UNIVERSITEIT VAN UTRECHT

Een Speltheoretisch Model van de Oorspronkelijke Positie

door

Rijk Mercuur

Een scriptie van 7,5 EC voor het behalen van
de titel Bachelor of Science

in de
Geesteswetenschappen
Wijsbegeerte & CKI

Eerste begeleider: prof. dr. Henry Prakken
Tweede begeleider: prof. dr. Martin van Hees

20 februari 2013

Inhoudsopgave

Hoofdstuk 1

Inleiding

Een samenleving wordt evenzeer gekenmerkt door een belangenconflict als een belangenovereenkomst. Een overeenkomst omdat samenwerking een beter leven mogelijk maakt dan ieder op zichzelf kan garanderen; een conflict omdat personen niet onverschillig staan tegenover de verdeling van de opbrengst van hun samenwerking. Dit is in een notendop het probleem dat John Rawls in *A Theory of Justice* [8] behandelt. Rawls stelt twee *principes van rechtvaardigheid* voor die de distributie van goederen volgens hem moeten reguleren. De reden dat Rawls deze principes verkiest boven andere hedendaagse concepties van rechtvaardigheid (bijvoorbeeld utilitarisme) is omdat, zo claimt Rawls, deze principes worden gekozen in een hypothetische initiële keuzesituatie: de *oorspronkelijke positie*.

In deze scriptie bekijk ik de mogelijkheid tot een *speltheoretisch modellering* van de oorspronkelijke positie. Dit roept de vragen op (1) wat Rawls' oorspronkelijke positie inhoudt, (2) wat een adequaat model van deze oorspronkelijke positie zou zijn en (3) wat speltheorie op dit vlak kan bieden. Mijn doel is om deze vragen te beantwoorden en zo uiteindelijk tot een adequaat speltheoretisch 'oorspronkelijke positie'-model te komen.

Ik wil dit doen door in hoofdstuk 2 aan de hand van [8], [3] en [10] Rawls' oorspronkelijke positie te beschrijven. Ik blijf dicht bij Rawls' filosofische taalgebruik maar eis, met in het achterhoofd een modellering, vereenvoudiging en precisie. In hoofdstuk 3 beschrijf ik wat, naar mijn mening, de eisen zijn waar een speltheoretisch model van de oorspronkelijke positie aan zou moeten voldoen. In hoofdstuk 4 onderzoek ik de mogelijkheden van een coöperatief-speltheoretische benadering aan de hand van Anthony Ladens 'Games, Fairness and Rawls's A Theory of Justice' [4]. In hoofdstuk 5 geef ik een niet-coöperatief-speltheoretisch model van de oorspronkelijke positie, gebaseerd op een formeel raamwerk van Pattanaik en Suzumaru [6], waarvan ik claim dat deze voldoet aan de eisen gesteld in hoofdstuk 3.

Ik schrijf deze scriptie ter voltooiing van mijn bachelor in de kunstmatige intelligentie. Kunstmatige intelligentie houdt zich onder andere bezig met het formeel modelleren van menselijke intelligentie en interactie. De relevantie van mijn scriptie voor kunstmatige intelligentie ligt in een precieze en formele benadering van een concept dat hier onlosmakelijk mee verbonden is: rechtvaardigheid.

Hoofdstuk 2

Rawls' oorspronkelijke positie

2.1 Rechtvaardigheid als eerlijkheid

John Rawls beschrijft in *A Theory of Justice* zijn *conceptie van rechtvaardigheid* die hij *rechtvaardigheid als eerlijkheid* ('justice as fairness') noemt [8]. Deze conceptie wordt gedefinieerd door Rawls' *principes van rechtvaardigheid*. Deze principes moeten volgens Rawls de basisstructuur van de samenleving reguleren, dat wil zeggen, de manier waarop de hoofdinstituties van de samenleving rechten en plichten toekennen en de voordelen van sociale samenwerking verdelen. De principes van rechtvaardigheid als eerlijkheid zijn (vrij vertaald van [9, p. 43-44]):

1. Eerste principe: Elk persoon dient een gelijk recht te hebben op het meest uitgebreide stelsel van gelijke fundamentele vrijheden, dat verenigbaar is met een vergelijkbaar stelsel van vrijheden voor anderen.
2. Tweede principe: Sociale en economische ongelijkheden dienen zo te worden geordend dat ze zowel
   (a) verbonden zijn aan ambten en posities die voor allen toegankelijk zijn *(principe van gelijkheid van mogelijkheden)*, en
   (b) het meest ten goede komen aan de minst bevoordeelden *(het verschilprincipe)*.

De principes zijn in een seriële rangorde geplaatst: het eerste principe heeft voorrang op het tweede principe en het principe van gelijkheid van mogelijkheden op het verschilprincipe.[1] Het eerste principe garandeert voor elk persoon fundamentele vrijheden. Op deze vrijheden kan, door de genoemde rangorde, geen inbreuk worden gemaakt door een beroep op het tweede principe. Het tweede principe is het principe dat stelt dat sociale en economische ongelijkheden worden geordend zodat deze voldoen aan (a) het principe van gelijkheid van mogelijkheden en, waarna hieraan is voldaan, (b) het verschilprincipe. Het principe van gelijkheid van mogelijkheid valt buiten de vereenvoudiging van deze scriptie. Echter, het verschilprincipe vormt een belangrijk onderdeel van Rawls' theorie van rechtvaardigheid. Het stelt dat we de voordelen van

[1] Zie [8, p. 38] en paragraaf 2.4 voor toelichting op deze seriële rangorde.

sociale samenwerking verdelen volgens een maximin-verdeling. In een maximin-verdeling ordenen we alle mogelijke verdelingen op hun minimale component en kiezen de verdeling waar het minimum het hoogst is.

De hoofdreden dat Rawls' zijn principes van rechtvaardigheid boven andere hedendaagse concepties van rechtvaardigheid verkiest is dat hij van mening is dat deze principes zouden worden gekozen door rationele actoren in een hypothetische initiële situatie genaamd de *oorspronkelijke positie* [8, h. 3]. De oorspronkelijke positie is een gedachtenexperiment; het is ontworpen om de restricties die rechtvaardigheid plaatst op argumentatie te verduidelijken. Rawls vindt dat men, wanneer men spreekt over een rechtvaardige verdeling van goederen, alleen argumenten mag aandragen die *moreel* relevant zijn. De beschrijving van de oorspronkelijke positie belichaamt deze beperking en geeft de gekozen principes daarmee een morele basis. Eenmaal binnen dit raamwerk kan puur rationeel (zonder verdere morele redenen) worden bepaald welke principes worden gekozen. Het doel is om deze initiële situatie zo te karakteriseren dat de gekozen principes acceptabel zijn vanuit een moreel oogpunt, wat ze ook blijken te zijn. Deze scriptie legt de focus op deze rationele keuze en niet op de *verantwoording* van de restricties in de oorspronkelijke positie. We vervolgen onze weg door eerst, door een beschrijving van de oorspronkelijke positie, deze morele restricties te beschrijven. In paragraaf 2.3 zullen we de overgebleven basis voor de keuze van beschrijven. We sluiten dit hoofdstuk af met de keuze van de actoren in de oorspronkelijke positie voor Rawls' principes van rechtvaardigheid.

2.2 Een beschrijving van de oorspronkelijke positie

De oorspronkelijke positie is een keuzesituatie waarin vrije en gelijke actoren tot een besluit komen over de principes van rechtvaardigheid. Deze actoren, door Rawls partijen genoemd, dragen de *sluier van onwetendheid* ('veil of ignorance'): deze sluier ontneemt de partijen alle kennis over zichzelf en beperkt de basis voor hun keuze tot algemene overwegingen [8, par. 24]. In het bijzonder zijn de partijen niet bekend met hun sociaal-economische positie, karakter en *conceptie van het goede*, maar wel met algemene feiten over de maatschappij en de menselijke psyche. De sluier dient om de moreel irrelevante toevalligheden (die mensen verleiden hun sociale en natuurlijke voordelen te misbruiken) te niet te doen en zo een *eerlijke* ('fair') keuzesituatie te creëren. (Vandaar de naam 'rechtvaardigheid als eerlijkheid'.)

De rol van de partijen kan het best worden gezien als die van individuen die hun eigen belangen behartigen. De partijen zijn daarin *wederzijds ongeïnteresseerd rationeel*: de partijen kiezen die principes die hun *eigen* conceptie van het goede het best bevorderen.

