van zelfbedrog worden, aangezien de capaciteit voor reflectieve redenatie ertoe kan leiden dat subjecten geloven dat ze gerechtvaardigd zijn in dat we ze geloven. In zo'n geval bedriegen subjecten zichzelf in de manier waarop ze op een overtuiging zijn gekomen (Scott-Kakures 2002, 587). Waarom "wishful thinking" verschilt van voorbeelden van zelfbedrog waarin reflectieve kritische redenatie wel aanwezig is, wil Scott-Kakures laten zien aan de hand van een voorbeeld:

> "Ingrid greatly fears the significance of recurring pains in her lower back. She is of an age when such matters should be taken seriously, and she does know that her aunt died of ovarian cancer. Still, she thinks to herself that the odds of her actually having the disease must certainly be quite small. Her desire that she [does] not have cancer causes her to generate more benign hypotheses as to the significance of her discomfort. Her search and her reasoning are biased in the sense that Ingrid focuses upon various data and downplays the significance of others. She gives, for example, great importance to the fact that her maternal grandmother had arthritis, while downplaying the significance of a history of reproductive cancers. She notes that she has been under great stress at work and has, as a result, been a rare visitor to the gym. This, too, she thinks, must certainly be implicated in her distress. Considerations that cut against her favored hypothesis and point toward more dreadful possibilities are discounted as irrelevant, or are subject to intensive scrutiny. Ingrid, for example, discounts recent unexplained fevers, noting that she has been prone to these since childhood. She may, as well, think to herself that if she, in fact, had ovarian cancer she would certainly be troubled by various other symptoms, and these, she notes, are absent. Recalcitrant data may then be downplayed by appeal to the benign hypotheses she has now accepted." (Scott-Kakures 2002, 588)

Zo lang de overgang van de relevante overweging naar het maken van een conclusie wordt bemiddeld door het reflectieve bewustzijn van de gerechtvaardigde relatie tussen de twee, is de *agent* actief de richting van haar cognitie aan het vormen op manieren die gevoelig zijn voor haar begrip van wat haar redenatie haar aanraadt. Door reflectief te redeneren vormt Ingrid haar mentale houdingen in overeenstemming met haar epistemische oordelen (Scott-Kakures 2002, 589). Zelfbedrog omvat activiteiten van een *agent*. Ingrid redeneert zelfbewust dat ze gezond is. Ze evalueert het beschikbare bewijs en ze speelt een actieve rol in het maken van een fout. Ze is daarnaast bewust van de stappen die ze neemt die haar leiden tot zelfbedrog (uiteraard voor haar niet onder die noemer) (Scott-Kakures 2002, 590).

Waar bij zelfbedrog deelname van een *agent* vereist is, is die deelname absent in "wishful thinking". Deze claim omvat volgens Scott-Kakures het overgrote deel van wat belangrijk is in intentionalistische benaderingen, terwijl we de excessen van die benaderingen vermijden. Motivationalisten zullen echter niet ontkennen dat gevallen waarin reflectieve kritische redenatie naar een conclusie aan bod komt gevallen van zelfbedrog zijn. Toch zullen ze wel stellen dat zo'n redenatie een noodzakelijke voorwaarde is voor zelfbedrog, maar waarom zou dit noodzakelijk zijn (Scott-Kakures 2002, 591)?

Het korte antwoord op deze vraag is dat om zelfbedrogen te zijn, de zelfbedrieger een actieve rol moet spelen in zijn bedrog. Daarnaast kunnen we de claim verdedigen dat zelfbedrog deelname van de *agent* vereist vanuit methodologische overwegingen. Als we een benadering kunnen vinden die een onderscheid maakt tussen zelfbedrog en *wishful thinking* in termen van de processen of mechanismen die erbij betrokken zijn, en die beantwoordt aan de intuïtie dat dit onderscheid er is, zouden we daar de voorkeur aan moeten geven, meent Scott-Kakures (Scott-Kakures 2002, 591).

