Scriptie voor de Academische Master Wijsbegeerte

De Rationaliteit van Moreel Handelen in het Prisoner's Dilemma

Constrained Maximization of Tit-for-Tat?

Door: **Gert Crielaard**

Studentnummer: **3039641**

Eerste begeleider: **Prof. Dr. Marcus Düwell**

Tweede begeleider: **Dr. Thomas Müller**

Oktober 2010

Inhoud

Voorwoord

Ik ben voor het eerst in aanraking gekomen met het Prisoner's Dilemma tijdens een cursus speltheorie dat onderdeel was van mijn masterprogramma in economie. De maatschappelijke relevantie van het model, in het bijzonder de kracht die het heeft om tot de kern te komen van bepaalde milieu gerelateerde vraagstukken, heeft mij altijd geïntrigeerd. In deze scriptie, welke overwegend van filosofische aard is, heb ik opgedane kennis van mijn studie economie geprobeerd te verbinden met kennis en vaardigheden die ik opgedaan heb tijdens mijn filosofiestudie. Ik hoop dat deze multidisciplinaire insteek een interessante benadering van het Prisoner's Dilemma biedt voor zowel de filosofisch als speltheoretisch geïnteresseerde lezer.

Ik wil Prof. Dr. Marcus Düwell hartelijk bedanken voor de uitstekende begeleiding tijdens het schrijven van deze scriptie. Zijn tip om *Morals by Agreement* van David Gauthier als uitgangspunt te nemen voor een filosofische scriptie over speltheorie en ethiek was een schot in de roos. Verder waren de gesprekken die we hadden over de scriptie altijd inhoudelijk waardevol en heb ik ons contact altijd als ontspannen en zeer prettig ervaren. Ook wil ik mijn tweede begeleider Dr. Thomas Müller danken voor een aantal waardevolle kritieken die ik van hem heb ontvangen. Zonder de hulp van beide begeleiders zou deze scriptie niet zijn geworden tot dat wat het nu is.

Deze scriptie markeert voor mij tevens het einde van een prachtige periode waarin ik het voorrecht heb gehad om filosofie te mogen studeren. Het heeft mij in belangrijke mate gevormd tot wie ik nu ben. Filosofie wordt wel eens een eenzame studie genoemd, ik heb echter het tegendeel ervaren. Ik wil bij deze al mijn mooie en inspirerende medestudenten bedanken die de afgelopen vijf jaar mede tot een fantastische tijd hebben gemaakt.

Utrecht, oktober 2010

Introductie

Praktische rationaliteit is de capaciteit die ons in staat stelt te bepalen hoe we moeten handelen. Onder de vele theorieën die er zijn over praktische rationaliteit en rationele keuze is de nutsmaximaliserende conceptie van praktische rationaliteit ongetwijfeld één van de meest breed geaccepteerde theorieën. Deze conceptie van praktische rationaliteit houdt in dat een rationeel individu kiest voor die handeling waarmee het zijn doelen, ongeacht wat deze ook mogen zijn, maximaal realiseert. *Nut* geeft aan in welke mate het individu in staat is zijn doelen te realiseren of zijn persoonlijke behoeften te bevredigen. Soms spreekt men ook wel van *geluk* als synoniem voor nut. Deze filosofisch belangrijke visie op praktische rationaliteit is ondermeer erg invloedrijk in hedendaagse economische disciplines.

Binnen deze theorie van rationele keuze kunnen de keuzes van andere individuen ofwel als onafhankelijke omgevingsvariabelen worden verondersteld ofwel als zijnde afhankelijk van de eigen keuze. Indien dit laatste het geval is handelt het individu rationeel wanneer het kiest voor die handeling die hem het hoogst verwachte nut oplevert, waarbij het effect dat de eigen keuze heeft op de keuze van het andere individu in acht genomen moet worden. Een situatie waarin sprake is van wederzijds afhankelijke keuzes maakt het rationele besluitvormingsproces dus extra complex. Speltheorie is de economische discipline die dit soort situaties systematisch bestudeerd.

Speltheorie wordt gebruikt om menselijk gedrag te verklaren, te voorspellen en te evalueren in situaties waarin de keuzes die men maakt effect hebben op de keuzes van anderen. Speltheorie is daarom ook relevant voor ethiek en politieke filosofie. Er kunnen grofweg drie verschillende onderzoeksgebieden onderscheiden worden waarin een relatie wordt gelegd tussen filosofie en speltheorie. 1) Speltheorie wordt gebruikt om de functie van moraal te identificeren. 2) Speltheorie wordt gebruikt om de uitkomst van

een sociaal contract te verklaren of voorspellen en het naleven van het sociaal contract te rechtvaardigen. 3) Evolutionaire speltheorie wordt gebruikt om het bestaan en de (historische) ontwikkeling van morele principes en gebruiken te verklaren.[1]

Het werk *Morals by Agreement* van David Gauthier is een invloedrijk werk in deze traditie. In dit werk geeft Gauthier onder andere zijn visie op hoe een rechtvaardig sociaal contract er uit ziet en hoe dit tot stand komt. Daarnaast tracht hij een rationele rechtvaardiging te geven voor het naleven van een rechtvaardig sociaal contract in situaties waarin er geen instituties zijn die het sociaal contract bindend kunnen maken. Dit is tevens het onderscheidende aspect van *Morals by Agreement*. Speltheorie wordt niet alleen gebruikt om het sociaal contract te analyseren maar ook om een rechtvaardiging te vinden voor het handelen conform het sociaal contract. Hierin onderscheid Gauthier zich van andere contractualisten zoals Rawls en Harsanyi.[2]

In het voorwoord van *Morals by Agreement* schrijft Gauthier dat het zogenaamde Prisoner's Dilemma een belangrijke aanleiding is geweest voor het schrijven van dit werk. Het Prisoner's Dilemma is een speltheoretische situatie die laat zien dat er een spanningsveld bestaat tussen moraliteit en rationaliteit. Een sociaal contract tussen twee gelijkwaardige rationele spelers in een Prisoner's Dilemma zal bestaan uit een afspraak dat beide spelers coöperatief zullen handelen. De uitkomst die bereikt wordt door wederzijds coöperatief handelen is namelijk de uitkomst waarbij beide spelers hun nut maximaliseren. Vanuit contractualistisch perspectief schrijft de morele imperatief dus voor dat men kiest voor de coöperatieve handeling omdat dit de handeling is die conform is met het sociaal contract.

Rationaliteit schrijft echter voor dat spelers afwijken van de afspraak omdat voor elke speler individueel geldt dat afwijken van de afspraak in overeenstemming is met individuele nutsmaximalisatie. Echter, wanneer beide spelers afwijken van het sociaal contract resulteert dit in een uitkomst waarbij beide een lager nut realiseren dan wanneer beide coöperatief gehandeld

[1] Bruno Verbeek and Christopher Morris, 'Game theory and Ethics,' in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*,
URL = <http://plato.stanford.edu/archives/fall2010/entries/game-ethics/>.
[2] Ibid.

hadden. Gauthier claimt echter dat er een rationele rechtvaardiging bestaat voor de morele handeling in een Prisoner's Dilemma.

In deze scriptie zal ik trachten antwoord te geven op de volgende vragen: 1) Is er binnen conventionele speltheorie een rationele rechtvaardiging te vinden voor moreel handelen in een Prisoner's Dilemma? 2) Als dit het geval is, waarom zoekt Gauthier dan naar een alternatieve rechtvaardiging voor moreel handelen in een Prisoner's Dilemma? 3) Is Gauthier's rechtvaardiging voor moreel handelen in een Prisoner's Dilemma houdbaar?

In hoofdstuk 1 gaan we verder in op het Prisoner's Dilemma. We zullen een uiteenzetting geven van de speltheoretische aspecten van het Prisoner's Dilemma en er zal gekeken worden welke verschillende soorten Prisoner's Dilemmas er zijn. Ook zullen we zien of er binnen conventionele speltheorie een rechtvaardiging gevonden kan worden voor coöperatief handelen binnen een Prisoner's Dilemma. Daarnaast zullen er ook een aantal praktische voorbeelden gegeven worden die de vorm hebben van een Prisoner's Dilemma. Dit teneinde de relevantie van het vraagstuk te benadrukken. In hoofdstuk 2 volgt een uiteenzetting van Gauhtier's theorie. De nadruk zal hier liggen op de rationele rechtvaardiging die Gauthier claimt te geven voor het handelen conform de morele imperatief in een Prisoner's Dilemma. Hoofdstuk 3 zal antwoord moeten geven op de vraag waarom Gauthier zoekt naar een ander soort rechtvaardiging dan die binnen conventionele speltheorie wordt gegeven. In hoofdstuk 4 zal de houdbaarheid besproken worden van een tweetal belangrijke veronderstellingen die Gauthier maakt. In hoofdstuk 5 zal tot slot geprobeerd worden aannemelijk te maken dat Gauthier's rechtvaardiging voor het handelen conform de morele imperatief geen toegevoegde waarde biedt ten opzichte van conventionele speltheorie.

Hoofdstuk 1

Het Prisoner’s Dilemma

Er zijn twee gevangenen die beide zijn gearresteerd wegens het beroven van een bank. De openbaar aanklager die beide gevangenen graag veroordeeld ziet legt beide gevangenen de volgende keuze voor. “Als jullie beide een bekentenis afleggen worden jullie beide aangeklaagd. Ik zal er echter op toe zien dat jullie beide strafvermindering krijgen als beloning voor jullie bekentenis en slechts 3 jaar eisen. Als één van jullie een bekentenis aflegt en de ander niet zal ik alle aanklachten tegen degene die de bekentenis heeft afgelegd laten vallen en de bekentenis gebruiken om de maximale straf van 5 jaar te eisen tegen degene die geen bekentenis heeft afgelegd. Als jullie beide geen bekentenis afleggen zal ik genoegen moeten nemen met een lagere strafeis en leg ik jullie slechts wapenbezit ten laste waarvoor de maximale straf 1 jaar is.” Aangezien beide gevangenen in een aparte cel zitten is het voor hen niet mogelijk met elkaar te communiceren.

De volgende situatie doet zich nu voor. Als gevangene y geen bekentenis aflegt is het voor gevangene x rationeel om wel een bekentenis af te leggen omdat dan alle aanklachten tegen hem ingetrokken zullen worden. Als y wel een bekentenis aflegt is het voor x ook rationeel om een bekentenis af te leggen omdat dit resulteert in een eis van 3 jaar in plaats van 5 jaar. Een bekentenis afleggen is een zogenaamde dominante strategie in het Prisoner’s Dilemma. Ongeacht wat de andere gevangene doet is een bekentenis afleggen altijd de beste strategie.

Het dilemma is nu dat voor beide gevangenen individueel geldt dat ze beter af zijn wanneer ze een bekentenis afleggen. Maar wanneer beide gevangenen een bekentenis afleggen krijgen beide een gevangenisstraf van 3 jaar, terwijl beide 1 jaar hadden gekregen indien beide niet bekend hadden.

Wat nu als beide gevangenen in dezelfde cel hadden gezeten en af hadden kunnen spreken om beide geen bekentenis af te leggen? Indien beide gevangenen coöperatief handelen en zich aan de afspraak houden zouden beide wegkomen met slechts een eis van 1 jaar. Het had voor de uitkomst echter niet uitgemaakt. Beide gevangenen hebben namelijk een rationeel motief om van de afspraak af te wijken. Ongeacht of de ander afwijkt van de afspraak of niet is het immers altijd beter om een bekentenis af te leggen. Aangezien er daarnaast ook geen mogelijkheden zijn om de afspraak bindend te maken zullen beide spelers, indien rationeel, ook daadwerkelijk van de afspraak afwijken als het moment van handelen daar is. Het volgen van de rationele strategie leidt dus tot een collectief ongewenste situatie.