De oorspronkelijke positie is dus een keuzesituatie waarin partijen onder de sluier van onwetendheid proberen die principes te kiezen die het best hun eigen conceptie van het goede bevorderen. De sluier biedt de morele component van de gekozen principes. De rationaliteit van de partijen biedt de rationele component.

2.3 De basis voor de keuze van de partijen

In deze paragraaf beantwoord ik de vraag hoe de partijen bepalen welke principes het best hun eigen conceptie van het goede bevorderen. Immers, Rawls veronderstelt dat de partijen door

de sluier van onwetendheid hun conceptie van het goede niet kennen. Ik volg Freeman [3] in de interpretatie van Rawls dat er drie 'hogere orde'-belangen zijn van de partijen, deze 'hogere orde'-belangen bepalen welke principes worden gekozen in de oorspronkelijke positie.

1. De partijen streven hun conceptie van het goede na, ook al weten ze niet wat deze conceptie inhoudt [3, par. 4].

Omdat ze niet weten wat hun conceptie van het goede inhoudt bestaat dit 'hogere order'-belang, naar mijn interpretatie, uit twee delen:

a Het nastreven van condities in de maatschappij die (het nastreven van) diverse concepties van het goede mogelijk maken;

b Het nastreven van condities die in dienst staan van *ieders* conceptie van het goede.

In dienst van dit eerste 'hogere orde'-belang staan twee 'hogere orde'-belangen die bekend staan als de twee *morele vermogens*. Volgens Freeman definieert Rawls deze morele vermogens als "the features of human beings in virtue of which they are to be treated in accordance with the principles of justice" en onderscheidt hij twee soorten:[2]

2. Het eerste morele vermogen dat de partijen nastreven noemt Rawls *rationaliteit*. De partijen streven condities in de maatschappij na die ze het rationele vermogen geven om hun conceptie van het goede te vormen, hervormen en na te streven. Alhoewel de partijen hun eigen specifieke conceptie van het goede niet weten, willen de partijen de capaciteit hebben om een stelsel van doelen te kunnen nastreven, wat deze doelen ook blijken te zijn. Op deze manier geeft dit rationele vermogen de mogelijkheid tot het nastreven van diverse concepties van het goede (1a). Echter, zo claimt Rawls, is een rationeel vermogen ook inherent verbonden met onze menselijkheid en maakt daarom deel uit van ieders conceptie van het goede (1b).

3. Het tweede morele vermogen dat de partijen nastreven noemt Rawls *redelijkheid*. De partijen streven condities in de maatschappij na die ze het redelijke vermogen geven om hun *gevoel van rechtvaardigheid* te ontwikkelen en op te volgen. Rawls claimt dat mensen van nature mee willen werken met verplichtingen als deze passen bij hun gevoel van rechtvaardigheid. Het ontwikkelen en willen handelen vanuit dit gevoel is inherent aan mensen en, in het bijzonder, menselijke samenwerking. Redelijkheid is om deze reden onderdeel van ieders conceptie van het goede (1b): ben je namelijk niet redelijk, dan sluit je jezelf af van de voordelen van menselijke samenwerking.

Deze drie 'hogere-orde'-belangen vormen de basis voor de keuze van de partijen in de oorspronkelijke positie. Rawls vereenvoudigt deze belangen waar nodig door gebruik te maken van de *primaire sociale goederen*. Freeman beschrijft deze primaire goederen als de universele sociale middelen die nodig zijn voor de uitvoering en ontwikkeling van de morele vermogens en (mede daardoor) diverse rationele levensplannen mogelijk maken [3, par. 4]. In deze scriptie vereenvoudigen we de verschillende primaire goederen tot twee categorieën: primaire goederen van vrijheid en economische primaire goederen.

[2] Freeman doet deze claim in [3, par. 4] en baseert deze voornamelijk op [8, p. xii, xiii, 44 en 441].

Sterk met de 'hogere orde'-belangen verbonden is Rawls notie van *stabiliteit*. De belangrijkste basis waarop de partijen hun keuze in de oorspronkelijke positie baseren is gegeven, maar Rawls besteedt een groot deel van *A Theory of Justice* om te beredeneren dat zijn conceptie van rechtvaardigheid stabiel *genoeg* is [8, p. 441]. Ik onderscheid in Rawls' *Theory of Justice* twee vormen van stabiliteit en reduceer deze tot *rationele* en *psychologische* stabiliteit:

1. Een samenleving is wat ik *rationeel stabiel* zal noemen, als *alle* burgers hun eigen conceptie van het goede kunnen nastreven ondanks (of dankzij) hun toewijding aan de principes van rechtvaardigheid. Rawls eist van de partijen dat zij rekening houden met deze eis en de principes niet in kwade trouw kiezen, dat wil zeggen, geen principes kiezen waarvan ze weten dat ze deze misschien niet willen naleven omdat ze niet samengaan met hun conceptie van het goede. (Wat deze conceptie dan ook mag blijken te zijn.) We zien een sterke connectie met het morele vermogen tot rationaliteit: het vermogen je eigen conceptie van het goede te vormen en na te streven [8, p. 126, 153-154].

2. Een samenleving is wat Rawls *psychologisch stabiel* noemt, als alle burgers zich aan de regulerende principes kunnen toewijden vanuit hun gevoel van rechtvaardigheid. Rawls eist van de partijen dat zij principes kiezen waarvan zij denken dat men, wanneer men opgroeit in een samenleving gereguleerd door de gekozen principes, deze ondersteunt vanuit een gevoel van rechtvaardigheid en niet ondermijnt uit gevoelens van jaloezie of haat. We zien een sterke connectie met het morele vermogen redelijkheid: het vermogen om je *gevoel van rechtvaardigheid* te vormen en na te streven. [8, p. 154, 398, 400].

Ik reduceer, gebaseerd op deze paragraaf, de basis van de keuze nu tot twee belangen:

1. Ten eerste kiezen de partijen principes die elke burger de mogelijkheid geven zijn rationele vermogens te ontwikkelen en na te streven. De samenleving zal hierdoor voldoen aan de rationele stabiliteitseis.

2. Ten tweede kiezen de partijen principes die elke burger de mogelijkheid geven zijn redelijke vermogens te ontwikkelen en na te streven. De samenleving zal hierdoor voldoen aan de psychologische stabiliteitseis.

'

2.4 De keuze voor de principes van rechtvaardigheid

In deze paragraaf wil ik duidelijk maken waarom de partijen in de oorspronkelijke positie Rawls' principes van rechtvaardigheid verkiezen boven andere principes. De partijen maken hun keuze, volgens Rawls, uit een gegeven lijst van hedendaagse concepties van rechtvaardigheid.[8, par. 21]. Rawls beperkt zijn argumentatie echter tot twee fundamentele vergelijkingen [8, p. xiv,par. 29].

Ten eerste vergelijken we Rawls' twee principes van rechtvaardigheid met het principe van *gemiddeld utilitarisme*. Het principe van gemiddeld utilitarisme vereist dat we de maatschappij richten op een zo'n hoog mogelijk gemiddeld nut (of versimpeld: welzijn) [8, p. 140]. Rawls claimt dat, gegeven het belang van de partijen bij rationele vermogens en rationele stabiliteit de

partijen gelijke basale vrijheden veiligstellen voor *elke* burger. In een maatschappij waarin basale (primaire goederen van) vrijheid gegarandeerd zijn heeft iedereen namelijk het vermogen zijn conceptie van het goede te ontwikkelen en na te streven. Zo is er *niemand* die zijn conceptie van het goede niet kan nastreven en is een maatschappij rationeel stabiel.
Rawls claimt dat zijn eerste principe deze vrijheden veilig stelt voor *elke* burger. Vrijheden worden door de prioriteit van het eerste principe nooit uitgewisseld ten behoeve van een economisch voordeel. Dit in tegenstelling tot bij gemiddeld utilitarisme, waar de vrijheid van een burger kan worden afgenomen (bijvoorbeeld in ruil voor economische voordelen) als dit het gemiddelde welzijn van een maatschappij bevordert [8, par. 29].