Een van deze intuïties is dat zelfbedrog een vreemde vorm van irrationaliteit omvat waarin de *agent* op een of andere manier bewust normen schendt die hij aanhoudt. Hier kunnen we een vergelijking trekken met wat Fernández de normativiteit van zelfbedrog noemt. Fernández meent namelijk dat we zelfbedrog verwerpelijk vinden juist omdat een *agent* zijn epistemische normen schendt. Wat centraal staat in de gedachte van Scott-Kakures, is dat in het geval van zelfbedrog, rede gebruikt wordt tegen iemand zelf: de *agent* geeft zichzelf reden(en) om dat te geloven waarvan hij gelooft dat hij voldoende reden(en) heeft om het niet te geloven (Scott-Kakures 2002, 591).

Er is ook bewijs dat reflectief redeneren zelf een rol kan spelen in onze doxastische verleidingen. Wilson, Hodges en Lafleur (Wilson, Hodges & Lafleur 1995) stellen dat het proces van reflectief redeneren onze cognitie kan vervormen. In een aantal experimenten werden deelnemers gevraagd of ze een initiële impressie van een persoon konden maken door 14 beschrijvingen te gebruiken, zoals “Fran is quite high-strung”, of “Fran lent money to an acquaintance who had lost his wallet”, etc. (Scott-Kakures 2002, 592). Vervolgens ondergingen de deelnemers geheugenmanipulatie, waarbij het herinneren van positieve of negatieve kenmerken verhoogd werd. Hierna werden de deelnemers gesplitst en werd ze gevraagd een vragenlijst in te vullen. Groep 1 werd gevraagd waarom ze zich voelden zoals ze zich voelden ten opzichte van een persoon, terwijl groep 2 werd gevraagd zo veel mogelijk feiten op te noemen als ze konden over de persoon. Uiteindelijk werden alle deelnemers gevraagd wat ze van de persoon vonden (Scott-Kakures 2002, 592).

De geheugenmanipulatie beïnvloedde significant groep 1, wat ervoor zorgde dat zij de desbetreffende persoon positiever beschouwde, maar groep 2 niet. Dit effect was zelfs aanwezig wanneer de deelnemers niet werd gevraagd om initieel al een impressie te geven van de persoon. De geheugenmanipulatie maakt een vooringenomen set van relevante data toegankelijk. Wilson, Hodges en Lafleur stelden dat redeneren dient om de beoordeelde bruikbaarheid van deze toegankelijke informatie een zetje in de rug te geven (Scott-Kakures 2002, 592).

Dit onderzoek is relevant voor Scott-Kakures’ punt, omdat er goede reden is om te geloven dat reflectief redeneren zelf een rol kan spelen in het vormen van onze overtuigingen in de richting van onze interesses en verlangens. Voor iedereen die redeneert geldt dat mensen die zichzelf bedriegen

geneigd zijn om te beginnen met het testen van hypotheses met gunstige mogelijkheden. Door het redeneren is het waarschijnlijk dat ze deze hypotheses bevestigen. Deze tendens wordt niet geremd zelfs als subjecten van tevoren gewaarschuwd worden over de mogelijke onrepresentatieve aard van de redenen waar ze toegang tot hebben. Subjecten die van tevoren worden gewaarschuwd lijkt juist meer de neiging te hebben om hun redenen te zien als niet-vooringenomen. We lopen allemaal permanent risico om onszelf te bedriegen, omdat we in het redeneren te snel bewogen worden door krachten waar we ons niet bewust van zijn (Scott-Kakures 2002, 592-3).

3.2.3 Link met zelfkennis

Wanneer we zelfbewust redeneren, laten we onszelf leiden door de kracht van onze redenen. Het doel van dit reflectief redeneren is om controle uit te oefenen over de richting van onze cognitie, op een manier die ons begrip van onze redenen reflecteert (Scott-Kakures 2002, 593).

Het lijkt erop dat Ingrid uit het eerdergenoemde voorbeeld bewust is van de hypotheses die ze test, ze evalueert het beschikbare bewijs, overweegt verschillende hypotheses, en toch lijkt het alsof ze haar redenering aanpast om aan haar verlangens te beantwoorden. Intentionalisten zouden zeggen dat het gedrag van Ingrid bemiddeld moet worden door een intentie om haar cognitie vooringenomen in de richting van de gewenste overtuiging te sturen. Volgens Scott-Kakures is dit niet een vereiste voor zelfbedrog: wat vereist is voor zelfbedrog is dat de zelfbedrieger lijdt aan een falen van zelfkennis, bijvoorbeeld dat Ingrid fout zit in wat het maakt dat haar testen van hypotheses de richting aanneemt die het op gaat (Scott-Kakures 2002, 593).

