Universiteit Utrecht

Faculteit Geesteswetenschappen

Kunstmatige Intelligentie

Moraliteitsmetriek in het Snowdrift spel

Auteur
B.C.G. Knüppe

Begeleider
Dr. Gerard Vreeswijk

April 2019

Samenvatting

In deze paper wordt er gekeken naar de toepassing van moraliteitsmetriek in het Snowdrift spel. Moraliteit in de werkelijk wereld is een moeilijk begrip om een vaste definitie van te maken. Daarom wordt er in deze paper gekeken en vergeleken naar een bestaand onderzoek over moraliteitsmetriek in het gevangenendilemma. Deze resultaten worden naast de resultaten gelegd van een Snowdrift toernooi die ook dezelfde moraliteitsmetriek toepast. De resultaten van dit onderzoek wezen uit dat de gebruikte moraliteitsmetriek dezelfde rangorde in morele waarde produceerde. De absolute waarden van de vergeleken moraliteitswaarden bleek echter verschillend. In het Snowdrift spel werden hogere waarden aangetroffen dan in het gevangenendilemma. Dit zorgt ervoor dat moraliteit sterker tot uiting komt in het Snowdrift spel. In dit onderzoek wordt doorverwezen naar meerdere manieren om vergelijkbare onderzoeken voort te zetten. Zo kunnen er nog meer 2x2 spellen gebruikt worden, of er kan geprobeerd worden om een formele definitie voor de moraliteit van botsystemen te geven met de resultaten van dit onderzoek.

Inhoudsopgave

1 Introductie

Moraliteit in de echte wereld is altijd al een moeilijk te definiëren begrip geweest. Dit komt onder andere doordat de definitie van een moreel persoon of morele handeling subjectieve kanten heeft. Daarom is het zeer lastig om een vast begrip of harde definitie te geven van moraliteit. Dit probleem vloeit ook door naar de informatica en kunstmatige intelligentie. De kwestie om moreel gedrag te definiëren is al complex genoeg, maar om dat gedrag dan te simuleren in een agent of botsysteem is nog een stap complexer.

Echter zijn er wel wat voorbeelden in de praktijk die een vleugje van moraliteit in de informatica laten zien. Zo is er bijvoorbeeld de zoekmachine Google. Deze zoekmachine gebruikt een onderliggend concept, dat sterkt lijkt op moraliteit, om de beste zoekresultaten voort te brengen. Dit concept is dat Google een hogere waardering geeft aan pagina's die zelf veel doorverwijzingen geven naar andere pagina's. Zo worden als het waren die pagina's beloond voor hun morele gedrag. Dit is een noodzakelijk concept aangezien de pagina's met veel doorverwijzingen zichzelf eigenlijk zwak opstellen door de concurrerende pagina's een handje te helpen. (Rogers, Ian. 2002)

De theorist Scott Aaronnson heeft in zijn blog een definitie gegeven voor een volgens hem moreel persoon of agent. "Een moreel persoon is iemand die samenwerkt met een ander moreel persoon, en die weigert samen te werken met een immoreel persoon.". Deze definitie lijkt sterk op het beschreven concept van Google. Tyler Singer-Clark (2014) heeft deze definitie concreet gemaakt in eigen onderzoek, waarin hij verschillende bots het gevangenendilemma tegen elkaar liet spelen. Na afloop geeft hij twee verschillende morele maatstaven voor het gedrag van de bots, die ons een beter inzicht moeten geven in de moraliteit van bots.

Deze meegegeven maatstaven zijn al een solide begin van het onderzoek naar moraliteit van bots. Echter kunnen er nog wel een paar punten worden aangevuld. Zo zijn er nog tal van bots die Singer-Clark niet gebruikt, en er zouden naar meer 2x2 spellen gekeken kunnen worden om zo een beter beeld te verkrijgen van de morele metriek. Door deze mogelijke opties tot vervolgonderzoek luidt mijn onderzoeksvraag als volgt:

Hoe verhouden de moraliteitsmetrieken, die worden toegepast in het geïtereerde gevangenendilemma, zoals beschreven door Tyler Singer-Clark en Scott Aaronson, zich tot de moraliteitsmetrieken die we toepassen op een ander 2x2 spel, zoals het Snowdrift spel, en is het hiervoor noodzakelijk dat de definities van de moraliteitsmetrieken aangepast moeten worden om zodanig een goed beeld van moraliteit te krijgen in het Snowdrift spel?