1.1 Het One-Round Prisoner's Dilemma

Het oorspronkelijke model van het Prisoner's Dilemma zoals we zojuist beschreven, is het one-round Prisoner's Dilemma, vanaf nu afgekort als 'PD'. In dit model zijn er twee spelers[3] die slechts één keer interacteren. Beide spelers kennen de structuur van het spel, het aantal speelronden, de verschillende keuzemogelijkheden en de daarbij behorende uitkomsten. Het is voor beide spelers niet mogelijk om bindende afspraken te maken. Het PD kan als volgt schematisch worden weergegeven.

		speler y	
		coöperatief	**non-coöperatief**
speler x	**coöperatief**	(R , R)	(S , T)
	non-coöperatief	(T , S)	(P , P)

T>R>P>S en R>(T+S)/2

(…,…) = (Resultaat speler x, Resultaat speler y)

[3] 'Speler' verwijst overigens niet persé naar een menselijk individu maar kan bijvoorbeeld ook naar een organisatie, land of andere bestuurlijke eenheid verwijzen.

Voor beide spelers geldt dat de non-coöperatieve handeling de dominante strategie vormt. Dit wil zeggen dat, ongeacht de keuze van de andere speler, de non-coöperatieve strategie altijd het beste resultaat geeft. Voor x geldt dat als y voor de coöperatieve handeling kiest non-coöperatief handelen de beste keuze is omdat T>R. Als y kiest voor de non-coöperatieve handeling, is voor x non-coöperatief handelen ook de beste keuze omdat P>S. Voor y geldt hetzelfde als voor x dus beide spelers zullen voor de non-coöperatieve handeling kiezen. Dit resulteert in een Pareto-inefficiënte[4] uitkomst. Voor een Pareto-inefficiënte uitkomst u_i geldt dat alle rationele spelers een andere uitkomst u_e verkiezen boven u_i. Rationeel strategische interactie kan dus resulteren in een uitkomst die collectief ongewenst is. Dit is de kern van het PD, non-coöperatief gedrag domineert coöperatief gedrag en wederzijds coöperatief gedrag wordt door beide spelers verkozen boven wederzijds non-coöperatief gedrag.

1.2 Het Iterated Prisoner's Dilemma

Een ander model is het Iterated Prisoner's Dilemma (IPD). In dit model interacteren de spelers meerdere keren, het PD wordt als het ware herhaald. De twee belangrijkste basisvormen van het IPD zijn het gedetermineerde IPD en het ongedetermineerde IPD. In het gedetermineerde IPD is het aantal speelronden (*n*) bij elke speler bekend. Spelers kennen voor de eerste speelronde de waarde van *n*. In het ongedetermineerde IPD is *n* niet bekend bij de spelers. De spelers kennen alleen een waarschijnlijkheidswaarde (*p*) die aangeeft met welke waarschijnlijkheid er een volgende speelronde komt.[5]

[4] Pareto-efficiëntie: situatie waarin geen enkele speler een hoger resultaat kan behalen zonder dat een andere speler daardoor slechter af is. Een Pareto-inefficiënte situatie is een situatie waarin één of meerdere spelers een hoger resultaat zou kunnen behalen zonder dat één van de andere spelers daardoor slechter af is. In speltheorie wordt efficiëntie gedefinieerd aan de hand van Pareto-efficiëntie.

[5] Deze *p* staat ook wel bekend als 'the shadow of the future'. Een alternatieve interpretatie hiervan is om *p* als discontofactor te beschouwen hetgeen er voor zorgt dat toekomstige resultaten gedevalueerd worden. Dit maakt verder geen verschil voor de uitkomst van het IPD.
Bron: Steven Kuhn, 'Prisoner's Dilemma,' in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*,
URL = <http://plato.stanford.edu/archives/fall2010/entries/prisoner-dilemma/>.

Het interessante van een IPD is dat een speler, het effect dat zijn keuze heeft op de keuze van zijn opponent in de daarop volgende ronden, in acht moet nemen. Zo kan speler x er voor kiezen in de eerste ronde coöperatief te handelen. Wanneer speler y er vervolgens ook voor kiest om coöperatief te handelen (en besluit speler x niet uit te buiten) dan kan x er voor kiezen om speler y in ronde 2 te belonen door wederom coöperatief te handelen. Mocht y er echter voor kiezen misbruik te maken van het coöperatieve gedrag van x dan kan x besluiten om het gedrag van y in ronde 1 te bestraffen door in ronde 2 niet coöperatief te handelen. Omdat men door middel van het volgen van een bepaalde strategie de strategie van de ander in toekomstige speelronden kan beïnvloeden is non-coöperatief handelen niet persé een rationele keuze meer.

Het voorgaande heeft voornamelijk betrekking op het ongedetermineerde IPD. Er is namelijk een argument dat bekend staat als het 'backward induction' argument dat laat zien dat het nooit rationeel kan zijn om in een gedetermineerd IPD coöperatief te handelen. Dit argument verloopt als volgt. Wanneer een gedetermineerd IPD uit *n* ronden bestaat en beide spelers weten dit, dan is het voor beide spelers rationeel om in ronde *n* non-coöperatief te handelen. Maar als spelers van elkaar weten dat ze in ronde *n* non-coöperatief zullen handelen dan is het ook rationeel om in ronde *n*-1 non-coöperatief te handelen. Het zelfde geldt voor ronde *n*-2, *n*-3...etc. en ook voor de eerste ronde.

Voor het ongedetermineerde IPD gaat dit argument niet op. In een ongedetermineerd IPD kan het wel rationeel zijn om coöperatief te handelen. Dit hangt af van de strategie van de andere speler. Een strategie kan men karakteriseren aan de hand van verschillende kenmerken. Hieronder volgen een aantal kenmerken.

- *unconditional defection* = Speler kiest altijd de non-coöperatieve handeling.
- *unconditional cooperation* = Speler kiest altijd de coöperatieve handeling.
- *random* = Speler kiest op willekeurige basis voor de coöperatieve of non-coöperatieve handeling.

- *negation* = Speler kiest in ronde *r* altijd de handeling die tegenovergesteld is aan de handeling van opponent in ronde *r*-1.
- *nice* = Speler kiest nooit als eerste de non-coöperatieve handeling
- *retaliatory* = Als opponent in ronde *r*-1 non-coöperatief heeft gehandeld dan handelt speler in ronde *r* non-coöperatief.
- *forgiving* = Als opponent in ronde < *r*-1 ten minste 1 keer non-coöperatief gehandeld heeft maar in ronde *r*-1 coöperatief heeft gehandeld dan handelt speler in ronde *r* coöperatief.
- *clear* = Het is gemakkelijk voor opponent om de strategie van speler te voorspellen

De vraag die uiteraard meteen op komt is: 'wat is de beste strategie?'. Hier is echter geen eenduidig antwoord op te geven. Gegeven dat *p* hoog genoeg is, is er geen enkele strategie die het beste resultaat oplevert onafhankelijk van haar omgeving.[6] [7] *Unconditional defection* geeft een speler het maximale resultaat wanneer zijn opponent *unconditional cooperation* speelt. Wanneer zijn opponent een strategie volgt die *retaliatory* is, is het wellicht verstandiger om een andere strategie te spelen.

Men kan zich echter wel afvragen welke strategieën succesvol zijn in een omgeving waar andere spelers strategieën gebruiken die er ook op gericht zijn een zo goed mogelijk resultaat te behalen. Robert Axelrod heeft in de jaren tachtig toonaangevend werk verricht[8] met als doel een antwoord te geven op deze vraag. Axelrod organiseerde twee opeenvolgende toernooien waarin professionele speltheoretici werd gevraagd verschillende strategieën te ontwerpen voor IPD's. Vervolgens werd gekeken hoe elke strategie presteerde door elke strategie eenmaal in te zetten tegen iedere andere strategie en eenmaal tegen zichzelf. De strategieën werden vervolgens beoordeeld op basis van de cumulatieve score over het gehele toernooi. De resultaten uit het eerste toernooi werden vervolgens gebruikt voor het daaropvolgende toernooi. De strategie die in beide toernooien het beste scoorde was 'Tit-for-Tat' (TFT).

[6] Robert Axelrod, 'The Emergence of Cooperation among Egoists,' in *The American Political Science Review*, Vol. 75, No. 2 (1981), p. 309.
[7] Wanneer *p* onder een bepaalde drempelwaarde valt zijn de verwachte opbrengsten van coöperatie niet hoog genoeg om de kosten van coöperatie te compenseren en is *unconditional defection* altijd de dominante strategie.
[8] Robert Axelrod, *The Evolution of Cooperation*, New York: Basic Books, 1984.

TFT wordt gekenmerkt door de eigenschappen *nice, retaliatory, forgiving* en *clear*. Dit betekent dat bij TFT in de eerste speelronde altijd gekozen wordt voor coöperatie en dat in alle daaropvolgende ronden de actie van de opponent in de vorige ronde geïmiteerd wordt.

Ondanks de logica van het 'backward induction' argument blijkt TFT in de praktijk overigens ook een succesvolle strategie te zijn in gedetermineerde IPD's. Dit wijst er op dat strikte speltheoretische aannames met betrekking tot rationaliteit niet altijd realistisch zijn.[9] Wanneer men realistische en minder strikte aannames maakt met betrekking tot rationaliteit zou men kunnen argumenteren dat ook TFT een rationeel acceptabele strategie is binnen een gedetermineerd IPD.[10]

In een omgeving waar spelers strategieën gebruiken die er op gericht zijn een zo goed mogelijk resultaat te behalen geldt voor een IPD dat *unconditional defection* dus niet noodzakelijk de enige rationele strategie is. Binnen een IPD is er een rationele basis om een coöperatieve TFT strategie te volgen. Wanneer beide spelers in een IPD een TFT strategie volgen resulteert dit in wederzijds coöperatief handelen in alle speelronden van het IPD. In tegenstelling tot een PD kunnen twee rationele spelers in een IPD dus wel tot een Pareto-efficiënte uitkomst komen.

In de praktijk zijn er tal van situaties te bedenken die de vorm hebben van een IPD. Een goed voorbeeld van een IPD is het prijsbeleid van twee producenten van een homogeen product in een duopolistische markt. Stel dat beide producenten in de uitgangpositie dezelfde prijs hanteren voor hun product. Voor beide producenten geldt dat het in dit geval aantrekkelijk is de prijs te verlagen en zo de winst te verhogen door een grote toename in de afzet, een toename die ten koste gaat van de afzet van de concurrerende

---

[9] Steven Kuhn, 'Prisoner's Dilemma,' in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*,
URL = <http://plato.stanford.edu/archives/fall2010/entries/prisoner-dilemma/>.

[10] Een vergelijkbare vorm van rederneren vinden we bij de 'surpirse exam paradox'. Studenten die weten dat ze in de volgende week op een niet nader gespecificeerde dag een onverwacht tentamen zullen krijgen kunnen als volgt rederneren. Als we op donderdag geen tentamen hebben gehad weten we dat we deze op vrijdag zullen krijgen. Maar als het een 'onverwacht' tentamen is, kan deze dus niet op vrijdag gegeven worden. Maar voor donderdag, woensdag, dinsdag en maandag geldt vervolgens hetzelfde. Volgens deze vorm van redeneren zou een onverwacht tentamen dus niet mogelijk zijn. De logica van dit argument is uiteraard omstreden. In dat licht zou men ook vraagtekens kunnen zetten bij de logica van het 'backward induction' argument.

producent. Wanneer beide producenten deze strategie volgen komt men echter in een neerwaartse prijsspiraal terecht die nadelig is voor beide producenten. In de uitgangspositie zullen beide producenten dus na moeten denken wat het effect van hun prijsbeleid zal zijn op het prijsbeleid van de concurrent. Een denkbare strategie zou dan *nice* en *retaliatory* kunnen zijn waarbij een producent alleen tot prijsverlaging overgaat als zijn concurrent dit ook doet, in de hoop dat zijn concurrent dezelfde strategie volgt en een neerwaartse prijsspiraal voorkomen wordt.

De wapenwedloop tussen de V.S. en de voormalige Sovjet-Unie zou men ook kunnen zien als een voorbeeld van een IPD. Beide landen volgden een strategie die resulteerde in een voor beide landen collectief ongewenste situatie. Nadat er afspraken gemaakt werden over ontwapening volgden beide landen een *nice*, *retaliatory* en *forgiving* strategie waarbij er wisselend sprake was van ontwapening en bewapening.