Waarom dit zo is, legt Scott-Kakures uit aan de hand van meerdere onderzoeken van een aantal psychologen, waaronder Friedrich, Trope en Liberman. Volgens deze psychologen is het testen van hypotheses gemotiveerd door pragmatische belangen van het subject, oftewel, hoe intensief en op welke manier een individu een hypothese test reflecteert zijn interesses en waarden. Dit hebben we al kunnen zien bij de bespreking van het FTL-model (Scott-Kakures 2002, 593). Volgens deze kijk is cognitie geschikt om de beloningen of voordelen veilig te stellen en wat niet verlangt wordt te vermijden. Neem nu iemand die een hypothese wil testen, zoals Ingrid. Samen met haar interesses, waarden, verlangens en andere overtuigingen overweegt ze de vraag "*p* of ~*p*?". Het aanpakken van zo'n vraag zal onvermijdelijk tijd en energie kosten voor Ingrid. Wanneer er niet zo veel op het spel staat, is er ook niet veel reden om de hypothese intensief te testen. In zulke gevallen lijkt het erop dat subjecten vertrouwen op een snellere strategie (Scott-Kakures 2002, 594).

Deze pragmatische manier van hypotheses testen brengt nog een ander cruciaal punt met zich mee, namelijk de 'kosten' voor eventuele fouten in het testen van hypotheses. De motivatie voor het testen van hypotheses wordt voorzien door hoeveel het subject het uitmaakt of hij fout zit (Scott-

Kakures 2002, 594). Denk hierbij aan de acceptatie- en de zekerheidsdrempel die een subject voor zichzelf kan stellen.

De vooringenomenheid van Ingrid is het resultaat van de asymmetrie in de drempels die ze heeft voor verschillende hypotheses. Deze drempels zijn weer het resultaat van haar interesses en verlangens. Dit is de rol van het verlangen of de motivatie in haar strategische redeneerproces (Scott-Kakures 2002, 595).

Als dit zo is, hoeven we niet aan te nemen dat zulk gedrag bemiddeld wordt door een leidende intentie. Het is veeleer zo dat Ingrid redenen zoekt die haar toestaan om haar onderzoek te beëindigen. Benaderingen van deze pragmatische manier van het testen van hypotheses laten ons zien dat in het geleid worden dat wat haar rede haar aanraadt, Ingrid geleid wordt door haar verlangens en interesses. Ingrid gelooft dat het aannemen van haar conclusie een resultaat is van haar epistemische evaluaties van wat haar redenen haar aanraden. In feite is het zo dat haar doxastische activiteit gevormd wordt door verlangens en interesses. Ze gelooft dat haar onderzoek en haar conclusie geleid worden door haar epistemische oordelen. Ze denkt ten onrechte dat haar onderzoek door iets anders geleid wordt. Verlangen, wat ervoor zorgt hoe grondig ze haar overwegingen evalueert, vormt haar redenen en bepaalt de conclusies die ze bereikt. Verlangen of interesse sturen Ingrid om haar redenen aan te nemen als redenen om te geloven. Dit is de kenmerken fout van zelfkennis in zelfbedrog (Scott-Kakures 2002, 595-6).

De fout die wordt gemaakt in zelfkennis die Ingrids zelfbedrog mogelijk maakt is een misvatting van wat haar doxastische of cognitieve activiteiten bezielt. Net zoals elke andere reflectieve redeneerder, zal Ingrid haar onderzoeken zien als gestuurd door haar zelf, door haar begrip van wat haar rede haar aanraadt; haar zoektocht is een zoektocht naar redenen die haar toestaan om een bevredigend einde te brengen aan haar onderzoeken. Toch is het zo dat, onafhankelijk van haar eigen evaluaties, oordelen en activiteiten, haar onderzoeken gestuurd worden door haar verlangens of interesses (Scott-Kakures 2002, 599).

3.3 Analyse van de argumenten

Na beide papers uiteengezet te hebben, zal ik in deze sectie nog een keer de argumenten benoemen en analyseren.