Ten tweede vergelijken we Rawls' twee principes van rechtvaardigheid met dezelfde principes met één belangrijke verandering: het verschilprincipe wordt vervangen door het principe van gemiddelde utilitarisme. Alhoewel basale vrijheden nu zijn gegarandeerd voor elke burgers worden sociale en economische ongelijkheden, in tegenstelling tot bij Rawls' principes, geordend ten behoeven van het gemiddelde welzijn. (We zullen naar deze conceptie van rechtvaardigheid verwijzen met 'begrensd utilitarisme'.)
Rawls claimt dat, gegeven het belang van de partijen bij redelijke vermogens en redelijke stabiliteit, de partijen principes willen die samengaan met hun gevoel van rechtvaardigheid. Rawls verdiept zich in de morele psychologie en zegt dat een hoofdeigenschap van dit gevoel van rechtvaardigheid *reciprociteit* is: 'voor wat, hoort wat'.[8, Deel III]
Het verschilprincipe bevredigt deze behoefte aan reciprociteit als het principe waarbij de slechtsten nooit nog slechter af zijn ten behoeve van zij die al beter af waren. Dit in tegenstelling tot een maatschappij gereguleerd door begrensd utilitarisme, waar men economische voordelen van de minderbedeelden af kan nemen en kan geven aan zij die het al beter hadden, zolang dit het gemiddelde welzijn maar verhoogt.
Dus verkiezen de partijen Rawls' principes van rechtvaardigheid boven begrensd utilitarisme. Dit omdat zij Rawls' principes kunnen ondersteunen uit hun gevoel van rechtvaardigheid, terwijl begrensd utilitarisme, door zijn gebrek aan reciprociteit, een onstabiele samenleving (door bijvoorbeeld jaloerse gevoelens) zal opleveren [8, par. 29].

Hoofdstuk 3

Het modelleren van de oorspronkelijke positie

In dit hoofdstuk wil ik de vragen beantwoorden (1) wat een model is, (2) wat een adequaat 'oorspronkelijke positie'-model zou zijn en (3) op welke manier speltheorie een rol heeft in zo'n model.

3.1 Wat is een model?

Een model is een gestileerde weergave van de werkelijkheid. Een model forceert precisie in terminologie, openheid in aannames en helderheid in deductie en conclusie. De waarde van een model ligt in zijn nut, een adequate representatie van de wereld staat hier in dienst van.[1]

3.2 Wat is een 'oorspronkelijke positie'-model?

Om het nut van een 'oorspronkelijke positie'-model duidelijk te maken beschouwen we eerst, verdiepend op paragraaf 2.1, het nut van de oorspronkelijke positie zelf. Ik beschouw de rol van de oorspronkelijke positie in Rawls' *Theory* als volgt:

Stap 1. Rawls onderzoekt onze intuïtieve aannames over een eerlijke initiële keuzesituatie. (Zie paragraaf 2.2 en paragraaf 2.3.)

Stap 2. Rawls bepaalt welke principes in deze keuzesituatie, de oorspronkelijke positie, worden gekozen gegeven deze aannames. (Zie paragraaf 2.3 en paragraaf 2.4.)

Stap 3. Rawls houdt deze principes tegen het licht van onze weloverwegen oordelen ('considered judgements'). (De invulling van deze stap ligt buiten het bereik van deze scriptie, zie [8, p. 17-19, 104, 507-509].)

Rawls neemt onze intuïtieve aannames over een eerlijke keuzesituatie noch onze (intuïtieve) weloverwogen oordelen als leidend. Op sommige punten zullen we onze oordelen moeten schikken

[1] Ik ben dank verschuldigd aan [2] voor inspiratie op dit punt.

naar de principes. Wijken de principes te ver af van onze dagelijkse weloverwogen oordelen dan zullen we onze aannames over een eerlijke keuzesituatie moeten aanpassen om andere principes in de oorspronkelijke positie te concluderen [8, p.18-19].

Met dit in ons achterhoofd is de kwestie waar wij ons in deze scriptie voornamelijk mee bezighouden stap 2. Ik beschouw stap 1 en 3 als vragen voor de empirie, maar in stap 2 kan de wiskunde een uitkomst bieden. Wiskunde die Rawls streven beantwoordt naar "a kind of moral geometry, with all the rigor which this name connotes" [8, p.105].

Ik beschouw de rol van een 'oorspronkelijke positie'-model dan als volgt: een 'oorspronkelijke positie'-model representeert de rol van de oorspronkelijke positie in de bepaling van rechtvaardige principes; het forceert precisie en openheid in Rawls' aannames over een eerlijke keuzesituatie (in het bijzonder de restricties op de partijen en de belangen van de partijen); en het maakt een heldere deductie van deze aannames naar de keuze van de partijen voor Rawls' principes van rechtvaardigheid.

Het nut van zo'n model ligt dan in zijn verhelderende rol: wat is precies de rol van de oorspronkelijke positie in de bepaling van een rechtvaardige maatschappij? Wat zijn precies de aannames van Rawls over een eerlijke keuzesituatie? Alsmede in zijn verklarende rol: hoe deduceert men van deze aannames naar de principes van rechtvaardigheid? Zijn Rawls' aannames genoeg om te concluderen dat de principes van rechtvaardigheid worden gekozen?

3.3 Wat is een speltheoretisch 'oorspronkelijke positie'-model?

Nu ik duidelijk heb gemaakt wat mijns inziens de rol is van een 'oorspronkelijke positie'-model, wil ik verklaren welke rol speltheorie heeft in zo'n model. Echter, tot nu toe is er weinig gezegd over de *modus operandi* in de oorspronkelijke positie: hoe komen de partijen in de oorspronkelijke positie tot een besluit?

Een cruciale vraag voor deze besluitvorming is of we de partijen (1) allen tezamen als één entiteit beschouwen of (2) beschouwen als individuele actoren die samen tot overeenstemming moeten komen.

In de eerste visie zou de keuze in de oorspronkelijke positie adequaat kunnen worden gemodelleerd met behulp van 'eenvoudige' individuele *'rationele keuze'-theorie*. In deze wiskundige discipline worden keuzeproblemen bestudeerd waar wordt aangenomen dat de actoren rationeel zijn: ze maximaliseren een stelsel van doelen. Dit postulaat komt overeen met Rawls' omschrijving van de partijen die wederzijds ongeïnteresseerd hun conceptie van het goede nastreven. (Zie paragraaf 2.2.)

In deze scriptie bekijken we of er ook een rol voor speltheorie is in het modelleren van de oorspronkelijke positie. *Speltheorie* is de tak van 'rationele keuze'-theorie die keuzeproblemen bestudeert in (complexere) situaties waar strategische interactie plaatsvindt tussen actoren. Deze situaties worden *spellen* genoemd en de actoren *spelers*. Speltheorie is een adequaat gereedschap als we de tweede visie adopteren: de partijen zijn individuele actoren die samen tot overeenstemming moeten komen.

Ik claim dat de visie op de partijen als individuele actoren in (een deel van) het keuzeproces adequaat is en dat er als zodanig een rol voor speltheorie is in het modelleren van de oorspronkelijke

positie. De visie op de partijen als individuele actoren is van belang omdat er in de oorspronkelijke positie een zekere *quid pro quo* is: een uitwisseling van goederen of diensten. Ik denk, Freeman volgend in [3, par. 6.2], dat de partijen hun toewijding aan de principes (ook als deze tegen kortetermijn-eigenbelang in gaat) ruilen voor een vruchtbare samenwerking. Voor Rawls is de keuze in de oorspronkelijke positie ultiem: de partijen beschouwen de gekozen principes als principes waar zij zich de rest van hun leven aan zullen toewijden.[2] De partijen maken dus *binnen* de oorspronkelijke positie een keuze tussen samenwerking (met voor elke burger een onvergankelijke toewijding aan de gekozen regulerende principes) en onthouding van samenwerking. Het is op dit vlak dat er strategische interactie plaatsvindt tussen de partijen en er een rol is voor speltheorie in de modellering van de oorspronkelijke positie.

Ik beschouw de rol van een speltheoretisch 'oorspronkelijke positie'-model dan als volgt: een speltheoretisch model van de oorspronkelijke positie representeert de rol van de oorspronkelijke positie in de bepaling van rechtvaardige principes. Het beschouwt de partijen als individuele actoren met geformaliseerde belangen en restricties die overeenstemming moeten bereiken over principes van rechtvaardigheid. Met behulp van een speltheoretisch raamwerk wordt er een heldere deductie gemaakt van Rawls' aannames over de belangen van de partijen, en restricties op de partijen, naar Rawls' principes van rechtvaardigheid.

De toegevoegde waarde van een *speltheoretisch* model is dat het verheldering kan bieden in de afweging van partijen tussen de toezegging tot principes en een onthouding van samenwerking.