1.3 Het Multiplayer Prisoner's Dilemma

Er is ook een vorm van het Prisoner's Dilemma mogelijk waarin meerdere spelers participeren. In het Multiplayer Prisoner's Dilemma (MPD) gelden wederom dezelfde voorwaarden als in het PD. Alle spelers kennen de structuur van het spel, het aantal speelronden, de verschillende keuzemogelijkheden en de daarbij behorende uitkomsten. Het is voor spelers niet mogelijk om bindende afspraken te maken. Een MPD kan er als volgt uit zien.

		rest van de populatie	
		meer dan *n* coöperatief	**minder dan *n* coöperatief**
speler x	**coöperatief**	(B-C, B-C)	(D-C, D)
	non-coöperatief	(B , B-C)	(D , D)

B > B-C > D > D-C en (B-C)>(B+(D-C))/2

n = minimale niveau van effectieve coöperatie

Speler x staat voor het volgende dilemma. Als meer dan een bepaalde hoeveelheid *n* spelers binnen de populatie coöperatief handelt resulteert dit in een collectief positief resultaat (B) voor de gehele populatie. Wanneer minder dan een bepaalde hoeveelheid *n* spelers binnen de populatie coöperatief handelt resulteert dit in een collectieve rampspoed (D) voor de gehele populatie. Coöperatief handelen brengt bepaalde kosten (C) met zich mee. Hier geldt wederom dat non-coöperatief handelen de dominante strategie vormt. In het geval dat *n* overschreden wordt geldt B > B-C en in het geval dat *n* niet overschreden wordt geldt D > D-C. Wanneer aangenomen wordt dat alle spelers rationeel zijn leidt deze situatie onvermijdelijk tot het Pareto-inefficiënte resultaat waar iedere speler rampspoed treft.

In deze situatie is het Pareto-efficiënte resultaat overigens niet dat resultaat waarbij elke speler coöperatief handelt. Het Pareto-efficiënte resultaat is dat resultaat waarbij precies *n* spelers in de populatie coöperatief handelen. Wanneer meer spelers coöperatief zouden handelen zou dit immers resulteren in overbodige kosten (C). De spelers die in dit geval niet-coöperatief handelen en die wel profiteren van het coöperatieve gedrag van de andere spelers worden ook wel free-riders genoemd.

In de praktijk komen dergelijke problemen bijvoorbeeld voor wanneer er sprake is van gemeenschappelijk bezit. Garrett Hardin noemde dit ook wel ‘the tradegy of the commons’. Wanneer een populatie van rationele individuen vrijelijk toegang heeft tot een schaarse natuurlijke hulpbron die slechts een beperkte capaciteit heeft zal dit uiteindelijk resulteren in onduurzaam gebruik en uitputting van die hulpbron.

De overbevissing van de oceanen is een voorbeeld hiervan. Door een toename in de vraag naar bepaalde soorten vis staat de voortplanting van sommige vissoorten onder druk. De toekomst van de visserij die afhankelijk is van dergelijke soorten staat daardoor ook onder druk. Vissers zouden daarom beter coöperatief kunnen handelen en af kunnen spreken om zich aan bepaalde quota te houden waardoor duurzame visvangst gegarandeerd wordt. Iedere individuele visser heeft echter een rationeel motief om van die afspraak af te wijken. De nadelige gevolgen van zijn overschrijding worden immers gedragen door alle vissers terwijl de inkomsten van zijn overschrijding slechts en alleen aan hemzelf toekomen. Aangezien het lastig is om internationaal bindende afspraken te maken is overbevissing het gevolg.

Het klimaatprobleem is een ander voorbeeld van een MPD. Door een toenemende uitstoot van broeikasgassen warmt de aarde steeds verder op met alle nadelige gevolgen van dien. Er zijn daarom internationale afspraken gemaakt, zoals het verdrag van Kyoto en Kopenhagen voor het terugbrengen van de uitstoot van broeikasgassen. Maar omdat ieder land er individueel belang bij heeft om van die afspraken af te wijken en het lastig is om internationaal harde en bindende afspraken te maken is er nog steeds geen sprake van effectieve samenwerking. Daarnaast proberen de verschillende landen met allerlei argumenten de verantwoordelijkheid van zich af te schuiven en free-rides te krijgen ten koste van de andere landen.

Voor het MPD zijn uiteraard ook herhaalde (iterated) versies te bedenken. Men zou het klimaatprobleem ook kunnen zien als een herhaald MPD. Wanneer een land zich niet aan de afspraken houdt die bijvoorbeeld in het Kyoto of Kopenhagen verdrag zijn gemaakt, dan kan het daarvoor later mogelijk bestraft worden door andere landen. Het afwijken van de afspraak zou in dit geval minder voordelig kunnen zijn omdat een land in de toekomst mogelijk door andere landen uitgesloten kan worden van voordelige coöperatieve verbanden. Non-coöperatief handelen is dan niet noodzakelijk de dominante strategie.

1.4 Het Evolutionary Prisoner's Dilemma

Een andere veel besproken vorm van het Prisoner's Dilemma is het Evolutionary Prisoner's Dilemma (EPD). Stel er is een populatie van spelers die IPD's met elkaar spelen en verschillende strategieën volgen. Naarmate de tijd vordert zullen succesvolle strategieën in de populatie toenemen in aantal en minder succesvolle strategieën afnemen.[11] Iedere speler wil immers het beste resultaat behalen en zal daarom proberen de meest succesvolle strategie te volgen. Er vindt dus evolutie plaats in het spelgedrag van de spelers in de populatie.

In een EPD kan men het lange termijn succes van een bepaalde strategie beoordelen door te kijken naar haar zogenaamde evolutionaire stabiliteit. Dit concept veronderstelt een populatie die in haar geheel volgens één bepaalde strategie (B) handelt en één afwijkend individu die volgens een andere strategie (A) handelt.[12] Axelrod definieerd evolutionaire stabiliteit als volgt.

"A strategy is *collectively stable* if no strategy can invade it." [13]

"Strategy A is said to invade strategy B if V(A | B) > V(B | B) where V(A | B) is the expected payoff an A gets when playing a B, and V(B | B) is the expected payoff a B gets when playing another B." [14]

Een strategie B is volgens Axelrod stabiel wanneer er geen enkele andere strategie A bestaat waarvan het verwachte resultaat (V) hoger is dan het verwachte resultaat van strategie B. Dit betekent dat bij een collectief stabiele strategie B per definitie gekozen wordt voor de non-coöperatieve handeling wanneer het cumulatieve resultaat van A hoger dreigt te worden dan het cumulatieve resultaat van A. Op deze manier zorgt de speler van strategie B er voor dat het verwachte resultaat van strategie A nooit hoger wordt dan dat van

[11] Steven Kuhn, 'Prisoner's Dilemma,' in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*,
URL = <http://plato.stanford.edu/archives/fall2010/entries/prisoner-dilemma/>.
[12] Robert Axelrod, 'The Emergence of Cooperation among Egoists,' in *The American Political Science Review*, Vol. 75, No. 2 (1981), p. 310.
[13] Ibid.
[14] Ibid.

strategie B. Dit is zowel een noodzakelijke als een voldoende voorwaarde voor een collectief stabiele strategie.[15]

Axelrod laat zien dat TFT, gegeven dat *p* hoog genoeg is, in een EPD een collectief stabiele strategie is. TFT is een *nice* strategie welke alleen collectief stabiel kan zijn wanneer deze ook *retaliatory* is. Wanneer een *nice* strategie niet *retaliatory* zou zijn in een willekeurige ronde *r* zou er namelijk een strategie denkbaar zijn waarbij alleen gekozen wordt voor de non-coöperatieve handeling in ronde *r*-1 en welke dus een hoger cumulatief resultaat zou behalen, zo argumenteert Axelrod. TFT is beide *nice* en *retaliatory* en is daarom een collectief stabiele strategie. De strategie *unconditional defection* is echter ook een stabiele strategie. Hierbij wordt namelijk altijd gekozen voor de non-coöperatieve handeling en voldoet dus ook aan de definitie van een collectief stabiele strategie.[16]

Men kan zich nu afvragen hoe coöperatie tussen individuen tot stand komt wanneer we ons een soort Hobbesiaanse 'state of nature' voorstellen waarin iedereen de strategie *unconditional defection* volgt. Het feit dat *unconditional defection* een collectief stabiele strategie is impliceert dat wanneer een enkel individu TFT speelt in een wereld waar de rest de strategie *unconditional defection* volgt het individu dat TFT speelt per definitie slechter af is. Het individu dat TFT speelt behaalt in ronde 1 immers het S-resultaat terwijl zijn opponent die *unconditional defection* speelt het T-resultaat behaald.

Axelrod laat echter zien dat wanneer nieuwkomers die een TFT strategie volgen in clusters komen, deze een kans hebben om te overleven. Volgens Axelrod kan TFT, afhankelijk van *p*, al overleven bij clusters die slechts een zeer klein deel uitmaken van de populatie. Wanneer V(TFT | *unconditional defection*) > V(*unconditional defection* | *unconditional defection*) zal de evolutie van TFT uiteindelijk ten koste gaan van het bestaan van *unconditional defection*. Verder demonstreert Axelrod dat een *nice* strategie die collectief stabiel is zichzelf net zo goed kan wapenen tegen een cluster van nieuwkomers als tegen een enkel individu. Dus strategieën die *nice* zijn hebben niet de structurele zwakte ten opzichte van clusters van nieuwkomers

[15] Ibid., pp. 313-314.
[16] Ibid., pp. 314-315.

als de strategie *unconditional defection*. Axelrod geeft hiermee een evolutionaire verklaring en een rationele basis voor het bestaan van coöperatief gedrag binnen een populatie waar men IPD's met elkaar speelt.[17]

[17] Ibid., pp. 315-317.

Hoofdstuk 2

Constrained Maximization

2.1 Gauthier's rechtvaardiging van de morele handeling

In *Morals by Agreement* probeert Gauthier antwoord te geven op de vraag wat een rechtvaardig sociaal contract is en op de vraag waarom we ons aan een dergelijk contract zouden moeten houden. Of anders gezegd, waarom moreel handelen? In het beantwoorden van deze vragen maakt Gauthier gebruik van speltheorie. Gauthier geeft eerst zijn visie op de natuurlijke toestand van het menselijk individu en de praktische rede. Zijn beschrijving van het menselijk individu komt veelal overeen met het idee van het rationele individu dat streeft naar nutsmaximalisatie zoals we deze beschreven in de inleiding. Vervolgens wordt een antwoord gegeven op de vraag hoe een rechtvaardig sociaal contract er uit ziet. Gauthier ziet een rechtvaardig sociaal contract als een hypothetische verzameling van handelingprincipes waartoe een groep rationele individuen zou besluiten. De totstandkoming van een dergelijk sociaal contract heeft volgens Gauthier het karakter van een onderhandelingsproces zoals we deze kennen uit de speltheorie. Daarna geeft Gauthier antwoord op de vraag waarom we ons aan een sociaal contract zouden moeten houden. Door een alternatieve theorie van praktische rationaliteit te geven tracht Gauthier te bewijzen dat het rationeel is om te handelen conform het sociaal contract. Tot slot argumenteert Gauthier dat de handelingsprincipes die rationele individuen op gelijkwaardige basis overeenkomen ook morele principes zijn. Gauthier denkt zo dus een rationele rechtvaardiging te hebben gevonden voor morele principes en moreel handelen.

Deze scriptie heeft vooral betrekking op de vraag waarom we moreel zouden moeten handelen. Gauthier's antwoord op deze vraag is 'Constrained Maximization'. Wat dit precies inhoudt zullen we nu bespreken.

We hebben gezien dat wanneer twee rationele spelers af kunnen spreken over hoe te handelen in een PD situatie beide spelers met elkaar af zullen spreken die handeling te kiezen die de Pareto-efficiënte uitkomst geeft. Een rationele speler probeert immers zijn nut te maximaliseren en zou van alle mogelijke uitkomsten nooit instemmen met een uitkomst die een hoger nut oplevert voor zijn opponent maar een lager nut oplevert voor hemzelf. Anders gezegd, in een PD waar twee rationele spelers zonder coöperatief handelen het P-resultaat behalen kan het nooit zo zijn dat beide spelers instemmen met de (T,S)-uitkomst, in dit geval is één van beide spelers immers slechter af. Beide spelers zullen dus overeenkomen voor de (R,R)-uitkomst te gaan. Voor deze uitkomst geldt dat beide spelers beter af zijn zonder dat hun opponent slechter af is en dus is deze uitkomst Pareto-efficiënt.