3.3.1 Fernández

Fernández geeft antwoord op de vraag welke eigenschap van zelfbedrogen subjecten het moeilijk maakt voor ons om overtuigingen aan hen toe te schrijven, en wat de indruk wekt dat hun omstandigheden verwerpelijk zijn. Deze benadering van zelfbedrog wordt gekarakteriseerd door het

conflict en de normativiteit van zelfbedrog. In mijn optiek is dit een goede benadering voor zelfbedrog, aangezien dit twee algemeenheden zijn die we meestal terugvinden in voorbeelden van zelfbedrog. Het komt natuurlijk voor dat er voorbeelden worden gegeven van zelfbedrog waar bijvoorbeeld de normativiteit van zelfbedrog ontbreekt, maar dat zou met deze theorie niet uit hoeven maken. Wat deze theorie namelijk zo sterk maakt is dat het de variëteit van zelfbedrog kan verklaren die gekarakteriseerd wordt door conflict en normativiteit. Intentionalisme en motivationalisme kunnen deze variëteit niet verklaren.

Daarnaast is het zo dat er in de literatuur over zelfbedrog een grote onenigheid bestaat in de definitie van zelfbedrog, maar meestal omvatten benaderingen van zelfbedrog wel deze twee kenmerken die Fernández noemt. Doorgaans omvat zelfbedrog namelijk een subject die een foutieve overtuiging vormt terwijl er bewijs is voor het tegenovergestelde (conflict), en die gedrag vertoont dat lijkt te suggereren dat hij toch wel enig bewustzijn van de waarheid heeft (normativiteit).

De theorie van Fernández loopt tevens niet tegen de ‘statische’ en ‘dynamische’ paradoxen aan. Het ontwijkt de ‘statische’ paradox doordat Fernández een onderscheid maakt in eerste- en tweede-orde overtuigingen, en het ontwijkt de ‘dynamische’ paradox omdat het subject in Fernández’ theorie niet een intentie heeft om een overtuiging te vormen, maar juist geleid wordt door verlangens en interesses.

De eigenschap die zelfbedrogen subjecten hebben, is volgens Fernández dat ze een vergissing maken in zelfkennis waarin epistemische nalatigheid een rol speelt. Een voordeel van Fernández’ benadering is dat het gebruik maakt van weinig conceptuele middelen. Een centrale notie binnen zijn benadering is de notie van de ‘fundering’ van een overtuiging. Dit is een aantal mentale toestanden die een subject ertoe leidt om een overtuiging te hebben. Hierin gaat Fernández uit van een aanname, namelijk dat een subject gerechtvaardigd is in het vormen van een overtuiging als hij die overtuiging vormt op basis van een mentale toestand die voldoende steun biedt voor die overtuiging. Deze aanname klinkt intuïtief aantrekkelijk. Vervolgens stelt hij dat overtuigingen over onze overtuigingen die gevormd worden door het bypass model gerechtvaardigd zijn, en niet afhangen van gedragsobservatie of redenatie. Dit is echter nog maar de vraag, want dit zou betekenen dat een subject gerechtvaardigd is in het zelf toeschrijven van een tweede-orde overtuiging bij gratie van het bewijs dat de eerste-orde overtuiging genereert. Je zou namelijk ook kunnen zeggen dat een tweede-orde overtuiging gerechtvaardigd is bij gratie van de eerste-orde overtuiging zelf.

Al met al lijkt de benadering van zelfbedrog erg aantrekkelijk, terwijl het bij Fernández’ bypass model nog maar de vraag is of dat wat gesteld wordt wel klopt. Als dit klopt, biedt het een goed inzicht in de manier waarop zelfkennis en zelfbedrog aan elkaar gerelateerd zijn.

3.3.2 Scott-Kakures

Bij Scott-Kakures ontstaat zelfbedrog door reflectieve kritische redenatie, gepaard met een specifieke fout in onze zelfkennis. Zijn benadering kent intentionalistische, maar ook motivationalistische kenmerken. Zo heeft zijn theorie motivationalistische kenmerken in de zin dat zelfbedrog niet intentioneel is, en dat we worden geleid door verlangen(s) in het testen van hypotheses (volgens het FTL-model) en in onze capaciteit als reflectieve kenners. Zijn theorie is intentionalistisch in de zin dat zelfbedrog vereist dat er een *agent* is die een actieve rol speelt in het genereren van het bedrog. Zelfbedrog is namelijk een zelfbewuste kwestie. Aan deze intentionalistische claim voegt Scott-Kakures toe dat wanneer reflectieve kritische redenatie een rol speelt in zelfbedrog, ook al heeft de *agent* niet de intentie om zichzelf te bedriegen, hij toch zijn zelfbedrog door zijn activiteit uitvoert.