Onderzoek naar een ander 2x2 spel en het gedrag van de agents in dat spel zou ons een beter en globaler idee kunnen geven van moraliteit in een simpele

2x2 omgeving, en dus ook voor simpele Kunstmatige Intelligentie systemen. Ook zouden we hier eventueel fouten of aanpassingen in de definitie van de gegeven morele maatstaven kunnen aantonen en verbeteren. Dit alles zou kunnen leiden naar een beter begrip van de moraliteit binnenin agents, en hoe we die later dan weer kunnen toepassen op grotere Kunstmatige Intelligentie systemen binnen de informatica.

In dit onderzoek gaan we het gedrag van morele agenten bestuderen in een andere omgeving dan die van het gevangenendilemma. We kijken naar het Snowdrift 2x2 spel. Omdat het Snowdrift spel een anti-coördinatie spel is en het gevangenendilemma juist niet, zouden er verschillen kunnen bestaan tussen het morele gedrag van de verschillende agents tussen beide spellen.

De tekst begint met uitleg over de 2x2 spellen en wat de essentiële verschillen zijn tussen deze twee spellen. Hierna zullen we de al bestaande maatstaven van Singer-Clark bestuderen en uitleggen zodat er een goede basis aanwezig is voor het onderzoek. Hierna beschrijven we in de methode een manier om de bestaande moraliteitsmetriek om te zetten naar het Snowdrift spel. Na deze uitleg worden de bots geïntroduceerd en toegelicht. Na de introductie van de bots volgt het experiment. In het experimentsectie is beschreven hoe het uitgevoerde experiment tot stand is gekomen en met welke parameters gewerkt wordt. Na de uitleg van het experiment worden de resultaten getoond. Bij deze resultaten wordt een korte toelichting gegeven. We eindigen de tekst met de discussie en de conclusie. In de discussie lichten we alle merkwaardige resultaten toe met behulp van het gedrag van de bots. Ook worden de resultaten vergeleken met de resultaten van Singer-Clark (2014). We lichten de resultaten toe per moraliteitsmetriek. Zo is er een sectie voor de eigenjesusmetriek en eigenmosesmetriek. Na de discussie is er nog de conclusie. In de conclusie vatten we het gehele onderzoek samen en geven we voorzetten om nieuw onderzoek mogelijk te maken in dit gebied.

2 Moraliteitsmetriek

2.1 Geïtereerde Gevangenendilemma

Voor dit onderzoek worden resultaten bekeken en vergeleken uit een geïtereerd gevangenendilemma toernooi. Het gevangenendilemma is een 2x2 spel waarbij twee spelers moeten kiezen voor de optie C of D. De optie C staat voor bekennen en de optie D staat voor niet bekennen. In verband met moraliteit worden deze optie's samenwerken (C) en niet samenwerken (D) genoemd. Het gevangenendilemma is een 2x2 sociaal dilemma spel waarbij twee spelers een hoge scoren kunnen behalen door allebei te kiezen voor niet samenwerken. Dit is het Nash evenwicht van het gevangenendilemma. Als nu één speler zou kiezen voor C in de plaats van D, dan zal de speler die zelf voor C gekozen heeft een lagere score krijgen dan de speler die voor D blijft kiezen. Daardoor is het Nash evenwicht

van het gevangenendilemma allebei voor D kiezen.

	C	D
C	3,3	0,5
D	5,0	1,1

Figuur 1: Schematische weergave van het gevangenendilemma

De hoogste score voor beide spelers kan gehaald worden door beide voor C te kiezen. Maar als dit gebeurt, hoeft één speler maar te wisselen van keuze om zichzelf de hoogste score te geven en de tegenstander een lagere score. Zoals te zien in is Figuur 1. Het voorvoegsel geïtereerde betekent dat de spelers dit spel meerdere keren achter elkaar spelen. De spelers kunnen hun eigen keuzes na elke zet wijzigen en het is mogelijk om de vorige zetten van hun tegenstander te onthouden, zodat er een wijziging van de strategie kan worden toegepast.