Vanuit een contractualistisch perspectief is datgene wat twee gelijke rationele individuen in een neutrale situatie af zouden spreken moreel imperatief. In een situatie die de vorm heeft van een PD is de coöperatieve handeling dan ook de moreel juiste handeling. In het PD hebben beide spelers echter een rationeel motief om van de afspraak af te wijken. Non-coöperatie is immers de dominante strategie. Het bovenstaande in acht genomen moet men dus concluderen dat binnen een PD de morele handeling en de rationele handeling in strijd zijn met elkaar.

Gauthier claimt echter dat men kan handelen conform de morele imperatief in een PD zonder dat dit in strijd is met rationaliteit. Gauthier noemt dit *Constrained Maximization* (CM). Hij contrasteert dit met *Straightforward Maximization* (SM), de dominante strategie in een PD zoals we die kennen uit conventionele speltheorie. CM is de strategie waarbij de speler een dispositie heeft tot coöperatief handelen. Dit wil zeggen dat een speler alleen coöperatief handelt indien zijn opponent diezelfde dispositie tot coöperatief handelen heeft. Heeft zijn opponent een dispositie tot non-coöperatief handelen dan schrijft CM de non-coöperatieve handeling voor.[18]

[18] Als CM een rationele strategie is impliceert dit tevens dat CM de non-coöperatieve handeling voorschrijft indien een opponent onvoorwaardelijk coöperatief handelt. De

Stel speler x is een *Constrained Maximizer* (CM-er) en heeft een dispositie tot coöperatief handelen. Als speler y een *Straightforward Maximizer* (SM-er) is en dus een dispositie tot non-coöperatief handelen heeft zal x dus ook voor de non-coöperatieve handeling kiezen. Als speler y een CM-er is en net als x een dispositie heeft tot coöperatief handelen bestaat er een wederzijdse dispositie tot coöperatief handelen en is het voor beide spelers rationeel om coöperatief te handelen, aldus Gauthier.

CM rust op twee belangrijke aannames. 1) Er heerst transparantie, dit wil zeggen dat de speler elkaars dispositie kennen. 2) De handeling van de spelers is altijd conform hun dispositie. Beide aannames zullen nu achtereenvolgens besproken worden.

2.2 Transparantie

Gauthier realiseert zich dat zijn argument rust op de aanname dat beide spelers volkomen op de hoogte zijn van elkaars dispositie. Transparantie is een noodzakelijke voorwaarde voor coöperatief handelen. Het is immers alleen wanneer spelers op de hoogte zijn van elkaars dispositie dat rationele spelers een coöperatieve dispositie vormen. Door het vormen van een coöperatieve dispositie die door de opponent herkend wordt hoopt een speler zijn opponent er toe te bewegen ook een coöperatieve dispositie in te nemen. In het geval dat er geen transparantie heerst hebben spelers geen enkele reden om een coöperatieve dispositie te vormen. Wanneer er geen transparantie heerst is CM geen rationele strategie.

De vraag is uiteraard hoe deze aanname gerechtvaardigd wordt. Gauthier geeft hiervoor het volgende argument.

"Since our argument is to be applied to ideally rational persons, we may simply add another idealizing assumption, and take our persons to be *transparent*. Each is directly aware of the

---

coöperatieve handeling zou in dit geval immers niet rationeel zijn. Gauthier is hier echter niet expliciet over.

disposition of his fellows, and so aware whether he is interacting with straightforward or constrained maximizers. Deception is impossible[…]" [19]

Gauthier stelt voor dat we binnen een theoretisch kader vereenvoudigende assumpties mogen maken. Net zoals aangenomen wordt dat de spelers in een PD volledig rationeel zijn kan men ook aannemen dat er volledige informatie is en dat spelers elkaars dispositie kennen.

Gauthier realiseert zich echter dat een dergelijke aanname zijn argument minder interessant zou maken en dat dit niet zou bewijzen dat CM ook onder realistische omstandigheden een rationele strategie is. Om deze reden kiest hij voor een minder sterke aanname, een soort quasi-transparantie.

"We may appeal instead to a more realistic *translucency*, supposing that persons are neither transparent nor opaque, so that their disposition to co-operate or not may be ascertained by others, not with certainty, but as more than mere guesswork."[20]

Spelers zijn dus in staat elkaars dispositie in meer of mindere mate in te schatten maar kunnen nooit honderd procent zeker zijn van elkaars dispositie. Gauthier erkent verder dat het essentieel is voor een CM-er dat zijn capaciteit om de dispositie van de ander in te kunnen schatten goed ontwikkeld is en getraind moet worden.

"The ability to detect the dispositions of others must be well developed in a rational CM. Failure to develop this ability, or neglect of its exercise, will preclude one from benefiting from constrained maximization."[21]

In hoofdstuk 4 zullen we een kritische bespreking geven van deze aanname.

[19] David Gauthier, *Morals by Agreement*, New York: Oxford University Press 1986, pp. 173-174.
[20] Ibid., p.174.
[21] Ibid., p. 181.

2.3 Rationality of Perseverance Principle

Stel spelers x en y zijn beide CM-ers, in dit geval hebben beide spelers dus een dispositie tot coöperatief handelen. Beide spelers hebben er echter ook belang bij van deze dispositie af te wijken wanneer men tot daadwerkelijk handelen overgaat. Als y een dispositie tot coöperatief handelen laat zien en daadwerkelijk conform deze dispositie handelt is het voor x aantrekkelijk om non-coöperatief te handelen, dit levert hem immers het T-resultaat op. Ook wanneer y wel de coöperatieve dispositie laat zien maar niet coöperatief handelt is het voor x aantrekkelijker om niet coöperatief te handelen want dit levert hem het P-resultaat op in plaats van het S-resultaat. Wat voor x geldt, geldt uiteraard ook voor y en dit betekent dat non-coöperatief handelen in dit geval voor beide spelers nog steeds de dominante strategie is.

Om dit te doorbreken moet aannemelijk gemaakt worden dat spelers altijd conform hun dispositie handelen. Wat Gauthier lijkt te beweren is dat de dispositie van een rationeel persoon *ceteris paribus*[22] noodzakelijk een rationele handeling voortbrengt die conform is met die dispositie.

"The Fool rejects what would seem to be the ordinary view that, given neither unforeseen circumstances nor misrepresentation of terms, it is rational to comply with an agreement if it is rational to make it."[23]

"The dispositions of a fully rational actor issue in rational choices."[24]

In haar bespreking van Gauhtier noemt Holly Smith deze stelling het 'Rationality of Perseverance Principle' (RPP).

"RPP: If it is rational for an agent to form the intention to do A, then it is rational for the agent to actually do A when the time comes (assuming the agent acquires no new information, and has not altered her values)." [25]

[22] De noodzakelijkheid van de handeling conform de dispositie geldt alleen wanneer er geen relevante veranderingen plaats vinden. Mocht er bijvoorbeeld een verandering plaats vinden in het verwachte resultaat van de coöperatieve handeling dan kan de handeling afwijken van de oorspronkelijke dispositie.
[23] David Gauthier, *Morals by Agreement*, New York: Oxford University Press 1986, p. 165.
[24] Ibid., p. 187.

Wanneer het voor een speler rationeel is om een bepaalde intentie te vormen, is het *ceteris paribus* ook rationeel om conform die intentie te handelen.[26] Smith merkt terecht op dat Gauthier verder geen positief argument geeft voor de houdbaarheid van RPP. In hoofdstuk 4 komen we hierop terug en zal ook de houdbaarheid van deze aanname ter discussie gesteld worden.

2.4 CM als succesvolle strategie in het Prisoner's Dilemma

Onder de aanname van transparantie en RPP is CM volgens Gauthier superieur ten opzichte van SM. Ongeacht de strategie van de andere speler realiseert CM altijd een gelijkwaardig of beter resultaat dan SM. Gauthier formuleert het als volgt.

"What we have shown is that, if the straightforward maximizer and the constrained maximizer appear in their true colours, then the constrained maximizer must to better."[27]

Indien spelers rationeel zijn zullen zij dan ook altijd de strategie CM volgen. Dit resulteert er in dat in een PD beide spelers, conform de morele imperatief, coöperatief zullen handelen en dus een Pareto-efficiënte uitkomst realiseren. Daarbij moet benadrukt worden dat, wanneer de aanname van transparantie vervangen wordt door quasi-transparantie, het voorgaande alleen geldt indien de capaciteit van spelers om de disposities van hun opponenten te herkennen goed genoeg ontwikkeld is, aldus Gauthier.

CM is er op gericht een Pareto-efficiënt resultaat te realiseren in een PD. Met de aannames die Gauthier doet geldt voor CM echter ook dat het een Pareto-efficiënt resultaat realiseert in IPD's en MPD's. Een IPD kan men immers zien als simpelweg een aaneenschakeling van PD's waar voorafgaand

---

[25] Holly Smith, 'Deriving Morality from Rationality,' in Peter Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on David Gauhtier's Morals by Agreement* (Cambridge: Cambridge University Press, 1991), p. 244.

[26] Omdat Gauthier spreekt van disposities moeten we intenties hier interpreteren als *voorwaardelijke intenties*: intenties die alleen in actie resulteren als aan een bepaalde voorwaarde wordt voldaan.

[27] David Gauthier, *Morals by Agreement*, New York: Oxford University Press 1986, p. 173.

aan iedere ronde de spelers elkaars dispositie weer inschatten. Wanneer transparantie verondersteld wordt betekent dit dat het rationeel is om CM te volgen hetgeen resulteert in wederzijds coöperatief handelen en een Pareto-efficiënt resultaat. In MPD zal na een onderhandelingsproces vastgesteld worden hoe de verschillende spelers zullen moeten handelen om tot een Pareto-efficiënt resultaat te komen. Wanneer er transparantie heerst en spelers perfect op de hoogte zijn van elkaar disposities is het rationeel om te handelen conform de afspraken. In een MPD wordt dus net als in een PD een Pareto-efficiënt resultaat gerealiseerd.

Gauthier claimt verder dat CM ook een collectief stabiele strategie is in een EPD.

"We have claimed that a population of constrained maximizers would be rationally stable; no one would have reason to dispose herself to straightforward maximazation."[28]

In Axelrod's termen zou men dit als volgt kunnen formuleren: V(CM | CM) > V(SM | CM). In een populatie die in zijn geheel uit CM-ers bestaat loont het niet om als enige de strategie SM te volgen. Wanneer er transparantie heerst is het immers niet mogelijk om als SM-er te profiteren van de coöperatieve handelingen van andere spelers.

Verder concludeerd Gauthier dat SM ook een collectief stabiele strategie is en dat het voor een enkel individu niet loont om, in een populatie die volledig uit SM-ers bestaat, coöperatief te handelen.

"If a population of reciprocal altruists is genetically stable, surely a population of egoists is also stable."[29]

"In a world of Fooles, it would not pay to be a constrained maximizer, and to comply with one's agreements. In such circumstances it would not be rational to be moral."[30]

"A small proportion of CMs might well suffer more from exploitation by undetected SMs than by co-operation among themselves unless their capacities for detecting the dispositions of others were extraordinarily effective. Similarly, a mutant reciprocal altruist would be at

[28] Ibid., p. 187.
[29] Ibid., p. 188.
[30] Ibid., pp. 181-182.

disadvantage among egoists; her attempts at co-operation would be rebuffed and she would lose by her efforts in making them."[31]

In deze context refereert Gauthier met de termen 'egoists' en 'Fooles' naar SM-ers en met de term 'reciprocal altruists' naar CM-ers.[32] We formuleren dit dan ook als V(SM | SM) = V(CM | SM) in het geval dat er volledige transparantie heerst en als V(SM | SM) > V(CM | SM) in het geval dat er quasi-transparantie heerst. Als iedereen in de populatie de strategie SM volgt biedt een voorwaardelijk coöperatieve strategie als CM geen enkel voordeel. Wanneer een CM-er perfect op de hoogte is van de disposities in de rest van de populatie zal dit er in resulteren dat de CM-er altijd voor de non-coöperatieve handeling kiest hetgeen uiteindelijk hetzelfde resultaat geeft als SM. Wanneer we uit gaan van quasi-transparantie zal een CM-er de dispositie van zijn opponenten nu en dan verkeerd inschatten en afwisselend het P-resultaat en het S-resultaat behalen. De CM-er is hier dus in het nadeel ten opzichte van de SM-er die altijd op zijn minst het P-resultaat haalt.