Omdat wij de capaciteit hebben om reflectief en kritisch te redeneren, kunnen we in het testen van hypotheses volgens het FTL-model verschillende cognitieve strategieën toepassen om ongunstige bewijsstukken weg te redeneren en daarvoor in de plaats 'vriendelijke' theorieën te vormen, zoals Scott-Kakures dat noemt. Wij worden, als we onszelf bedriegen, continu geleid door een motivatie in het genereren van overtuigingen. Deze benadering klinkt aantrekkelijk en wordt ondersteund door psychologisch onderzoek (Wilson, Hodges en Lafleur 1995).

Zelfbedrog is meer dan alleen de generatie van een overtuiging als een resultaat van een gemotiveerde vooringenomen cognitie. Dit komt omdat in elk stadium van het redeneerproces, de *agent* deelneemt in het genereren van de overtuiging die hij voor ogen heeft. Elke stap die het subject onderneemt wordt reflectief op beredeneerd om de gewenste conclusie te bereiken. Ook al weet het subject niet hoe zwaar verschillend bewijs weegt, moet hij het bewijs toch evalueren met vragen in de vorm van "*p* of ~*p*?".

De rol van zelfkennis in dit hele verhaal is dat het subject een misvatting heeft van wat zijn doxastische of cognitieve activiteiten bezielt. We zagen in het voorbeeld van Ingrid dat zij haar onderzoeken ziet als gestuurd door haar zelf, door haar begrip van wat haar rede haar aanraadt. Maar dit is precies waar haar misvatting over haar zelf plaatsvindt. Zij is van mening dat ze haar onderzoeken nauwkeurig evalueert, juiste oordelen maakt en het is precies in die activiteiten dat ze fout zit, aangezien haar onderzoeken worden geleid door verlangens of interesses.

Deze benadering van zelfbedrog als een fout die plaatsvindt in onze zelfkennis, is minder eisend dan Fernández' benadering, maar beide zijn plausibel. Daarnaast kunnen er kanttekeningen geplaatst worden bij het model van zelfkennis dat Fernández aanhoudt. Scott-Kakures baseert zijn benadering van zelfbedrog niet op een model van zelfkennis, maar voor beide auteurs geldt dat ze denken dat aan zelfbedrog een probleem in onze zelfkennis ten grondslag ligt. In beide benaderingen blijkt dat

we in zelfbedrog een epistemische norm schenden. We denken ten onrechte dat we, in het testen van hypotheses, een gerechtvaardigd oordeel maken en tot juiste conclusies komen, terwijl dat in beide gevallen niet zo is. In Fernández' geval is dit een misvatting van onze tweede-orde overtuigingen, en in Scott-Kakures' geval is dit een misvatting van wat onze cognitieve acitiviteit leidt.

4 Conclusie

In de analyse is gekozen voor een voornamelijk motivationalistische benadering van zelfbedrog, aangezien zelfbedrog geleid wordt door een motivatie. Meestal is deze motivatie een verlangen of meerdere verlangens. De problemen met intentionalisme blijven toch de twee paradoxen waar tegenaan gelopen wordt. Alhoewel ik in dit paper wel een intentionalistische benadering heb toegelicht van Sorensen waarin niet wordt aangelopen tegen de 'statische' paradox, blijft de 'dynamische' paradox toch een probleem. Om dit probleem te vermijden, is er gekozen voor een benadering waarbij we in zelfbedrog niet worden geleid door een intentie, maar door een motivatie.

De meest omvattende definitie van zelfbedrog is in mijn optiek een definitie die gekenmerkt wordt door, zoals Fernández het noemt, het conflict en de normativiteit van zelfbedrog. In de literatuur over zelfbedrog worden ook voorbeelden behandeld die alleen worden gekenmerkt door het conflict van zelfbedrog, maar als een model beide kenmerken kan verklaren, zou daar de voorkeur naar uit moeten gaan.