Singer-Clark (2014) gebruikt het geïtereerde gevangenendilemma om spelende bots een moraliteitmetriek mee te geven. Singer-Clark hecht hierbij waarde aan de zet waarin beide spelers kiezen voor C ofwel samenwerken. In de volgende sectie wordt gekeken naar het Snowdrift spel. Dit spel zal in dit onderzoek de rol overnemen van het gevangenendilemma.

2.2 Snowdrift Spel

Het Snowdrift spel is een 2x2 spel dat op het eerste oog erg lijkt op het gevangenendilemma. Dit spel kan worden gezien als een scenario waarin twee auto's op elkaar afrijden. Nu hebben beide auto's de keuze om uit te wijken (C), of door te rijden (D). Als er twee keer voor optie D wordt gekozen zullen de spelers punten verliezen. In plaats van dat spelers een voordeel krijgen uit het kiezen van dezelfde strategie, zal in het Snowdrift spel dezelfde strategie een nadeel opleveren qua score, en daarom wordt het Snowdrift spel beschouwd als een anti-coördinatie spel.

	C	D
C	0,0	−1,1
D	1,−1	-10,-10

Figuur 2: Schematische weergave van het Snowdrift spel

Vanuit een oogpunt gericht op moraliteit is dit spel een interessante ontwikkeling vanuit het gevangenendilemma. In het Snowdrift spel zullen acties die

resulteren in het beide kiezen van D bestraft worden. Ook zullen spelers die kiezen voor zichzelf door altijd D te kiezen, bestraft worden in het evalueren van de moraliteit.

2.3 Eigenjesus en Eigenmoses

De morele maatstaven die in dit onderzoek worden bekeken zijn de eigenjesuswaarden en eigenmoseswaarden. Deze waarden zijn als volgt gedefinieerd door (Singer-Clark 2014). De eigenjesus- en eigenmoseswaarden komen voort uit een idee dat een matrix kan worden opgesteld uit de interactie tussen verschillende bots in een 2x2 spel. De oorspronkelijke eigenjesus en eigenmoses definities komen uit het gevangenendilemma. Singer-Clark probeert een matrix op te stellen waarin voor elke cel terug te zien is wat de 'goedaardigheid' is van Bot A naar Bot B. In het gevangenendilemma wordt Bot A moreel of samenwerkend beschouwd als Bot A kiest voor samenwerken als Bot B dat ook doet. Als Bot A dus altijd samenwerkt als die tegen een bepaalde Bot B speelt, dan zal de cel in de matrix een hoge 'goedaardigheid' meegeven van Bot A ten opzichte van Bot B. Als Bot B ook voor samenwerken kiest als Bot A dat doet, dan zal de 'goedaardigheid' van Bot B richting Bot A ook oplopen.

De coöperatiegraad is dan te berekenen door naar het percentage te kijken dat Bot A heeft meegewerkt als die tegen Bot B speelt. Dit is voor alle bots te berekenen om zo de totale matrix te krijgen met cellen (i, j) die de coöperatiegraad van Bot i naar Bot j laten zien. Deze matrix wordt voor het gevangenendilemma beschreven als de coöperatiematrix. Voor het gevangenendilemma zijn er twee soorten coöperatiematrix nodig. Er zijn immers twee verschillende moraliteitsmetrieken die gedefinieerd worden.

Om de eigenjesuswaarde van een morele bot te vinden hebben we de coöperatiematrix C nodig. Beschouw een toernooi met een n aantal bots, dan is de coöperatiematrix $C \in [0, 1]^{n \times n}$, met $C = (c_{ij})$. c_{ij} staat voor de coöperatiegraad van de interactie tussen Bot i en Bot j. We definiëren nu de eigenvector van de coöperatiematrix C als $\vec{v} = [v_1, ..., v_n]$. De eigenvector is een vector die niet veranderd als de dimensies van de oorspronkelijke matrix veranderen. De eigenvector wordt hierdoor ook gebruikt omdat er zo meerdere interacties tussen meerdere bots kunnen plaatsvinden zonder dat de definitie van de moraliteitsmetriek moeten worden aangepast. Nu is de eigenjesuswaarde van bijvoorbeeld een bot b_i gelijk aan v_i van de eigenvector.