Men kan zich hier ook afvragen hoe coöperatie tussen individuen tot stand komt wanneer we ons een Hobbesiaanse 'state of nature' voorstellen waarin iedereen de strategie SM volgt. Gauthier redeneert in de lijn van Axelrod, het succes van het volgen van CM hangt af van het deel van de populatie dat CM volgt. Hoe meer mensen binnen de populatie CM volgen hoe groter het succes van die strategie zal zijn.

"As we have seen, the argument for the rationality of constrained maximization turns on the proportion of CMs in the population."[33]

Verder concludeert Gauthier dat CM uiteindelijk een evolutionair voordeel heeft ten opzichte van SM.

---

[31] Ibid., p. 188.

[32] De term 'Foole' ontleent Gauthier aan de terminologie van Thomas Hobbes. The Foole representeert alles wat de CM-er niet is. De manier van denken van de Foole is tegengesteld aan de manier van denken van de CM-er en wordt dus vaak gebruikt als symbool voor de SM-ers. 'Reciprocal altruists' en 'egoists' zijn synoniemen voor respectievelijk CM-ers en SM-ers. "Similarly, if we think of constrained and straightforward maximization as parallel to genetic tendencies to reciprocal altruism and egoism[…]" Bron: David Gauthier, *Morals by Agreement*, New York: Oxford University Press 1986, p. 187.

[33] David Gauthier, *Morals by Agreement*, New York: Oxford University Press 1986, p. 188.

"Does it then follow that we should expect both groups of reciprocal altruists and groups of egoists to exist stably in this world? Not necessarily. The benefits of co-operation ensure that, in any given set of circumstances, each member of a group of reciprocal altruist should do better than a corresponding member of a group of egoists. Each reciprocal altruist should have a reproductive advantage. Groups of reciprocal altruists should therefore increase relative to groups of egoists in environments in which the two come into contact. The altruists must prevail–not in direct combat between the two (although the co-operation possible among reciprocal altruists may bring victory there), but in the indirect combat for evolutionary survival in a world of limited resources."[34]

De positieve effecten van de onderlinge coöperatieve interactie tussen CM-ers zijn groter dan de nadelige effecten van het verkeerd inschatten van dispositie. Dit zorgt er voor dat een groep CM-ers beter in staat is om te overleven dan een groep SM-ers. Groepen met CM-ers nemen toe ten koste van groepen met SM-ers, iedere CM-er heeft een evolutionair voordeel ten opzichte van een SM-er. Om deze reden wint CM de evolutionaire strijd van SM, aldus Gauthier.

[34] Ibid.

Hoofdstuk 3

CM versus TFT

In hoofdstuk 1 hebben we gezien dat er binnen conventionele speltheorie voor verschillende vormen van het Prisoner's Dilemma een rationele basis te vinden is voor een handeling die conform is met de morele imperatief. Van TFT is gebleken dat het een succesvolle strategie is in IPD's en EPD's en dus een rationele basis biedt voor coöperatief handelen. In hoofdstuk 2 hebben we gezien hoe Gauthier argumenteert dat CM een rationele strategie is in het Prisoner's Dilemma. Wanneer we aannemen dat er transparantie heerst is CM in iedere vorm van het Prisoner's Dilemma een rationele strategie. Als we aannemen dat er sprake is van quasi-transparantie is de rationaliteit van CM enerzijds afhankelijk van de capaciteit van spelers om de disposities van hun opponenten in te schatten en anderzijds van het aantal CM-ers in de populatie.

TFT en CM zijn beide voorwaardelijke coöperatieve strategieën, strategieën die alleen de coöperatieve handeling voorschrijven indien de opponent een dispositie tot coöperatief handelen laat zien. Een TFT speler handelt coöperatief onder de voorwaarde dat zijn opponent de vorige speelronde coöperatief heeft gehandeld en dus zijn bereidheid tot coöperatief handelen heeft laten zien. Een CM-er handelt coöperatief onder de voorwaarde dat hij bij zijn opponent een dispositie tot voorwaardelijk coöperatief handelen waarneemt.

Voor een IPD geldt dat men de coöperatieve handeling, de handeling die conform is met de morele imperatief, zowel via CM als via TFT een rationele rechtvaardiging kan geven. Een CM-er zal bij elke speelronde een voorwaardelijke coöperatieve dispositie moeten laten zien omdat zijn opponent deze dispositie herkent en alleen coöperatief handelt indien deze dispositie van voorwaardelijke coöperatieve aard is. Een TFT speler kan er voor kiezen coöperatief te handelen omdat dit een optimale strategie is in een

omgeving waar andere spelers ook strategieën gebruiken die er op gericht zijn een zo goed mogelijk resultaat te behalen. Alleen via een voorwaardelijke coöperatieve strategie kan men een andere rationele speler er toe bewegen ook coöperatief te handelen en zo een Pareto-efficiënt resultaat behalen.

In een evolutionaire setting trekt Gauthier dezelfde conclusies met betrekking tot CM als Axelrod trekt met betrekking tot TFT. Beide concluderen dat: 1) In een populatie waarin iedereen een dispositie heeft tot coöperatief handelen het geen zin heeft om een non-coöperatieve strategie te volgen. 2) In een populatie waarin iedereen onvoorwaardelijk non-coöperatief handelt het geen zin heeft om een voorwaardelijk coöperatieve strategie te volgen. 3) De mate van succes van een voorwaardelijke coöperatieve strategie af hangt van het deel van de populatie dat een voorwaardelijke coöperatieve strategie volgt. 4) Voorwaardelijke coöperatieve strategieën zich beter kunnen wapenen tegen non-coöperatieve strategieën en daarom evolutionair voordeel hebben ten opzichte van non-coöperatieve strategieën.

Het voorgaande doet de vraag rijzen of we CM niet moeten zien als een strategie die parallel is aan TFT. Op een aantal vlakken komen Gauthier en Axelrod immers tot exact dezelfde conclusies. Gauthier geeft echter duidelijk aan dat men CM niet moet zien als een TFT-achtige strategie.

"[…]constrained maximization is not straightforward maximization in its most effective disguise. The constrained maxmizer is not merely the person who, taking a larger view than her fellows, serves her overall interest by sacrificing the immediate benefits of ignoring joint strategies and violating co-operative arrangements in order to obtain the long-run benefits of being trusted by others." [35]

"Thus constrained maximization is not parallel to such strategies as 'tit-for-tat' that have been advocated for so-called iterated Prisoner's Dilemmas. Constrained maximizers may co-operate even if neither expects her choice to affect future situations. Thus our treatment of co-operation does not make the appeal to reciprocity necessary to Robert Axelrod's account; see 'The Emergence of Cooperation among Egoists'[…]."[36]

Volgens Gauthier onderscheidt CM zich in een belangrijk opzicht van TFT. Een speler die een TFT strategie volgt handelt alleen coöperatief omdat hij

[35] Ibid., p. 169.
[36] Ibid., pp.169-170: voetnoot 19.

verwacht daardoor in de toekomst een beter resultaat te halen. TFT is dus gebaseerd op reciprociteit. Men hoopt dat door coöperatief te handelen men de opponent beweegt tot coöperatief handelen in de toekomst. Volgens Gauthier is dit een verkapte vorm van SM, 'straightforward maximization in its most effective disguise'. Een TFT speler kijkt verder dan een enkele speelronde en beoordeelt vervolgens dat coöperatief handelen hem op lange termijn het hoogste verwachte resultaat oplevert. Een CM-er daarentegen kan ook kiezen voor de coöperatieve handeling zelfs als die handeling geen enkel effect heeft op het resultaat in toekomstige speelronden. CM is gebaseerd op transparantie en RPP. Men neemt een voorwaardelijk coöperatieve dispositie aan welke herkend wordt door de opponent die daardoor ook coöperatief zal handelen.

De vraag is nu waarom CM volgens Gauthier een betere rechtvaardiging biedt voor het handelen conform de morele imperatief dan TFT. Er zijn een aantal denkbare redenen mogelijk.

1) TFT biedt geen rechtvaardiging voor de coöperatieve handeling binnen een PD en MPD. Aangezien PD's en MPD's slechts uit een enkele speelronde bestaan is er geen basis voor reciprociteit en is het niet rationeel om een TFT strategie te volgen. In Hoofdstuk 1 zagen we al dat de non-coöperatieve handeling altijd de dominante strategie vormt in deze vormen van het Prisoner's Dilemma.

2) Als *p* niet hoog genoeg is in een herhaalde vorm van het Prisoner's Dilemma is er eveneens geen rationele basis om TFT te spelen. De toekomstige voordelen van coöperatief handelen zijn dan te onzeker of van te geringe waarde om de negatieve aspecten van de coöperatieve handeling te kunnen compenseren. Non-coöperatief handelen vormt in dit geval altijd de dominante strategie.

3) In een Prisoner's Dilemma waarbij er sprake is van incidentele asymmetrische informatie bestaat de mogelijkheid tot 'vals spelen'. We kunnen ons een IPD voorstellen waarin een TFT speler verwacht dat een incidentele afwijking van de coöperatieve handeling in ronde *r* onopgemerkt zal blijven door zijn opponent. In deze situatie is het voor een TFT speler rationeel om te kiezen voor de non-coöperatieve handeling. We noemen dit ook wel vals spelen, een speler simuleert een consistente TFT strategie maar wijkt hier vanaf wanneer dit geen verdere nadelige gevolgen heeft. De TFT

speler die vals speelt krijgt hierdoor een incidenteel T-resultaat, zonder dat dit nadelige gevolgen heeft voor zijn toekomstige resultaten. Door dit vals spelen behaalt de opponerende TFT speler het S-resultaat maar omdat deze hier geen weet van heeft zal deze in ronde *r*+1 niet *retaliatory* maar gewoon coöperatief handelen. Binnen een herhaalde vorm van het MPD met een groot aantal spelers is asymmetrische informatie in het bijzonder relevant. De effecten van een incidentele afwijking van de coöperatieve handeling van een enkele speler binnen een grote groep spelers zullen over het algemeen immers minder goed voelbaar zijn dan wanneer één van de twee spelers binnen een IPD niet coöperatief handelt.

Omdat we er alleen van uitgaan dat er incidenteel sprake is van asymmetrische informatie maar verder wel uitgaan van een symmetrische situatie, heeft iedere speler over het geheel van alle ronden evenveel kansen om vals te spelen en heeft iedere speler evenveel kans om het slachtoffer te worden van vals spelen. Beide spelers hebben dus een gelijkwaardige kans om een aantal keer afwisselend het T-resultaat en het S-resultaat te behalen. Voor beide spelers geldt eveneens dat de kans op een T-resultaat even groot is als de kans op een S-resultaat. Omdat $R>(T+S)/2$ betekent dit dat iedere speler hierdoor slechter af is en dus niet het Pareto-efficiënte resultaat wordt behaald.

Het is goed om op te merken dat we hier de oorspronkelijke aannames van het Prisoner's Dilemma verlaten. De aanname dat er sprake is van volledige informatie wordt vervangen door de aanname dat er sprake is van asymmetrische informatie. TFT is een strategie die oorspronkelijk ontworpen is voor IPD's met volledige informatie. Het is in deze setting dat TFT een succesvolle strategie blijkt te zijn. Uit experimenten blijkt ook dat wanneer de aannames van het model veranderen TFT een minder succesvolle strategie is. In IPD's met foutkans bijvoorbeeld, waar spelers de handeling van hun opponent verkeerd kunnen percipiëren of anders handelen dan bedoeld, blijkt TFT een minder succesvolle strategie te zijn. Andere strategieën zoals 'generous TFT' waarin een enkele afwijking van de coöperatieve handeling niet meteen bestraft wordt blijken in een dergelijk model beter te werken.[37]

[37] Steven Kuhn, 'Prisoner's Dilemma,' in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*,
URL = <http://plato.stanford.edu/archives/fall2010/entries/prisoner-dilemma/>.