Speltheorie kent het onderscheid tussen *coöperatieve* en *niet-coöperatieve* speltheorie. Het verschil tussen deze twee typen spellen is dat bij coöperatieve spellen bindende afspraken tussen de spelers gemaakt kunnen worden terwijl dit niet mogelijk is bij de niet-coöperatieve spellen. In het algemeen uit dit verschil zich echter vooral in de modelleringstechnieken van een spel. In het volgende hoofdstuk bekijken we een coöperatieve benadering van Anthony Laden [4]. In hoofdstuk 5 bekijken we mijn eigen modellering van de oorspronkelijke positie via een niet-coöperatief raamwerk van Suzumaru en Pattanaik [6].

[2] Dit wordt duidelijk wanneer Rawls in [8, par. 29] praat over de *onherroepelijke* en *bindende* overeenkomsten binnen de oorspronkelijke positie, zo denkt ook Freeman in [3, par. 6.2]. Dit is op zichzelf een interessante kwestie: als de afspraken in de oorspronkelijke positie niet als bindend worden beschouwd kan de toewijding van burgers aan de principes ter discussie worden gesteld (zoals in [1]). De rol van speltheorie wordt dan verschoven van strategieën van de partijen in de oorspronkelijke positie naar strategieën van burgers in de maatschappij.

Hoofdstuk 4

Een coöperatief-speltheoretische benadering

Coöperatieve speltheorie is een tak van speltheorie waar de aanname wordt gemaakt dat spelers coalities kunnen vormen en bindende afspraken maken over de verdeling van de opbrengst van zulke coalities. In dit hoofdstuk beantwoorden we de vraag: wat kan coöperatieve speltheorie betekenen voor een modellering van de oorspronkelijke positie?

We beantwoorden deze vraag aan de hand van Ladens essay 'Games, Fairness and Rawls's Theory of Justice' [4]. In dit essay claimt Laden een 'coöperatief speltheoretisch'-model te geven van de oorspronkelijke positie dat, volgens hem, de betekenis weet te vangen van Rawls' oorspronkelijke positie [4, p.210]. Ik denk zelf dat een modellering van de oorspronkelijke positie, zoals beschreven in hoofdstuk 3, via Ladens weg niet mogelijk is. De modelleringstechnieken die Laden gebruikt gaan namelijk in tegen (1) Rawls' visie van de oorspronkelijke positie en (2) het doel van een speltheoretisch model beschreven in hoofdstuk 3.

Laden maakt in zijn modellering impliciet twee keuzes. Ik maak deze keuzes expliciet en laat op die manier twee valkuilen zien van een coöperatief-speltheoretische benadering van de oorspronkelijke positie. Ik stel echter, na deze valkuilen expliciet te hebben gemaakt, een lezing van Laden voor die laat zien hoe verder onderzoek in de toepassing van coöperatieve speltheorie op Rawls' oorspronkelijke positie zijn vruchten kan afwerpen. Mijns inziens zorgen formele definities van de gebruikte concepten voor een beter begrip, deze zijn te vinden in bijlage A.

Ten eerste beperkt Laden zich tot *TU-spellen* ('transferable utility'-spellen, Def. 2). Dit zijn spellen waarin de aanname wordt gemaakt dat de opbrengst kan worden uitgedrukt in een getal. Dit getal staat voor één medium, bijvoorbeeld geld, waarin de utiliteit van de spelers lineair is, dat wil zeggen, waar iedere speler aan dezelfde stijging dezelfde waarde toekent.[1]

Ik denk dat deze beperking in Ladens model, de opbrengst uitdrukken op één lineaire schaal, niet samen gaat met een cruciale claim van Rawls over de oorspronkelijke positie. Rawls claimt dat in de oorspronkelijke positie primaire goederen van vrijheid en economische primaire goederen op een andere schaal worden gemeten. Laden drukt echter *alle* primaire goederen uit in dezelfde karakteristieke functie (de functie die elke coalitie de waarde toekent die hij zelf kan garanderen,

[1] Laden doet deze aanname impliciet doordat zijn definitie van *core* 5 en *least core* 6 veronderstellen dat een opbrengst kan worden verdeeld en overschot **??** met elkaar kan worden vergeleken op dezelfde lineaire schaal.[4, p. 196-198]

Def. 2) en maakt zo een (lineaire) vergelijking tussen beide goederen mogelijk. Zo'n functie strookt niet met Rawls' claim; zoals we in paragraaf 2.4 zagen ruilen de partijen namelijk nooit primaire goederen van vrijheid voor meer economische goederen. Dit wijst ons op een eerste conclusie over een coöperatieve benadering van de oorspronkelijke positie: een 'oorspronkelijke positie'-model kan niet slechts één lineaire karakteristieke functie hebben over het domein van *alle* primaire goederen.

Ten tweede beschouwt, zoals vaak wordt gedaan bij coöperatieve spellen, Laden geen *individuele* spelers. Hij analyseert een spel op een abstracter niveau. Het niveau waar een spel gedefiniëerd is door de verzameling spelers en een karakteristieke functie (zie Def. 2). Laden veronderstelt vervolgens op dit niveau dat alle spelers samenwerken en zo de *grote coalitie* vormen. Hij beperkt zich door deze beide keuzes tot de vraag: hoe moet de opbrengst van deze grote coalitie worden verdeeld over de spelers?[2] Laden geeft vervolgens een aanzet tot een model zoals geschetst in hoofdstuk 3, met individuele spelers en strategieën, maar vertrouwt uiteindelijk op de, zoals hij het zelf noemt, "spirit of its arguments"[4, p. 191].

Deze visie strookt niet met mijn doel een speltheoretisch 'oorspronkelijke positie'-model om de *keuze* van *individuele* partijen voor samenwerking onder de principes van rechtvaardigheid te modelleren. In plaats van aan elke individuele speler een formeel gedefinieerde functie toe te bedelen die de belangen van de partijen representeert, bekijkt Laden vanuit een abstracter en externer perspectief alleen *eigenschappen van verdelingen* van de opbrengst van de grote coalitie. In zo'n modellering maken de spelers geen keuzes: niet of zij zullen samenwerken en niet met welke strategie zij het beste hun belangen nastreven. Dit wijst ons op een tweede conclusie over een coöperatieve benadering: als we een speltheoretische modellering van de 'oorspronkelijke positie'-model zoals beschreven in hoofdstuk 3 willen maken kunnen we ons niet beperken tot de vraag wat eigenschappen van verdelingen zijn van de opbrengst van de grote coalitie.

Mijn voorstel is dat we Ladens coöperatieve benadering (niet als een modellering in termen van hoofdstuk 3 zien, maar) beschouwen als een zoektocht naar nieuwe argumenten voor Rawls' principes van rechtvaardigheid. We kunnen Ladens conclusie in dit licht herformuleren tot twee claims:

1. Volgens Laden kunnen burgers utilitarisme afwijzen (m.a.w. kiezen zij het eerste principe van Rawls' twee principes) omdat deze conceptie hen niet de basale vrijheden geeft die ze kunnen garanderen in de oorspronkelijke positie.

2. Volgens Rawls kunnen burgers begrensd utilitarisme afwijzen (m.a.w. kiezen het verschil-principe) omdat ze een bepaalde reciprociteit verlangen van de principes: reciprociteit te vinden in een maximin-verdeling van goederen. Laden ziet een overeenkomst met het afwijzen van distributies buiten de *least core*: de least core is een oplossingsconcept van coöperatieve speltheorie (zie Def. 6). In de least core zitten alleen distributies waar het maximale *overschot* (het verschil tussen wat een coalitie zelf kan garanderen en krijgt, Def. 4) wordt geminimaliseerd [4, p. 217].

Met zijn eerste claim stelt Laden, geïnspireerd door de *core* (een oplossingsconcept in coöperatieve speltheorie waar elke coalitie tenminste krijgt wat hij zelf kan garanderen, Def. 5), dat de basale vrijheden die een partij zelf kan garanderen in de oorspronkelijke positie relevant zijn

[2] Laden beperkt zich niet bewust tot deze vraag, maar dit is naar mijn mening het gevolg van zijn aannames op[4, p. 196-198, p. 211]

voor een rechtvaardige verdeling van goederen. In Ladens tweede claim stelt hij het maximaliseren van het minimale aantal goederen gelijk aan het minimaliseren van het maximale overschot. Ook hier claimt Laden, door het overschot als maatstaaf te nemen, dat het relevant is voor rechtvaardigheid wat een coalitie zelf kan garanderen.

In beide van Ladens claims is het dus relevant wat een partij in de oorspronkelijke positie zelf (of met een coalitie) kan garanderen; dit is niet verassend nadat we in hoofdstuk 3 vaststelden dat de rol van speltheorie zich beperkt tot de vraag of partijen in de oorspronkelijke positie samenwerken of zich onthouden van samenwerking.