We kunnen nu een antwoord bieden op de onderzoeksvraag van deze scriptie. De onderzoeksvraag luidde: "hoe zijn zelfkennis en zelfbedrog aan elkaar gerelateerd?". Aan de hand van voorgaande analyse is gebleken dat de relatie tussen zelfkennis en zelfbedrog als volgt is: ten grondslag aan zelfbedrog ligt een falen van zelfkennis. Dit falen kan een misvatting over onze tweede-orde overtuigingen inhouden, of een misvatting over wat onze cognitieve of doxastische activiteit leidt. Zelfbedrog ontstaat vervolgens wanneer deze misvatting plaatsvindt, zoals een conflict tussen eerste- en tweede-orde verlangens, of ten onrechte denken dat we in het testen van hypotheses geleid worden door onze eigen evaluaties, oordelen en activiteiten, terwijl we in feite gestuurd worden door onze verlangens of interesses.

Bibliografie

Burge, T. 1998. “Our Entitlement to Self-Knowledge”, in P. Ludlow en N. Martin (eds.), *Externalism and Self-Knowledge* (Stanford: CSLI Publications): 239-64.

Cassam, Q. 2014. *Self-Knowledge for Humans*. New York: Oxford University Press.

Descartes, R. 1985a. *Philosophical Writings of Descartes*, vol. 1, vertaald door John Cottingham, Robert Stoothoff, Dugald Murdoch, en Anthony Kenny. Cambridge: Cambridge University Press.

Descartes, R. 1985b. *Philosophical Writings of Descartes*, vol. 2, vertaald door John Cottingham, Robert Stoothoff, Dugald Murdoch, en Anthony Kenny. Cambridge: Cambridge University Press.

Evans, G. 1982. *The Varieties of Reference*. New York: Oxford University Press.

Fernández, J. 2013. “Self-deception and self-knowledge”, in *Philosophical Studies*, 162(2): 379-400.

Gazzaniga, M.S. 1995. “Principles of Human Brain Organization Derived From Split-Brain Studies”, in *Neuron* 14: 217-228.

Goldman, A.I. 2006. *Simulating Minds*. New York: Oxford University Press.

Hirstein, W. 2000. “Self-Deception and Confabulation”, in *Philosophy of Science*, 67: 418-429.

Horgan, T. 2012. “Introspection About Phenomenal Consciousness”, in *Introspection and Consciousness*: 405-421.

Korsgaard, C.M. 2009. “The Activity of Reason”, in *Proceedings and Addresses of the American Philosophical Association*, 83(2): 23-43.

Kurzban, R. *Why Everyone (Else) Is a Hypocrite: Evolution and the Modular Mind*. Princeton: Princeton University Press.

Mele, A.R. 2001. *Self-Deception Unmasked*. Princeton: Princeton University Press.

Ramachandran, V.S. 1995. “Anosognosia in Parietal Lobe Syndrome”, in *Consciousness and Cognition* 4: 22-51.

Scott-Kakures, D. 2002. “At “Permanent Risk”: Reasoning and Self-Knowledge in Self-Deception”, in *Philosophy and Phenomenological Research*, 65(3): 576-603.

Siewert, C. 2012. “On the Phenomenology of Introspection”, in *Introspection and Consciousness*: 129-168.

Sorensen, R.A. 1985. “Self-Deception and Scattered Events”, in *Mind*, 94(373): 64-69.

Sperry, R.W. 1985. “Consciousness, Personal Identity, and the Divided Brain”, in D.F. Benson en Eran Zaidel (eds.), *The Dual Brain: Hemispheric Specialization in Humans* (New York: Guilford): 11-25.

Taylor, S.E. en Brown, J.D. 1988. “Illusion and well-being. A social psychological perspective on mental health”, in *Psychological Bulletin*, 103: 193-210.

Trivers, R. 2011. *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*. New York: Basic Books.

Tulving, E. 1983. *Elements of Episodic Memory*. Oxford: Clarendon Press.

Vendler, Z. 1972. *Res Cogitans: An Essay in Rational Psychology*. Ithaca: Cornell University Press.

Wilson, T., Hodges, S. en Lafleur, S. 1995. “Effects of Introspecting About Reasons: Inferring Attitudes from Accessible Thoughts”, in *Journal of Personality and Social Psychology*, 69:16-28.