De eigenmoseswaarde van een morele bot wordt bijna op een identieke manier gevonden. Er wordt echter een andere matrix gebruikt om de eigenvector en dus ook de eigenmoseswaarde te vinden. Bij de eigenmoses waarde wordt er gebruik gemaakt van een coöperatiematrix D. Deze coöperatiematrix D gaat ook uit van de coöperatiegraad, net zoals de coöperatiematrix C, alleen de manier waarop de cellen in matrix D zijn gedefinieerd is anders. Coöperatiematrix C gebruikt voor de cellen de waardes van c_{ij}. Dit is de coöperatiegraad van de

interactie tussen Bot i en Bot j. Om coöperatiematrix D op te stellen moeten we de waarden van de cellen als volgt aanpassen. We krijgen nieuwe cellen d_{ij} met $d_{ij} = 2(c_{ij} - 0,5)$. Deze nieuwe cellen zorgen ervoor dat coöperatiematrix D er als volgt uitziet. $D \in [-1,1]^{n \times n}$. De eigenvector van deze matrix is de vector $\vec{v} = [v_1, ..., v_n]$. De eigenmoseswaarde van een bot b_i is gelijk aan v_i in deze eigenvector.

Het verschil tussen deze twee coöperatiewaarden in de cellen van de twee verschillende matrices geeft een groot verschil aan binnenin het begrip van moraliteit. Dit is ook te zien in de resultaten van Singer-Clark. De coöperatiegraad kan in coöperatiematrix C niet onder nul gaan, hierdoor zal de uiteindelijke eigenjesuswaarde ook niet onder de nul komen. Dit heeft als resultaat dat er niet een bepaalde straf of sanctie bestaat op het besluit om niet mee te werken met je tegenstander. Een bot die er bijvoorbeeld voor kiest om niet mee te werken, krijgt slechts een lagere eigenjesuswaarde op het eind. Bekeken vanuit het gevangenendilemma betekent dit dat een bot altijd een goed moreel besluit maakt als die besluit mee te werken, dus om zichzelf als bot in een zwakke rol te zetten. Want er is een kans aanwezig dat de tegenstander ook zal kiezen voor samenwerken, wat resulteert in een goede score en goede coöperatiegraad. Maar er is ook een kans aanwezig dat de tegenstander niet kiest voor samenwerken, waardoor de meewerkende bot geen opbrengst krijgt. Dit heeft ook als gevolg dat de bot wordt beloond door mee te werken met een bot die tegen alle andere bots immoreel doet. Dus de eigenjesus komt voort uit een zachtere aanpak die de eigen positie van een immorele bot versterkt, terwijl er altijd nog een veilig optie is om mee te werken.

De eigenmoses gaat echter uit van een mogelijk negatieve score in de coöperatiematrix D. Dit heeft als resultaat dat de eigenmoseswaarde onder de nul kan belanden. Met deze definitie wordt het meewerken met immorele bots niet beloond, maar juist gestraft. In het gevangenendilemma betekent dit dat er geen veilige optie is die de bot altijd wel een hogere eigenmoseswaarde zal opleveren, want samenwerken met een bepaalde bot die per definitie alles probeert tegen te werken, zal voor de meespelende bot een negatieve eigenmoseswaarde opleveren.

Deze twee verschillende metrieken werpen dus beide een andere blik op moraliteit. Eigenjesuswaarde gaat uit van goedheid en het idee dat meewerken altijd loont, ook als degene waarmee moet worden samengewerkt een immorele bot is. Daarentegen geeft de eigenmoseswaarde een harder beeld van het werkelijke gedrag van de bots, en coöperatie met per definitie immorele bots wordt afgestraft.