Verder onderzoek in deze hoek kan nieuwe argumenten aandragen voor Rawls' principes van rechtvaardigheid. Uiteindelijk zou dit kunnen leiden tot een coöperatief model. Dit model zou niet zozeer een representatie van Rawlsiaanse partijen zijn die kiezen voor principes in de oorspronkelijke positie, maar een abstracter coöperatief model van de distributie van goederen over burgers. De karakteristieke functie van zo'n model representeert wat de burgers zelf kunnen bemachtigen; echter niet wat zij kunnen bemachtigen in een maatschappij onderhevig aan moreel irrelevante contingenties, maar in een eerlijke initiele keuzesituatie: de oorspronkelijke positie. Verder leert deze beschouwing van Laden ons dat in zo'n accuraat 'oorspronkelijke positie'-model niet *alle* primaire goederen worden geschaald aan één lineaire karakteristieke functie, maar er een onderscheid moet worden gemaakte tusse primaire goederen van vrijheid en primaire goederen van economie. Dit model zou kunnen laten zien dat eenzelfde argumentatie voor distributies in de core en least core kan opgaan voor distributies waar primaire goederen van vrijheid worden gegarandeerd en economische primaire goederen worden verdeeld via een maximin-distributie.

In deze scriptie zullen we zo'n model niet verder beschouwen. Het is niet een adequaat 'oorspronkelijke positie'-model in de termen van hoofdstuk 3: het representeert niet de keuze van de partijen, gebaseerd op de door Rawls' gestelde belangen, tot samenwerking onder de principes van rechtvaardigheid. In het volgende hoofdstuk beschouwen we of de niet-coöperatieve speltheorie wel zo'n model kan bieden.

Hoofdstuk 5

Een niet-coöperatief-speltheoretisch model

In dit hoofdstuk presenteer ik een niet-coöperatief-speltheoretisch model gebaseerd op een formeel raamwerk gegeven door Pattanaik en Suzumaru. Dit model voldoet naar mijns inziens aan de vorm geschetst in hoofdstuk 3.

In niet-coöperatieve speltheorie ligt de focus op individuele spelers en hun strategieën. De overeenstemming van de partijen kan in niet-coöperatieve speltheorie de vorm aannemen van een *Nash-equilibrium*. Een Nash-equilibrium is een verzameling strategieën, voor elke speler één, waar elke strategie de *beste reactie* is op de andere gespeelde strategieën, dat wil zeggen, in een equilibrium kan geen individuele speler voordeel behalen door zijn strategie te veranderen.

5.1 Het raamwerk van Pattanaik en Suzumura

In [6] stellen Pattanaik en Suzumura een raamwerk op voor de analyse van sociale keuze en welzijn. Dit raamwerk maakt gebruik van een informatiebasis van een individuele preferentieordening over het Cartesiaanse product van X en $\mathfrak{U}$. X is de verzameling van conventioneel gedefinieerde *sociale alternatieven*, waarbij een conventioneel gedefiniëerd sociaal alternatief een complete omschrijving is van de maatschappij met uitzondering van een *sociaal beslissingsmechanisme*. $\mathfrak{U}$ is de verzameling van sociale beslissingsmechanismes; een mechanisme, gewoonlijk bestaande uit een complexe verzameling regels, waarmee de maatschappij een sociaal alternatief bepaalt [6, 194].

In deze paragraaf zal ik een beperkte versie van dit raamwerk geven, toegespitst zodat ik in paragraaf 5.2 een modellering van de oorspronkelijke positie kan instantiëren in dit raamwerk. Het raamwerk in deze paragraaf wordt formeel gebracht, de intuïtie achter de gemaakte keuzes wordt echter helder in mijn instantiatie in de volgende paragraaf.

We beginnen met het definiëren van een verzameling rechtenstructuren $\mathfrak{G}$, een interpretatie van de sociaal beslissingsmechanismes $\mathfrak{U}$ met een focus op individuele rechten [6, 195].

Laat $\mathfrak{A}$ de verzameling van niet-lege eindige deelverzamelingen van X zijn. Elk element A van $\mathfrak{A}$ representeert een *kwestie* ('issue'), dat wil zeggen, een specifieke verzameling van sociale alternatieven [6, 195].

Definitie 1. Voor elk kwestie $A \in \mathfrak{A}$, wordt een A-gebaseerd spel G_A gedefinieerd door een (n+2)-tupel $(N, S_{1A}, \ldots, S_{nA}, g_A)$, waar, voor elke speler $i \in N$, S_{iA} staat voor de verzameling van toelaatbare strategieën van i en g_A voor de uitkomstfunctie die elke strategie-n-tupel $s^N = (s_1, \ldots, s_n) \in S_A^N := \times_{i \in N} S_{iA}$ afbeeldt op [een sociaal alternatief in] A. Een rechtenstructuur G is een specificatie van een A-gebaseerd spel G_A voor elke $A \in \mathfrak{A}$, hieruit volgt $G = \{G_A \mid A \in \mathfrak{A}\}$. $\mathfrak{G}$ staat voor de verzameling van alle denkbare *rechtenstructuren* [6, p. 196].

Om te bepalen welke strategie $s_i \in S_{iA}$ een speler $i \in N$ kiest, specificeren we de individuele preferenties van speler i over de verzameling van sociale alternatieven X. Laat voor een gegeven verzameling X, $\mathfrak{R}(X)$ de verzameling zijn van mogelijke ordeningen over X. Elk element van $\mathfrak{R}^n(X)$, het n-voudig Cartesiaans product van $\mathfrak{R}$(X), is een profiel van individuele prefentie-ordeningen over X [6, p. 196].

Een spelvorm G_A en een profiel $R^N = (R_1, .., R_n) \in \mathfrak{R}^n(X)$ van individuele preferentie-ordeningen beschrijven samen een spel (G_A, R^N). Laat $T(G_A, R_G^N)$ de verzameling, aangenomen als niet-leeg, van alle $x \in A$ zijn zodat x door de maatschappij wordt verwacht als mogelijke uitkomst van het spel (G_A, R_G^N). Deze uitkomst wordt bepaald door de verzameling $E(G_A, R_G^N)$, de verzameling van (pure) Nash-equilibria van het spel (G_A, R_G^N) en g_A (de uitkomstfunctie van het spel G_A) zodat $T(G_A, R_G^N) = g_A[E(G_A, R_G^N)]$ [6, p. 198-199].

5.2 Een model van de oorspronkelijke positie

In deze paragraaf wil ik een 'oorspronkelijke positie'-model geven, geinitieerd in dit raamwerk, dat de functie heeft zoals beschreven in hoofdstuk 3.

Dit model neemt de vorm aan van een beperkte versie van Pattanaik en Suzumura's 'sociale keuze'-procedure die zij voor ogen hadden in hun paper [6, p. 203-204]. Ik zal mij nu beperken tot een uitleg over de manier waarop het raamwerk wordt geïnitieerd zodat deze de oorspronkelijke positie representeert. In paragraaf 5.3 geef ik verdere toelichting over waarom ik dit model als een adequaat speltheoretisch 'oorspronkelijke positie'-model beschouw.

Stap 1: Identificatie van rechtenstructuur, kwestie en het profiel van preferentie-ordeningen
In deze stap worden door de samenleving een rechtenstructuur, kwestie en profiel van preferentie-ordeningen gepostuleerd als relevant voor de 'sociale keuze'-procedure.

Laat de rechtenstructuur $G \in \mathfrak{G}$ de 'oorspronkelijke positie'-rechtenstructuur G_{op} zijn. Laat de kwestie $A \in \mathfrak{A}$ worden bepaald door de lijst van concepties van rechtvaardigheid (zie paragraaf 2.3 en [8, p. 102-109]). Zodat elk sociaal alternatief x_ϕ uit de relevante kwestie A_c een maatschappij representeert gereguleerd door een conceptie van rechtvaardigheid ϕ staande op deze lijst. We beperken ons tot een vereenvoudiging en beschouwen de volgende concepties:

- pr:= Rawls' principes van rechtvaardigheid
- bu:= begrensd utilitarisme

- u:= gemiddeld utilitarisme
- a:= een andere conceptie uit de lijst

Ik wil aan de verzameling A_c nog één element toevoegen. Een staat z waarin er geen overeenstemming tussen de burgers is over een regulerende conceptie van rechtvaardigheid. Zodat $A_c = \{x_{pr}, x_{bu}, x_u, x_a, z\}$.