Onderzoek zal uit moeten wijzen of dit voor IPD's met asymmetrische informatie ook het geval is. Men zou evenwel kunnen argumenteren dat een TFT strategie waarin incidenteel vals wordt gespeeld omdat de situatie van asymmetrische informatie dit toelaat geen echte TFT strategie is. Daarentegen zou een rationele TFT speler nooit coöperatief handelen wanneer hij de zekerheid heeft dat zijn handeling onopgemerkt zal blijven door zijn opponent. Dit is immers in strijd met nutsmaximalisatie.

Hoe dan ook, het feit dat TFT geen Pareto-efficiënt realiseert in IPD's met asymmetrische informatie lijkt voor Gauthier een belangrijke tekortkoming te zijn van TFT. Dit blijkt indirect uit het volgende citaat.

"[...]"The altruism is the more efficiënt because it is *not* derived from calculated self-interest." This is exactly our point at the end of 2.1–constrained maximization is not straightforward maximization in its most effective guise. The constrained maximizer genuinely ignores the call of utility-maximization[…]. There is no simulation; if there were, the benefits of co-operation would not be fully realized."[38]

Gedeeltelijk citeert Gauthier hier Jon Elster[39] die het werk 'The Evolution of Reciprocal Altruism'[40] van Robert L. Trivers bediscussieert.

Ook volgens Trivers is het soort vals spelen dat zojuist besproken is, Trivers noemt dit 'subtle cheating', rationeel. Het loont om vals te spelen wanneer dit onopgemerkt blijft door de andere partij.[41] Maar, zo argumenteert Trivers, evolutie zal er voor zorgen dat de capaciteit van spelers om valsspelers en vals spelen te ontdekken beter ontwikkeld wordt.

"Selection should favor the ability to detect and discriminate against subtle cheaters. Selection will clearly favor detecting and countering sham moralistic aggression. The argument for the others is more complex. Selection may favor distrusting those who perform altruistic acts without the emotional basis of generosity or guilt because the altruistic tendencies of such individuals may be less reliable in the future. One can imagine, for example, compensating for a misdeed without any emotional basis but with a calculating, self-serving motive. Such an

[38] David Gauthier, *Morals by Agreement*. New York: Oxford University Press 1986, pp. 188-189.
[39] Jon Elster, *Ulysses and the Sirens: Studies in rationality and irrationality*, Revised edition, Cambridge: Cambridge University Press, 1993, (First published 1979), p. 145.
[40] Robert L. Trivers, 'The Evolution of Reciprocal Altruism,' in *The Quarterly Review of Biology*, Vol. 46, No. 1 (1971).
[41] Ibid., p. 48.

individual should be distrusted because the calculating spirit that leads this subtle cheater now to compensate may in the future lead him to cheat when circumstances seem more advantageous (because of unlikelihood of detection, for example, or because the cheated individual is unlikely to survive). Guilty motivation, in so far as it evidences a more enduring commitment to altruism, either because guilt teaches or because the cheater is unlikely not to feel the same guilt in the future, seems more reliable. A similar argument can be made about the trustworthiness of individuals who initiate altruistic acts out of a calculating rather than a generous hearted disposition or who show either false sympathy or false gratitude. Detection on the basis of the underlying psychological dynamics is only one form of detection. In many cases, unreliability may more easily be detected through experiencing the cheater's inconsistent behavior." [42]

Bij valsspelers is volgens Trivers sprake van een geveinsde moraliteit. Er is geen emotionele basis voor een coöperatieve dispositie, valsspelers worden niet geleid door gevoelens van generositeit of schuld. Bij valsspelers kunnen we ook wel spreken van een calculerend altruïsme dat alleen gericht is op het eigenbelang. Calculerende altruïsten simuleren een altruïstische houding en zullen proberen te profiteren van coöperatieve verbanden met oprechte altruïsten. Omdat calculerende altruïsten minder aantrekkelijke partners zijn dan oprechte altruïsten zal natuurlijke selectie er voor zorgen dat de capaciteit om valsspelers te herkennen ontwikkeld wordt. Het missen van een emotionele basis voor een coöperatieve dispositie en inconsistent gedrag zijn in deze belangrijke kenmerken om calculerend altruïsme te herkennen.

Oprechte altruïsten zijn aantrekkelijkere partners om coöperatieve verbanden mee aan te gaan. Oprecht altruïsme moet men in deze context overigens zien als een voorwaardelijk altruïsme. De term 'oprecht' duidt er op dat men niet vals speelt, niet dat men onvoorwaardelijk altruïstisch handelt. Gevoelens van generositeit en schuld zorgen er voor dat vals spelen voor de oprechte altruïst minder aantrekkelijk wordt. Wanneer de capaciteit om vals spelen te herkennen beter ontwikkeld is in een populatie zal calculerend altruïsme minder aantrekkelijk worden. Er is immers minder ruimte om vals te spelen en calculerende altruïsten zullen eerder uitgesloten worden van coöperatieve verbanden.

[42] Ibid., pp. 50-51.

In het citaat dat we eerder gaven zien we dat Gauthier een directe link legt tussen enerzijds oprecht altruïsme en CM. Net als een oprechte altruïst laat een CM-er zich niet leiden door nutsmaximalisatie en is er geen sprake van een simulatie van altruïstisch gedrag. Anderzijds legt Gauthier een directe link tussen calculerend altruïsme en SM. Calculerend altruïsme ziet Gauthier als een TFT strategie, als 'straightforward maximization in its most effective disguise'.

In het citaat zien we verder dat Gauthier zich aansluit bij de gedachte van Elster dat, oprecht altruïsme efficiënter is juist omdat het niet gebaseerd is op een op eigenbelang gericht calculerend altruïsme. Dit impliceert dat CM volgens Gauthier om dezelfde reden efficiënter is dan TFT. CM is een beter alternatief dan TFT omdat bij TFT de voordelen van coöperatief handelen, in het bijzonder in IPD's met asymmetrische informatie, niet volledig gerealiseerd worden. Spelers ontwikkelen immers de capaciteit om valsspelers te herkennen. De voordelen van vals spelen wegen uiteindelijk niet meer op tegen de nadelen, waardoor CM aantrekkelijker wordt. Waar TFT in IPD's met asymmetrische informatie uiteindelijk dus leidt tot een Pareto-inefficiënte uitkomst omdat rationele TFT spelers vals spelen wanneer ze de mogelijkheid hebben realiseert CM wel een Pareto-efficiënt resultaat omdat beide spelers niet de dupe worden van elkaars neiging tot vals spelen.

Het is belangrijk om nogmaals te benadrukken dat de noties van transparantie en RPP essentieel zijn. In Trivers bewoording betekent transparantie, het kunnen onderscheiden tussen voorwaardelijke oprechte altruïsten die niet geneigd zijn om vals te spelen en voorwaardelijke calculerende altruïsten die wel geneigd zijn om vals te spelen. Om het in Gauthier's bewoording te formuleren, het kunnen onderscheiden tussen CM-ers en SM-ers of SM-ers in 'most effective disguise' (TFT spelers). Beide komen op hetzelfde neer. Of CM een beter alternatief is dan TFT hangt af van de mate van transparantie binnen de populatie en het aantal CM-ers binnen de populatie. Alleen wanneer de capaciteit van spelers, om te kunnen onderscheiden tussen CM-ers en SM-ers of oprechte altruïsten en valspelers, goed genoeg ontwikkeld is en wanneer men er van uit mag gaan dat rationele spelers handelen conform hun dispositie, is CM een rationeel alternatief voor TFT.

Hoofdstuk 4

De houdbaarheid van het RPP en de transparantie aanname

Hoe houdbaar zijn de noties van RPP en transparantie eigenlijk? Deze vraag is cruciaal om een antwoord te kunnen geven op de vraag of CM een rechtvaardiging biedt voor coöperatief handelen.

4.1 De houdbaarheid van het RPP

Zoals we eerder al zagen impliceert RPP dat, wanneer het voor een speler rationeel is om een bepaalde intentie te vormen, het rationeel is om conform die intentie te handelen. In haar bespreking van Gauthier stelt Smith dit principe ter discussie. Volgens Smith gelden de redenen om een bepaalde intentie te vormen namelijk niet noodzakelijk ook als redenen om conform die intentie te handelen.

"[…]the agent's reasons for forming the intention to do A seem not to carry over *at all* as reasons to do A itself."[43]

In een PD kan het rationeel zijn voor speler x om een voorwaardelijke intentie tot coöperatief handelen te vormen in de hoop dat speler y hierdoor coöperatief zal handelen. De reden om de intentie te vormen is hier dus de hoop dat y coöperatief zal handelen. Dit is echter geen reden voor die speler om ook daadwerkelijk coöperatief te handelen. Het daadwerkelijke handelen

[43] Holly Smith, 'Deriving Morality from Rationality,' in Peter Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on David Gauhtier's Morals by Agreement* (Cambridge: Cambridge University Press, 1991), p. 246.

van x heeft immers geen enkele invloed op het handelen van y. Integendeel, x zal juist reden hebben om non-coöperatief te handelen aangezien dit ongeacht de handeling van de andere speler het hoogste resultaat oplevert. RPP schrijft echter voor dat het voor x rationeel is om coöperatief te handelen aangezien het rationeel was om die intentie te vormen.

RPP heeft volgens Smith onlogische implicaties en geeft ter illustratie het volgende voorbeeld. Stel er is een inbreker met telepathische krachten en deze inbreker dreigt alle waardevolle bezittingen van een huiseigenaar te stelen. Echter, de huiseigenaar heeft toevallig een lading gebruiksklare explosieven in zijn nachtkastje liggen. De huiseigenaar kan er bijna zeker van zijn dat wanneer hij de intentie vormt om het huis op te blazen (inclusief alle bezittingen, de inbreker en zichzelf) als de inbreker niet afziet van zijn plannen, dit de inbreker af zal schrikken en deze op de vlucht zal doen slaan. De huiseigenaar vormt de intentie, maar helaas, de inbreker wordt niet afgeschrokken. Volgens RPP is het nu rationeel voor de huiseigenaar om zichzelf op te blazen omdat er een rationele motivatie was om die intentie te vormen.

Smith concludeert dan ook dat RPP niet plausibel is. Ik denk echter dat Smith hier geen recht doet aan Gauthier's idee van disposities. Gauthier ziet een dispositie als een toestand die noodzakelijk gerealiseerd wordt als aan een bepaalde voorwaarde wordt voldaan. Dit wordt ook wel de 'Simple Condition Analysis' (SCA)[44] genoemd. SCA impliceert ook dat, wanneer er aan de betreffende voorwaarde wordt voldaan en die toestand niet gerealiseerd wordt, er geen sprake is van een dispositie.

CM betekent dat men coöperatief handelt indien de andere speler ook een dispositie tot coöperatief handelen heeft. Het betekent ook dat wanneer speler y een dispositie tot coöperatief handelen heeft en speler x niet coöperatief handelt, x geen CM-er is en geen dispositie tot coöperatief handelen heeft. In het geval dat een speler *ceteris paribus* afwijkt van zijn 'dispositie' betekent dit dat er eigenlijk geen sprake is van een echte dispositie maar van een soort schijndispositie. Een dispositie die er op gericht is om de andere speler te

[44] Michael Fara, 'Dispositions' in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*,
URL = <http://plato.stanford.edu/archives/fall2010/entries/dispositions/>.

misleiden. Gauthier neemt echter aan dat er transparantie heerst en dit betekent dat de spelers elkaars ware dispositie kennen. Blijk geven van een bepaalde dispositie en tegelijkertijd de intentie hebben om daar later van af te wijken is onmogelijk. Spelers zouden elkaar doorzien omdat er transparantie heerst. Het hebben van een dispositie tot coöperatief handelen betekent dat men ook daadwerkelijk de intentie heeft om coöperatief te handelen als bij de andere speler ook een dispositie tot coöperatief handelen waargenomen wordt.

Wat betreft het voorbeeld van de inbreker kunnen we het volgende zeggen. RPP impliceert eveneens dat, wanneer het niet rationeel is om conform een intentie te handelen het tevens niet rationeel is om die intentie in eerste instantie te vormen. Als het voor de huiseigenaar niet rationeel is om zichzelf op te blazen is het tevens niet rationeel om die intentie te vormen. Het zou voor de huiseigenaar rationeel zijn om de inbreker te laten denken dat hij zichzelf op zal blazen als de inbreker niet afziet van zijn plannen hetgeen de inbreker mogelijk af zal schrikken. Tegelijkertijd is het niet rationeel voor de huiseigenaar om ook daadwerkelijk de intentie te hebben om zichzelf op te blazen. Maar omdat de inbreker telepathisch is weet deze, net als in een situatie waar transparantie heerst, dat de huiseigenaar niet werkelijk de intentie heeft om zichzelf op te blazen en zal hierdoor dan ook niet afgeschrikt worden. De huiseigenaar wordt beroofd maar blijft desalniettemin levend achter.