Laat G_{A_c}. het volgende n+2-tupel $(N, S_{1A_c}, \ldots, S_{nA_c}, g_{A_c})$ zijn (zie figuur 5.1) :

- De verzameling spelers $N := \{1, \ldots, i, \ldots, n\}$, met $2 \leq n < \infty$ representeert de partijen in de oorspronkelijke positie.
- De verzameling strategieën voor een speler i, S_{iA_c}, wordt bepaald door A_c. Een strategie $s_{iA_c} \in S_{iA_c}$ is de keuze voor een specifieke conceptie van rechtvaardigheid. Gegeven onze vereenvoudigde A_c geldt $S_{iA_c} = \{pr, bu, u, a\}$.
- De uitkomstfunctie g_{A_c} beeldt elke strategie-n-tupel $s^N = (s_1, \ldots, s_n) \in S_{A_c}^N$ af op een sociaal alternatief in A_c volgens het volgende functievoorschrift:

$$g_A(s^N) = \begin{cases} x_\phi & \text{als } \exists \phi \in S_A : \forall i, j \in N : s_i = s_j = \phi \\ z & \text{in alle andere gevallen} \end{cases}$$

  dat wil zeggen als alle[1] spelers dezelfde principes ϕ kiezen leven ze in een maatschappij gereguleerd door die principes x_ϕ, kiezen ze verschillende principes dan leven ze in een maatschappij z waar er geen overeenstemming tussen de burgers is over een regulerende conceptie van rechtvaardigheid.

	pr	bu	u	a
pr	x_{pr}	z	z	z
bu	z	x_{bu}	z	z
u	z	z	x_u	z
a	z	z	z	x_a

Figuur 5.1: $G_A \in G_{op}$ vereenvoudigd tot een tweespeler-spelvorm.

Laat voor elke speler $i \in N$ zijn preferentie-ordening $R_{iG} \in \mathfrak{R}(X)$ afhangen van de basis voor de keuze in de oorspronkelijke positie beschreven in paragraaf 2.3. Merk op dat op dit punt in de analyse de strategische interactie tussen de spelers geen rol speelt en de preferentieordening kan worden bepaald door een voor elke speler (door de sluier van onwetendheid) identieke, utiliteitsfunctie u_{iG}. Deze functie beeldt elk sociaal alternatief $x \in A_c$ af op een reëel getal $u_{iG}(x)$ dat de waarde van een maatschappij x voor een speler representeert.[2] Elke maatschappij wordt dan beoordeeld door een partij op de twee relevante belangen behandeld in paragraaf 2.3, rationaliteit en redelijkheid:

- (1) := een maatschappij is rationeel stabiel en beantwoordt aan de behoefte van mensen om hun morele vermogen tot rationaliteit te ontwikkelen en na te streven.

[1] Rawls claimt dat de keuze van principes unaniem moet zijn [8, p. 106].
[2] Strict genomen hebben de partijen een preferentie-ordening over X, ik beperk me echter tot het relevante deel $A_c \subseteq X$.

- (2) := een maatschappij is psychologisch stabiel en beantwoordt aan de behoefte van de mensen om hun morele vermogen tot redelijkheid te ontwikkelen en na streven.

De functie u_{iG} die de preferentie-ordening $R_{iG} \in \mathfrak{R}(X)$ bepaalt maakt een onderscheid tussen vier gevallen:

$$u_{iG}(x) = \begin{cases} \delta & \text{als } x = z \\ \gamma & \text{als } x \neq z \text{ en } \neg(1) \\ \beta & \text{als } x \neq z, (1) \text{ en } \neg(2) \\ \alpha & \text{als } x \neq z, (1) \text{ en } (2) \end{cases}$$

Stap 2: Evaluatie van mogelijke sociale alternatieven In deze stap wordt gegeven de gepostuleerde G, A en R_G^N bepaalt wat de mogelijke sociale alternatieve $B \in \mathfrak{A}$ zijn.

Eerst wordt elk sociaal alternatief geëvalueerd aan de hand van u_{iG} en Rawls' evaluatie van gemiddeld utilitarisme, begrensd utilitarisme en de principes van rechtvaardigheid. (Of een maatschappij voldoet aan (1) en (2) is beschreven in paragraaf 2.4).

Het spel (G_A, R_G^N) (vereenvoudigd tot een tweespeler-spel) wordt nu gerepresenteerd door figuur 5.2.

	pr	bu	u	a
pr	α	δ	δ	δ
bu	δ	β	δ	δ
u	δ	δ	γ	δ
a	δ	δ	δ	γ

Figuur 5.2: Het spel (G_A, R_G^N), waar R_G^N wordt uitgedrukt door de functie u_G^N

De maatschappij bepaalt nu aan de hand van dit spel (G_A, R_G^N) de mogelijke sociale alternatieven $B \in \mathfrak{A}$ zodanig dat $B = T(G_A, R_G^N)$. De verzameling $T(G_A, R_G^N)$ van sociale alternatieven die door de maatschappij worden verwacht als mogelijk uitkomst van dit spel wordt, zoals behandeld in paragraaf 5.1, bepaald door de pure Nash-equilibria $E(G_A, R_G^N)$ en de functie g_A. Zodat: $B = T(G_A, R_G^N) = g_A[E(G_A, R_G^N)$.

In de bepaling van de pure Nash-equilibria van het spel (G_A, R_G^N) is van belang wat de ordening is op $\alpha, \beta, \gamma, \delta$. We behandelen twee representatieve gevallen:

Geval 1. Stel R_G^N is zodanig dat $\alpha > \beta > \gamma > \delta$. We laten een formeel bewijs over aan de lezer en markeren de pure Nash-equilibria $E(G_A, R_G^N)$ in figuur 5.3 met een streep:

	pr	bu	u	a
pr	$\bar{\alpha}$	δ	δ	δ
bu	δ	$\bar{\beta}$	δ	δ
u	δ	δ	$\bar{\gamma}$	δ
a	δ	δ	δ	$\bar{\gamma}$

Figuur 5.3: De pure Nash-equilibria $E(G_A, R_G^N)$ in het geval dat $\alpha > \beta > \gamma > \delta$.

Nu is de verzameling mogelijke sociale alternatieven:

$$B = T(G_A, R_G^N) = g_A[E(G_A, R_G^N)] = \{x_{pr}, x_{bu}, x_u, x_a\}$$

Geval 2. Stel $R_G^N{}'$ is zodanig dat $\alpha > \delta > \beta > \gamma$. We laten een formeel bewijs over aan de lezer en markeren de pure Nash-equilibria in figuur 5.4 met een streep:

	pr	bu	u	a
pr	$\bar{\alpha}$	δ	δ	δ
bu	δ	β	$\bar{\delta}$	$\bar{\delta}$
u	δ	$\bar{\delta}$	γ	$\bar{\delta}$
a	δ	$\bar{\delta}$	$\bar{\delta}$	γ

Figuur 5.4: De pure Nash-equilibria $E(G_A, R_G^N)$ in het geval dat $\alpha > \delta > \beta > \gamma$.

Nu is de verzameling mogelijke sociale alternatieven:

$$B = T(G_A, R_G^N) = g_A[E(G_A, R_G^N)] = \{x_{pr}, z\}$$

5.3 Toelichting bij mijn model

In deze laatste paragraaf licht ik toe waarom dit model een adequaat speltheoretisch 'oorspronkelijke positie'-model is zoals beschreven in hoofdstuk 3.

Ten eerste zou, volgens hoofdstuk 3, een speltheoretisch 'oorspronkelijke positie'-model de rol van de oorspronkelijke positie in de bepaling van rechtvaardige principes moeten representeren:

- In mijn model representeert de keuze van een samenleving voor de rechtenstructuur G_{op} (zie stap 1) de keuze van een samenleving voor de oorspronkelijke positie als geschikte initiele positie. De samenleving bepaalt met een rechtenstructuur $G \in \mathfrak{G}$ wat de strategieën zijn, wie er beslist over deze strategieën en hoe deze strategieën op een sociaal alternatief worden afgebeeld; net zoals de samenleving in Rawls' gedachtenexperiment met een geschikte initiële keuzesituatie bepaalt wat de keuzes zijn van de partijen, wie de handelende actoren zijn en hoe de keuze voor een conceptie van rechtvaardigheid verloopt.
- In mijn model representeert de kwestie A_c (zie stap 1) de lijst van concepties van rechtvaardigheid. Net zoals de samenleving met een kwestie $A \in \mathfrak{A}$ bepaalt wat de relevante sociale alternatieven zijn, bepaalt Rawls wat de relevante concepties van rechtvaardigheid zijn. De afhankelijkheid van $G_A \in G$ van $A \in \mathfrak{A}$ komt overeen met Rawls' claim dat "A argument for the principles, or indeed for any conception, is always relative to some list of alternative principles"[8, p. 109].
- In mijn model representeert de individuele preferentieordening $R_{iG} \in \mathfrak{R}(X)$ (zie stap 1) de evaluatie van de partijen van de concepties van rechtvaardigheid. Net zoals in het model de evaluatie ($R_{iG} \in \mathfrak{R}(X)$) van de sociale alternatieven $x \in A_c$ afhankelijk is van de rechtenstructuur G_{op} is in Rawls' oorspronkelijke positie de evaluatie van de concepties van rechtvaardigheid afhankelijk van de restricties (zoals de sluier van onwetendheid) in de oorspronkelijke positie.