Deze scriptie is er verder niet op gericht om te diep in te gaan op de vraag of RPP houdbaar is of niet. We concluderen slechts dat de kritiek van Smith in ieder geval geen stand houdt en we zullen RPP verder als gegeven beschouwen.

4.2 De houdbaarheid van de transparantie aanname

Volgens Gauthier spreken we van transparantie wanneer spelers binnen een populatie volledige kennis hebben van elkaars disposities. In een speltheoretische situatie waarin sprake is van transparantie kennen spelers de ware intenties van hun opponenten en weten spelers welke strategieën hun

opponenten volgen. Gauthier erkent dat transparantie een theoretisch concept is en beweert dat spelers in realiteit slechts quasi-transparant zijn. Spelers zijn dus slechts in staat om elkaars dispositie met een beperkte mate van zekerheid in te schatten. Verder is het vermogen om de dispositie van de ander redelijkerwijs in te kunnen schatten een capaciteit die ontwikkeld en getraind kan worden en meer dan slechts gokwerk.

De vraag is nu echter hoe men precies tot die kennis van de ander zijn dispositie komt. Gelet op het feit dat transparantie een erg belangrijke rol speelt is het opvallend dat Gauthier verder weinig aandacht besteed aan deze vraag. We zijn daarom ook genoodzaakt hier zelf een antwoord op te vinden. Mijns inziens zijn er twee mogelijke antwoorden denkbaar.

1) Spelers hebben kennis van de historie van elkaars gedrag en schrijven op basis daarvan een dispositie aan elkaar toe. Trivers geeft hier twee voorbeelden van. Spelers kunnen bijvoorbeeld door herhaalde interactie iets te weten komen over de emotionele configuratie van hun opponent. Men zou te weten kunnen komen of anderen in het handelen voornamelijk worden gedreven door gevoelens van generositeit en schuld of dat ze een meer calculerende aard hebben. Hieruit kan men afleiden of een opponent een oprechte dispositie tot coöperatief handelen heeft of dat deze de neiging heeft om vals te spelen indien mogelijk. Er is daarnaast uiteraard ook veel af te leiden uit het historisch spelgedrag van spelers. Als een speler simpelweg vaak non-coöperatief gedrag heeft laten zien bij opponenten die vaak coöperatief gehandeld hebben is het aannemelijk dat diegene een SM-er is. Spelers kunnen ook inconsistenties in elkaars gedrag waarnemen. Een speler die calculerend van aard is en probeert over te komen als een oprechte altruïst zal eerder inconsistenties in zijn gedrag laten zien dan iemand die oprecht altruïstisch is.

Om na te gaan of dit de soort van kennis is waar Gauthier op doelt is het belangrijk om te bedenken dat Gauthier benadrukt dat CM-ers voor de coöperatieve handeling kunnen kiezen zelfs als dit geen enkel effect heeft op toekomstige situaties.

"Constrained maximizers may co-operate even if neither expects her choice to affect future situations."[45]

Wanneer spelers alleen kennis hebben van de historie van elkaars dispositie is het moeilijk voor te stellen hoe de coöperatieve handeling ooit rationeel kan zijn in het geval dat deze geen enkel effect heeft op toekomstige situaties. Ongeacht het feit of een speler herkend wordt als een CM-er of SM-er is het in dat geval namelijk altijd rationeel om een non-coöperatieve dispositie in te nemen en non-coöperatief te handelen. Een opponent baseert zijn handeling immers op de historische dispositie van de speler. De dispositie die de speler inneemt met betrekking tot de eerstvolgende speelronde heeft geen enkel effect op de handeling in de eerstvolgende of toekomstige speelronden. De handeling van zijn opponent staat als het ware al vast en in dat geval is non-coöperatief handelen in een Prisoner's Dilemma de dominante strategie.

Een bijkomend probleem is dat wanneer spelers totaal onbekend zijn met elkaars historie er geen enkele basis is om coöperatief te handelen. In ieder geval, niet als we veronderstellen dat de handeling geen enkel effect heeft op toekomstige situaties. Het is überhaupt moeilijk voor te stellen hoe samenwerking in een dergelijke situatie tot stand zou kunnen komen. We moeten dan ook concluderen dat CM niet op deze notie van transparantie gebaseerd kan zijn.

2) Een ander mogelijk antwoord op de vraag hoe spelers tot kennis kunnen komen van elkaars dispositie is dat spelers directe toegang hebben tot elkaars mentale staat. Spelers hebben op dezelfde manier toegang tot elkaars mentale staat als zij zelf toegang hebben tot hun eigen mentale staat. De telepathische inbreker van Smith zou hier een voorbeeld van kunnen zijn. Spelers hebben een soort mentale link waardoor ze direct op de hoogte zijn van elkaars dispositie. We kunnen dit ook zien als een soort zesde zintuig waarmee we heel snel kunnen beoordelen of aanvoelen of iemand een coöperatieve of een non-coöperatieve dispositie heeft.

Dit is het idee van transparantie dat Gauthier lijkt te hebben. Als spelers directe toegang hebben tot elkaars mentale staat en directe kennis hebben van

[45] David Gauthier, *Morals by Agreement*, New York: Oxford University Press 1986, pp. 169-170: voetnoot 19.

elkaars disposities kan het wel rationeel zijn om te kiezen voor de coöperatieve handeling, zelfs als deze handeling geen enkel effect heeft op toekomstige situaties. De opponent baseert zijn dispositie dan namelijk op de dispositie die de speler heeft met betrekking tot de eerstvolgende speelronde. In dit geval is het rationeel voor een speler om een coöperatieve dispositie in te nemen waardoor het voor zijn opponent ook rationeel wordt om een coöperatieve dispositie in te nemen. Beide spelers kunnen bij deze vorm van transparantie een Pareto-efficiënt resultaat behalen zelfs als hun handelingen geen enkel effect hebben op toekomstige situaties.

Celeste M. Friend omschrijft Gauthier's notie van transparantie treffend als het idee dat spelers 'ramen' hebben. Spelers kunnen als het ware bij elkaar naar binnen kijken en elkaars motieven en intenties op een heel directe manier waarnemen. Friend verwerpt het idee dat spelers een dergelijke directe toegang hebben tot elkaars intenties.

"Gauthier seems to have in mind a picture of us in which we simply come equipped with windows (however imperfect or filmy they may be) into our inner lives, which simply reveal the contents of our intentions and motivations. I intend to show that this understanding of translucency, on which Gauthier's argument very much relies, is in fact quite wrong."[46]

Haar argument hiervoor verloopt als volgt. CM is gebaseerd op het idee dat transparantie of quasi-transparantie spelers reden geeft om elkaar te vertrouwen. Ze kunnen daardoor onderscheiden tussen betrouwbare CM-ers en onbetrouwbare SM-ers. Gauthier gaat er echter van uit dat transparantie enkel een voorwaarde is voor vertrouwen. Wat Gauthier echter zou vergeten volgens Friend is dat vertrouwen net zo goed een voorwaarde is voor transparantie. De relatie is dus niet zo unilateraal als Gauthier veronderstelt. Vertrouwen en transparantie ontstaan door herhaalde interactie en zijn de producten van sociale relaties. Transparantie is niet simpelweg een achtergrondvoorwaarde voor vertrouwen en coöperatief gedrag.

Ik denk dat Friend gelijk heeft wanneer ze er op wijst dat transparantie een product is van sociale interactie en het idee dat Gauthier heeft van transparantie verwerpt. Het is eenvoudigweg moeilijk voor te stellen hoe we

[46] Celeste M. Friend, 'Trust and the Presumption of Translucency,' in *Social Theory and Practice*, Vol. 27, Iss. 1 (2001), p9.

als het ware naar binnen kunnen kijken bij anderen en zoals we kennis kunnen nemen van onze eigen intenties ook kennis kunnen nemen van de intenties van anderen. Als er zoiets bestaat als transparantie zal dat eerder lijken op de eerste vorm die we besproken hebben. Door elkaars spelgedrag en emotionele configuratie te analyseren op basis van historisch gedrag, kunnen spelers elkaars strategieën trachten te achterhalen. Men kan te weten komen of een opponent oprecht of calculerend van aard is, of een opponent oprecht coöperatief handelt of vals speelt wanneer de kans zich voor doet. Wanneer spelers meer weten over elkaars strategieën zijn ze beter in staat een inschatting te maken van hoe de ander in de toekomst zal handelen en of er een rationele basis is voor coöperatief handelen. Wanneer de handeling echter geen enkel effect heeft op toekomstige situaties is er geen enkele rationele basis voor coöperatief handelen.

Hoofdstuk 5

De toegevoegde waarde van CM

Tot zover hebben we geconcludeerd dat er binnen conventionele speltheorie, voor bepaalde vormen van het Prisoner's Dilemma, een rationele rechtvaardiging is te vinden voor de coöperatieve handeling. Dit betekent dat wanneer spelers rationeel handelen ze tot een Pareto-efficiënt resultaat kunnen komen. Voor andere vormen zoals een PD, een MPD of een IPD met een lage waarde voor *p* of met asymmetrische informatie gaat dit niet op. Non-coöperatief handelen of vals spelen is in deze gevallen rationeel wat er toe leidt dat rationele spelers tot een Pareto-inefficiënt resultaat komen.

Volgens Gauthier biedt CM een beter alternatief omdat deze strategie ook in de laatst genoemde versies van het Prisoner's Dilemma rationeel is en een Pareto-efficiënte uitkomst realiseert. Een voorwaarde hiervoor is echter dat er een voldoende mate van transparantie heerst binnen de populatie en spelers handelen conform hun dispositie. Wanneer aan die voorwaarden is voldaan is CM superieur ten opzichte van SM, aldus Gauthier. Maar RPP is niet onomstreden en de vorm van transparantie die Gauthier veronderstelt is onhoudbaar wat CM tot een problematisch concept maakt.

Maar stel nu dat we RPP en transparantie wel aan zouden nemen. We moeten ons dan nog steeds afvragen of het inderdaad zo is dat CM een betere rechtvaardiging biedt voor coöperatief handelen in een Prisoner's Dilemma dan conventionele speltheorie. Om antwoord te geven op deze vraag moeten we ons eerst afvragen wat beide aannames eigenlijk in essentie betekenen voor een speltheoretische situatie.

RPP impliceert dat het voor spelers *ceteris paribus* rationeel is om te handelen conform hun intenties. Transparantie impliceert dat beide spelers volkomen op de hoogte zijn van elkaars intenties. Bij aanname van RPP en transparantie is het dus zo dat een speler de zekerheid heeft dat wanneer zijn

opponent rationeel is deze zal handelen conform zijn intenties. Beide spelers hebben dus als het ware een wederzijdse toezegging over hoe ze zullen handelen. Een toezegging waarvan ze zeker mogen zijn dat die nageleefd wordt.

De intenties van beide spelers zijn verder onderling afhankelijk van elkaar. Speler x handelt alleen coöperatief als speler y ook de intentie heeft om coöperatief te handelen. Voor y geldt hetzelfde. Als y niet de intentie heeft om coöperatief te handelen is het voor x ook niet rationeel om die intentie te hebben. Doordat de intenties van de spelers onderling afhankelijk zijn van elkaar kunnen spelers elkaar er toe bewegen tot een gedeelde intentie te komen. Het vormen van intenties werkt dus als een strategisch onderhandelingsproces. Nota bene, een strategisch onderhandelingsproces waarbij bindende afspraken gemaakt kunnen worden! Spelers hebben immers de zekerheid dat hun opponent handelt conform zijn intenties.