Ten tweede zou, volgens hoofdstuk 3, een speltheoretisch 'oorspronkelijke positie'-model de partijen modelleren als individuele actoren met geformaliseerde belangen en restricties.

- In dit model worden de partijen gemodelleerd als individuele spelers die rationeel hun utiliteits-functie u_{iG} maximaliseren. De belangen van de partijen, zoals beschreven in paragraaf 2.4 zijn geformaliseerd in deze functie u_{iG}. Ook de restricties zijn in deze functie verwerkt. u_{iG} representeert namelijk de evaluatie van een maatschappij vanuit de oorspronkelijke positie, terwijl zij de sluier van onwetendheid dragen. De functie u_{iG} is dan ook identiek voor elke speler.

Ten derde zou, volgens hoofdstuk 3, de rol van speltheorie duidelijk moeten worden in dit model. Speltheorie heeft, volgens hoofdstuk 3, een rol in de strategische interactie die gepaard gaat met het bereiken van overeenstemming over een conceptie van rechtvaardigheid:

- In dit model wordt er een formeler onderscheidt duidelijk tussen de individuele preferentie-ordening $R_{iG} \in \mathfrak{R}(X)$, waar strategische interactie geen rol heeft en de bepaling van de mogelijke sociale alternatieve $B \in G \in \mathfrak{G}$, waar strategische interactie wel een rol heeft.
- In dit model wordt de rol van speltheorie beperk tot het punt waar er zoals eerder gezegd een zekere *quid pro quo* is. De partijen kiezen tussen (strategieën die kunnen leiden tot) een maatschappij z waar er geen overeenstemming is over een conceptie van rechtvaardigheid of een maatschappij x_ϕ waar de maatschappij wordt gereguleerd door conceptie ϕ.
- Uit dit model blijkt dat het relevant is hoe een maatschappij z wordt beoordeeld door de partijen in hun individuele preferentie-ordening $R_{iG} \in \mathfrak{R}(X)$. Rawls geeft echter, in *A Theory of Justice*, naar mijn mening, geen uitsluitsel over hoe een maatschappij die niet wordt reguleert door principes moet worden beoordeeld. We beschouwen daarom in het model twee representatieve gevallen.
- De overeenstemming over de principes van rechtvaardigheid wordt in dit model gerepresenteerd door een puur Nash-equilibria.
  - In geval 1, waar een maatschappij z zo laag mogelijk wordt geëvalueerd, leiden deze Nash-equilibria tot mogelijke sociale alternatieven $\{x_{pr}, x_{bu}, x_u, x_a\}$, dat wil zeggen, de partijen kunnen een overeenstemming bereiken tot een maatschappij gereguleerd door welke conceptie van rechtvaardigheid dan ook.
  - In geval 2, waar alleen Rawls' principes van rechtvaardigheid hoger worden geëvalueerd dan een maatschappij z, leiden deze Nash-equilibria tot mogelijke sociale alternatieven $\{x_{pr}, z\}$, dat wil zeggen, de partijen kunnen een overeenstemming bereiken tot zowel een maatschappij gereguleerd door de principes van rechtvaardigheid als een maatschappij waar er geen overeenstemming is over een conceptie van rechtvaardigheid.

Dit model heeft een verhelderende en verklarende rol over de oorspronkelijke positie. Daarnaast leidt dit model tot twee interessante conclusies. Ten eerste moet, gegeven de aannames die zijn gedaan in deze modellering van de oorspronkelijke positie, er worden bepaald hoe een maatschappij z, waar er geen overeenstemming is over een conceptie van rechtvaardigheid, wordt geëvalueerd. Met andere woorden: hoe beoordelen de partijen een maatschappij waarin zij niet samenwerken onder regulerende principes?
Ten tweede laat dit model zien dat, ook al is een maatschappij gereguleerd door Rawls' principes van rechtvaardigheid het Pareto-optimale[3] sociale alternatief van de partijen, het niet evident is

[3]Met Pareto-optimaal bedoel ik in deze context het sociale alternatief dat elke speler bovenaan zijn preferentie-ordening heeft. In de twee gevallen die ik behandel zijn dit Rawls' principes van rechtvaardigheid.

dat zij overeenstemming bereiken over een maatschappij onder deze principes. Het is zelfs zo dat er geen (voor elke partij identieke) preferentie-ordening is waar Rawls' principes van rechtvaardigheid het enige Nash-equilibria vormen.

Ik wil eindigen met twee interessante punten. Ten eerste is er een verband tussen de coöperatieve benadering en de niet-coöperatieve benadering beschreven in deze scriptie. In beide benadering wordt het evident dat het relevant is wat er gebeurt als de partijen kiezen om niet samen te werken onder regulerende principes. In de coöperatieve benadering uit dit zich in een karakteristieke functie die iets moet vertellen over wat een partij zelf kan garanderen; in een niet-coöperatieve benadering uit zich dit in de evaluatie van een maatschappij z waarin er geen overeenstemming is over een conceptie van rechtvaardigheid.
Er zit echter ook een verschil tussen beiden benadering: in de coöperatieve speltheorie kunnen er bindende afspraken worden gemaakt tussen spelers. De grote coalitie N kan worden gevormd en de pareto-optimale keuze om de principes te verdelen volgens Rawls' principes van rechtvaardigheid kan worden gemaakt. In de niet-coöperatieve benadering kan adequaat de individuele afweging van partijen worden gemodelleerd tussen een samenwerking onder regulerende principes of geen samenwerking.
Ten tweede is Suzumaru's en Pattanaiks origineel bedoelt als raamwerk om de keuze voor een sociaal alternatief te baseren op een preferentie-ordening van paartjes bestaande uit een sociaal alternatief ($x \in X$) en een keuzemechanisme ($G \in \mathfrak{G}$). Ik heb dit buiten mijn beschouwing gelaten van het raamwerk; echter een uitbreiding van mijn model via deze weg zou de ruimte geven om de evaluatie van een initiële keuzesituatie en een (al dan niet uit deze keuzesituatie voortkomende) maatschappij te combineren. Dit is in essentie het doel wat Rawls voor ogen heeft in *A Theory of Justice*: met wiskundige zekerheid de evaluatie van verschillende geschikte keuzesituaties en de resulterende concepties van rechtvaardigheid te kunnen vergelijken.

Hoofdstuk 6

Conclusie en verder onderzoek

In deze scriptie onderzocht ik de mogelijkheid tot een speltheoretisch model van de oorspronkelijke positie. Ik deed dit door eerst in hoofdstuk 2 Rawls' oorspronkelijke positie te omschrijven en de basis van de keuze van de partijen te reduceren tot twee belangen: een belang in *rationaliteit* en een belang in *redelijkheid*.

In hoofdstuk 3 gaf ik aan wat volgens mij een adequate speltheoretische modellering van de oorspronkelijke positie is: een model dat de rol van de oorspronkelijke positie representeert in bepaling van rechtvaardige principes; een model dat de partijen beschouwt als *individuele* actoren, met geformaliseerde belangen, die overeenstemming moeten bereiken over principes van rechtvaardigheid; een model waarin de rol van speltheorie zich beperkt tot strategische interactie en verheldering biedt in de afweging van individuele partijen tussen toezegging tot principes of een onthouding van samenwerking.

In hoofdstuk 4 behandelde ik de rol van een coöperatieve benadering in een modellering van de oorspronkelijke positie. Ik heb laten zien dat een speltheoretische modellering zoals beschreven in hoofdstuk 3 niet mogelijk is in de door Laden gebruikte modelleringstechnieken van coöperatieve speltheorie.. Ik stel echter voor om Ladens coöperatieve benadering als een zoektocht naar nieuwe argumenten voor Rawls' principes van rechtvaardigheid te zien. Verder onderzoek kan worden gedaan naar een coöperatief model met een karakteristieke functie die (1) een onderscheid maakt tussen primaire goederen van vrijheid en economische primaire goederen en (2) representeert wat een burger kan bemachtigen in de oorspronkelijke positie.