De aannames van transparantie en RPP hebben dus dezelfde speltheoretische implicaties als het veronderstellen dat spelers kunnen onderhandelen met elkaar en bindende afspraken kunnen maken. In een PD waarin bindende afspraken kunnen worden gemaakt zullen twee rationele spelers altijd komen tot een Pareto-efficiënt resultaat. Coöperatief handelen is in dit geval de rationele uitkomst van het onderhandelingsproces. Is CM onder die aannames dan daadwerkelijk anders dan SM? Niet als we een SM-er begrijpen als een speler die simpelweg zijn nut probeert te maximaliseren. Een speler die simpelweg zijn nut maximaliseert zal in een Prisoner's Dilemma waarin bindende afspraken kunnen worden gemaakt immers ook gewoon kiezen voor de coöperatieve handeling. Wanneer we volledige transparantie en RPP aannemen is er binnen conventionele speltheorie dus ook een rechtvaardiging te vinden voor de coöperatieve handeling in een Prisoner's Dilemma. Ook voor vormen van het Prisoner's Dilemma die slechts uit één speelronde bestaan en herhaalde versies met een lage *p* of asymmetrische informatie. De coöperatieve handeling is immers altijd rationeel wanneer er bindende afspraken gemaakt kunnen worden.

Wanneer er sprake is van quasi-transparantie, geldt, net als voor CM, dat de rationaliteit van voorwaardelijk coöperatief handelen afhankelijk is van de mate van transparantie en het aantal spelers in de populatie dat een

voorwaardelijke coöperatieve strategie volgt. Minder transparantie betekent in speltheoretische termen dat het moeilijker is om te onderhandelen of dat er een grotere kans is dat een opponent afwijkt van de afspraak. Het volgen van een voorwaardelijke coöperatieve strategie wordt hierdoor minder aantrekkelijk.

We moeten dan ook vaststellen dat: 1) als we RPP en de problematische transparantie aanname verwerpen CM geen rationele rechtvaardiging biedt voor de morele coöperatieve handeling in een Prisoner's Dilemma en 2) wanneer we de transparantie en RPP aanname niet verwerpen conventionele speltheorie evengoed een oplossing biedt voor de morele coöperatieve handeling. CM biedt daarom geen enkele toegevoegde waarde aan conventionele speltheorie.

Conclusie

In de introductie hebben we gezien dat in een situatie die de vorm heeft van een Prisoner's Dilemma moreel handelen en rationeel handelen met elkaar in strijd kunnen zijn. Een sociaal contract tussen twee gelijkwaardige rationele spelers in een Prisoner's Dilemma zal bestaan uit een afspraak dat beide spelers coöperatief zullen handelen. Vanuit een contractualistisch perspectief is coöperatief handelen dus moreel imperatief. Rationaliteit schrijft echter voor dat spelers afwijken van de afspraak omdat voor elke speler individueel geldt dat afwijken van de afspraak in overeenstemming is met individuele nutsmaximalisatie.

In deze scriptie hebben we geprobeerd antwoord te geven op de volgende vragen: 1) Is er binnen conventionele speltheorie een rationele rechtvaardiging te vinden voor moreel handelen in een Prisoner's Dilemma? 2) Als dit het geval is, waarom zoekt Gauthier dan naar een alternatieve rechtvaardiging voor moreel handelen in een Prisoner's Dilemma? 3) Is Gauthier's rechtvaardiging voor moreel handelen in een Prisoner's Dilemma houdbaar?

We hebben gezien dat er binnen conventionele speltheorie een rationele basis te vinden is voor moreel handelen binnen verschillende vormen van het Prisoner's Dilemma. Van de strategie TFT is gebleken dat het een succesvolle strategie is in IPD's en EPD's en dus een rationele basis biedt voor coöperatief handelen. Toch pleit Gauthier voor een alternatieve rechtvaardiging van moreel handelen. De voorwaardelijke coöperatieve strategie CM biedt volgens Gauthier een rationele rechtvaardiging voor moreel handelen in het Prisoner's Dilemma. Gegeven dat er een voldoende mate van transparantie heerst en we het RPP aannemen.

De conclusies die Axelrod en Gauthier trekken met betrekking tot respectievelijk TFT en CM zijn echter in hoge mate vergelijkbaar. Dit doet de vraag rijzen of TFT en CM wellicht vergelijkbare strategieën zijn. Gauthier is

echter zeer duidelijk dat beide strategieën niet met elkaar verward mogen worden. TFT is een verkapte vorm van SM en mag dus zeker niet worden vergeleken met CM. CM is volgens Gauthier een beter alternatief dan TFT omdat bij TFT de voordelen van coöperatief handelen, in het bijzonder in IPD's met asymmetrische informatie, niet volledig gerealiseerd worden. Waar TFT in IPD's met asymmetrische informatie uiteindelijk dus leidt tot een Pareto-inefficiënte uitkomst omdat rationele TFT spelers vals spelen wanneer ze de mogelijkheid hebben realiseert CM wel een Pareto-efficiënt resultaat omdat beide spelers niet de dupe worden van elkaars neiging tot vals spelen. CM is efficiënter juist omdat het niet gebaseerd is op een op eigenbelang gericht calculerend altruïsme.

Wat CM echter problematisch maakt is dat het de niet onomstreden notie van RPP en een onhoudbare notie van transparantie veronderstelt. De kern van het transparantieprobleem is dat het historisch gedrag van een speler de enige mogelijkheidsvoorwaarde voor kennis over de dispositie of intenties van die speler lijkt te zijn. Het is eenvoudigweg moeilijk voor te stellen hoe we directe toegang kunnen hebben tot de mentale staat van de ander. Maar als transparantie gebaseerd is op historisch gedrag kan coöperatief handelen nooit rationeel zijn wanneer de handeling geen enkel effect heeft op toekomstige situaties. De dispositie die de speler inneemt met betrekking tot de eerstvolgende speelronde heeft dan immers geen enkel effect op de handeling in de eerstvolgende of toekomstige speelronden. De handeling van zijn opponent staat als het ware al vast en in dat geval is non-coöperatief handelen in een Prisoner's Dilemma de rationele strategie.

Maar zelfs als we RPP en transparantie wel aannemen biedt CM nog geen toegevoegde waarde ten opzichte van conventionele speltheorie. Beide aannames hebben namelijk dezelfde speltheoretische implicaties als het veronderstellen dat spelers kunnen onderhandelen met elkaar en bindende afspraken kunnen maken.

In een PD waarin bindende afspraken kunnen worden gemaakt zullen twee rationele spelers, ook volgens conventionele speltheorie, altijd komen tot een Pareto-efficiënt resultaat. De term 'dilemma' is in dit geval eigenlijk niet eens meer van toepassing. We concluderen dan ook dat CM geen rationele rechtvaardiging biedt voor moreel handelen in een Prisoner's Dilemma.

Nawoord

CM biedt geen rechtvaardiging voor de morele handeling in een Prisoner's Dilemma. Maar is er binnen conventionele speltheorie dan een rechtvaardiging te vinden voor de morele handeling in een Prisoner's Dilemma? We zagen dat voor een aantal vormen van het Prisoner's Dilemma TFT een rationele rechtvaardiging biedt voor coöperatief moreel handelen. Er zijn echter ook een aantal vormen waarin conventionele speltheorie geen rechtvaardiging biedt voor moreel handelen. In enkelvoudige vormen van het Prisoner's Dilemma zoals het PD en het MPD of in IPD's met een lage *p* of asymmetrische informatie.

Hierbij moet aangetekend worden dat niet alle vormen van het Prisoner's Dilemma evenveel praktische relevantie hebben. Een puur enkelvoudig PD of MPD is in de praktijk moeilijker te vinden dan herhaalde versies. Meestal heeft ons handelen effect op toekomstige situaties. De keuzes die we maken hebben vaak consequenties voor onze reputatie of het vertrouwen dat anderen in ons hebben en dus voor onze mogelijkheden om toekomstige coöperatieve voordelen te realiseren. Axelrod laat verder zien dat een coöperatieve TFT strategie al succesvol kan zijn bij relatief lage *p* waarden.[47]

Dit betekent uiteraard niet dat er in de praktijk altijd een rationele rechtvaardiging is te vinden voor moreel handelen in een situatie die de vorm heeft van een Prisoner's Dilemma. Er zijn tal van situaties te bedenken waarin TFT strategieën bijvoorbeeld minder goed werken. Neem bijvoorbeeld situaties waarin individuen minder goed in staat zijn om elkaars intenties in te schatten. Als men weinig kennis heeft over elkaars gedragshistorie kunnen coöperatieve strategieën minder succesvol zijn. Het is voor spelers moeilijker

[47] Robert Axelrod, 'The Emergence of Cooperation among Egoists,' in *The American Political Science Review*, Vol. 75, No. 2 (1981), pp. 316.

in te schatten welke strategie de ander volgt. We zouden kunnen zeggen dat een strategie niet *clear* is, hetgeen voorwaarde is voor een succesvolle TFT strategie. Ook in gevallen waarin meerdere individuen interacteren en waar sprake is van asymmetrische of onvolledige informatie zal TFT minder succesvol zijn. Zulke gevallen bieden in mindere mate een basis voor reciprociteit en dus zijn voorwaardelijke coöperatieve strategieën minder succesvol.

We moeten dan ook vaststellen er in theorie en praktijk niet altijd een rationele rechtvaardiging is te geven voor de morele handeling. Maar is dit nu niet juist de reden waarom er een onderscheid gemaakt wordt tussen moraliteit en rationaliteit? Ik sluit me aan bij de passage waar *Morals by Agreement* mee begint, Gauthier reflecteert hier op een quote van David Hume.

> "What theory of morals can ever serve any useful purpose, unless it can show that all the duties it recommends are also the true interest of each individual? David Hume, who asked this question, seems mistaken; such a theory would be too useful. Were duty no more than interest, morals would be superfluous. Why appeal to right or wrong, to good or evil, to obligation or to duty, if instead we may appeal to desire or aversion, to benefit or cost, to interest or to advantage?" [48]

Waarom zou men zich beroepen op morele principes als men zich ook kan beroepen op nutsmaximalisatie? Gauthier heeft gelijk wanneer hij zegt dat moreel handelen meer is dan simpwelweg SM. Als de morele handeling niets anders zou zijn dan die handeling die uiteindelijk het hoogste nut realiseert dan wordt ethiek een oppervlakkige bezigheid. Men kan zich dan beter bezig gaan houden met de wetenschap die nutsmaximalisatie bestudeert. Waar ik me tegen afzet is Gauthier's idee dat CM wel een rationele rechtvaardiging biedt voor moreel handelen in een Prisoner's Dilemma. Gegeven de contractualistische benadering van de morele handeling en de nutsmaximaliserende conceptie van praktische rationaliteit, kan ik het niet anders zien dan dat moreel handelen en rationeel handelen eenvoudigweg niet altijd samengaan.

[48] David Gauthier, *Morals by Agreement*, New York: Oxford University Press 1986, p. 1.

Literatuur

- Axelrod, Robert. 'The Emergence of Cooperation among Egoists,' in *The American Political Science Review*, Vol. 75, No. 2 (1981), pp. 306-318.
- Axelrod, Robert. *The Evolution of Cooperation*. New York: Basic Books, 1984.
- Elster, Jon. *Ulysses and the Sirens: Studies in rationality and irrationality*. Revised edition. Cambridge: Cambridge University Press, 1993. (First published 1979)
- Fara, Michael. 'Dispositions' in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*, URL = <http://plato.stanford.edu/archives/fall2010/entries/dispositions/>.
- Friend, Celeste M. 'Trust and the Presumption of Translucency,' in *Social Theory and Practice*, Vol. 27, Iss. 1 (2001), pp. 1-18.
- Gauthier, David. *Morals by Agreement*. New York: Oxford University Press 1986.
- Kuhn, Steven. 'Prisoner's Dilemma,' in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*, URL = <http://plato.stanford.edu/archives/fall2010/entries/prisoner-dilemma/>.
- Smith, Holly. 'Deriving Morality from Rationality,' in Peter Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on David Gauhtier's Morals by Agreement* (Cambridge: Cambridge University Press, 1991), pp. 229-253.
- Trivers, Robert L. 'The Evolution of Reciprocal Altruism,' in *The Quarterly Review of Biology*, Vol. 46, No. 1 (1971), pp. 35-57.
- Verbeek, Bruno and Christopher Morris. 'Game theory and Ethics,' in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*, URL = <http://plato.stanford.edu/archives/fall2010/entries/game-ethics/>.