In hoofdstuk 5 gaf ik een eigen niet-coöperatief model van de oorspronkelijke positie. Ik hoop dat dit model meer duidelijkheid biedt over het complexe werk van Rawls. Dit model leidt tot de conclusie dat (1) er meer moet worden gezegd over hoe de partijen in de oorspronkelijke positie een maatschappij beoordelen waar er geen overeenstemming is over de conceptie van rechtvaardigheid en (2) het niet evident is dat de partijen kiezen voor de principes van rechtvaardigheid.

Als laatste liet ik een overeenkomst zien tussen beide speltheoretische benadering zien en stelde ik een onderzoek voor tot de uitbreiding van mijn model uit hoofdstuk 5. Uiteindelijk zou dit model kunnen leiden tot Rawls doel in *A Theory of Justice*: met wiskundige zekerheid de evaluatie van verschillende geschikte keuzesituaties en de resulterende concepties van rechtvaardigheid combineren.

Bijlage A

Coöperatieve speltheorie

In deze bijlage definieer ik enkele concepten uit de coöperatieve speltheorie. Ten eerste volg ik [7, p. 123] in de definitie van een *TU-spel* en *uitbetalingsdistributie*. Laden gebruikt in 'Games, Fairness and Rawls's A Theory of Justice' [4] het concept *overschot* ('excess') en definieert aan de hand hiervan de *core* en de *least core*. Ik volg [7, p. 124,127] in de definitie van overschot en de core en [5] in de definitie van de least core. Echter bij gebrek aan standaardwerk op dit vlak, herdefinieer ik deze laatste drie zelf in termen van overschot aan de hand van Ladens suggesties.[1]

Om de lezer niet geheel aan zijn lot over te laten gebruiken we een voorbeeld.

Voorbeeld 1. In het wantenspel hebben Conny en Toos allebei drie wanten. Een paar wanten levert vijf euro op, een enkele want niets. Conny en Toos kunnen dus los van elkaar vijf euro verdienen. Ze kunnen er echter ook voor kiezen samen te werken en zo vijftien euro te verdienen, waarna ze dit bedrag onderling moeten verdelen.

TU-spel Een *TU-spel* bestaat uit twee delen, een verzameling spelers en de karakteristieke functie die ons vertelt welke opbrengst elke coalitie zelf kan garanderen.

Definitie 2. Een *coöperatief spel met transferable utility* (een *TU-spel*) is een paar (N, v), waar $N = \{1, 2, ..., n\}$ met $n \in \mathbb{N}$ de verzameling *spelers* is, en v een functie ($v : 2^N \to \mathbb{R}$) die aan elke *coalitie* S, *i.e.*, aan elke deelverzameling $S \subseteq N$, een reëel getal v(S) toekent, waarbij $v(\emptyset) = 0$. De functie v heet de *karakteristieke functie* en v(S) wordt de *waarde* van S genoemd.[2] De coalitie N heet de *grote coalitie*.

In voorbeeld 1 geldt:

$$N = \{1, 2\} \quad v(\{\emptyset\}) = 0 \quad v(\{1\}) = 5 \quad v(\{2\}) = 5 \quad v(\{1, 2\}) = 15$$

Uitbetalingsdistributie De uitkomst van een TU-spel wordt gepresenteerd door een *uitbetalingsdistributie*: een vector die staat voor de opbrengst per speler.

[1] Ik ben dank verschuldigd aan [11] op dit laatste punt.
[2] We zullen ook over v(S) spreken als de opbrengst die de coalitie S zelf kan garanderen.

Definitie 3. Een *uitbetalingsdistributie* $(x_1, ..., x_n) \in \mathbb{R}^n$ is een n-tupel van reële getallen $(x_i)_{i \in N}$. We gebruiken vaak de notatie $x(S) := \sum_{i \in S} x_i$ om de totale opbrengst aan te geven toegekend aan een coalitie $S \subseteq N$.

Enkele uitbetalingsdistributies in voorbeeld 1 zijn:

$$x^a = (0,0) \quad x^b = (1,14) \quad x^c = (100,0) \quad x^d = (7,8)$$

Overschot Het *overschot* ('excess') van een coalitie $S \subseteq N$ voor een uitbetalingsdistributie $x \in \mathbb{R}^n$ geeft aan hoeveel meer een coalitie S zelf kan garanderen dan zijn opbrengst $x(S)$.

Definitie 4. De *overschot* van een coalitie $S \subseteq N$ voor een uitbetalingsdistributie $x \in \mathbb{R}^n$ is gedefinieerd als $e(S, x) = v(S) - x(S)$

In voorbeeld 1 is het overschot van elke $S \subseteq N$ voor x^b:

$$e(\emptyset, x^b) = 0 - 0 = 0 \qquad e(\{1\}, x^b) = 5 - 1 = 4$$
$$e(\{2\}, x^b) = 5 - 14 = -9 \qquad e(\{1,2\}, x^b) = 15 - 15 = 0$$

Core De core is een oplossing voor een coöperatief spel. Daarmee wordt bedoeld dat er redenen zijn om uitbetalingsdistributies in de core te verkiezen boven andere uitbetalingsdistributies. De core is de verzameling distributies $x \in \mathbb{R}^n$ die (1) sommeren tot de waarde van de grote coalitie, waarbij (2) iedere speler minstens krijgt wat hij zelf kan garanderen en waarbij (3) iedere coalitie minstens krijgt wat hij zelf kan garanderen.

Definitie 5. Laat (N, v) een TU-spel zijn. Een vector $x \in \mathbb{R}^n$ heet een *imputatie* als deze voldoet aan de volgende voorwaarde:

1. x is *efficiënt, i.e.*, $e(N, x) = 0$
2. x is *individueel-rationeel, i.e.*, $\forall i \in N \; e(\{i\}, x) \leq 0$
3. x is *coalitie-rationeel i.e.* $\forall S \subseteq N e(S, x) \leq 0$

De core wordt genoteerd als $C(N, v)$.

Zo is in ons voorbeeld: $x^d \in C(N, v)$ en $x^a \notin C(N, v)$, $x^b \notin C(N, v)$, $x^c \notin C(N, v)$.

De ϵ-core en de least core Een uitbetalingsdistributie in de *ϵ-core* geeft iedere coalitie maximaal ϵ minder dan hij zelf kan garanderen.

Definitie 6. De ϵ-core is de verzameling van alle imputaties $x \in C(N, v)$ zodat
$\forall S \subseteq N \; e(S, x) \leq \epsilon$.

Bij de *least core* is deze epsilon minimaal.

Definitie 7. De ϵ-core met de kleinste ϵ waarvoor de ϵ-core niet leeg is heet de *least core*.

In ons voorbeeld is de -2,5-core de least core en is $x^e = (7.5, 7.5)$ de enige imputatie in de least core. Deze verdeling minimaliseert (over de imputaties) het maximale (over de coalities) overschot.

Bibliografie

[1] Ken Binmore. *Game Theory and the Social Contract*, volume 1: Playing Fair. The MIT Press.

[2] Morris P Fiorina. Formal models in political science. *American Journal of Political Science*, 19:133–159, 1975.

[3] Samuel Freeman. Original position. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2012 edition, 2012.

[4] Anthony Laden. Games, fairness, and Rawls's a theory of justice. *Philosophy and Public Affairs*, 20:189–222, 1991.

[5] M. Maschler, B. Peleg, and Lloyd S. Shapley. Geometric properties of the kernel, nucleolus, and related solution concepts. *Mathematics of Operations Research*, 4:303 tot 338, 1979.

[6] Prasanta K. Pattanaik and Kotaro Suzumura. Individual rights and social evaluation: A conceptual framework. *Oxford Economic Papers*, 48:194–212, 1996.

[7] Hans Peeters. *Game Theory: A Multi-Leveled Approach*. Springer, 2008.

[8] John Rawls. *A Theory of Justice: Revised Edition*. Harvard University Press, Cambridge, Mass., 1999.

[9] John Rawls. *Justice as Fariness: A Restatement*. Harvard University Press, Cambridge, Mass., 2001.

[10] John Rawls. *Een theorie van rechtvaardigheid*. Lemniscaat, 2008. trans. Frank Bestebreurtje, oorspr. uitg. *A Theory of Justice (Revised Edition, 1999)*.

[11] T.B. Rombouts. Solutions of argumentation in cooperative game theory. master thesis, University of Utrecht, August 2004.