3 Snowdrift toernooi met moraliteitsmetriek

De eigenjesus en eigenmoses metrieken zijn bedacht vanuit het idee van het gevangenendilemma. Dit spel heeft een focus op het samenwerken en dit is ook terug te zien in de manier waarop de eigenjesus en eigenmoses worden berekend, namelijk uit de coöperatiematrices. Omdat er gekeken wordt naar de eigenjesus- en eigenmoseswaarden in een ander 2x2 spel, namelijk het Snowdrift spel, moeten enkele elementen van de twee metrieken opnieuw gedefinieerd worden om zo een correcte weergave te maken van de moraliteitmetriek in het Snowdrift spel.

3.1 Methode

Er wordt gekeken naar de coöperatiematrix van het gevangenendilemma tegenover de coöperatiematrix van het Snowdrift spel. Voor het Snowdrift spel wordt de weergave van Figuur 2 gebruikt. Als Figuur 2 wordt vergeleken met Figuur 1, is er te zien dat optie 'C' kiezen een lagere opbrengst zal geven in het Snowdrift spel dan wanneer de bots optie 'C' kiezen in Figuur 1. Intuïtief zorgt dit ervoor dat er mogelijk al een groot verschil tussen de eigenjesuswaarde en eigenmoseswaarde ontstaat als de opbrengsten van het Snowdrift spel gekopieerd worden naar de opbrengsten van het gevangenendilemma. Want het Snowdrift spel is immers gericht op anti-coördinatie. Dit is echter niet zo. De berekening van de metrieken en de coöperatiematrix let niet op de absolute opbrengsten van het gespeelde 2x2 spel. De eigenjesus- en eigenmoseswaarde letten alleen op de coöperatiewaarde die in de matrices worden ingevoerd. Het verschil is alleen dat de totale scores van deze twee spellen verschillend zal uitkomen.

Er wordt gekeken naar een manier waarop de coöperatiematrix zodanig wordt aangepast dat als resultaat de eigenjesus- en eigenmoseswaarde een verschil vertonen. Omdat het gevangenendilemma een spel gericht op samenwerking is, en het Snowdrift spel een spel op anti-coördinatie, wordt er een nieuwe matrix gedefinieerd die een inzicht geeft van de mate waarop een speler voor eigenbelang kiest of zichzelf juist opoffert.

De coöperatiematrix van het gevangenendilemma wordt berekend door het aantal voorkomens van wederzijdse samenwerking te tellen, en daarvan het percentage te berekenen over het totaal aantal ontmoetingen. We definiëren de matrix van het Snowdrift spel echter als het percentage van de gevallen waarin Bot A kiest voor samenwerken in het Snowdrift spel, en waarin Bot B kiest voor het niet samenwerken. Deze nieuwe coöperatiewaarde is te omschrijven als een anti-coördinatiewaarde. Deze waarde laat zien hoe de morele bot zich opoffert door samenwerken te kiezen, als de andere bot niet samenwerken heeft gekozen. Dit is noodzakelijk in het Snowdrift spel want als Bot A ook zou kiezen voor niet samenwerken, dan is de straf voor de beide spelers hoger dan in het geval van het gevangenendilemma. Dus voor Bot A is er een zekere noodzakelijke aardigheid om samenwerken te kiezen.

Zo ontstaat er een compleet nieuw samenspel van de bots in het Snowdrift spel, en door dit nieuwe samenspel kan er gekeken worden naar nieuwe waarden van de moraliteitsmetriek.

Om dit nieuwe samenspel van de bots op te zetten en te analyseren wordt er gebruik gemaakt van een geïtereerd Snowdrift toernooi. In dit toernooi spelen verschillende algoritmen het Snowdrift spel tegen elkaar. Hieronder volgt een lijst van alle algoritmen die meedoen in het geïtereerde Snowdrift toernooi.

ALL_C en ALL_D

Deze twee algoritmen zijn van de meest simpele vorm en er wordt niet gelet op het gedrag van de tegenstander. ALL_C is een algoritme dat altijd zal samenwerken. ALL_D is het complement van ALL_C en zal daarom altijd voor niet samenwerken kiezen. Deze algoritmen zijn makkelijk uit te buiten door de andere algoritmen die wel letten op het gedrag van hun medespeler. Zo is de optimale strategie tegen de ALL_C namelijk altijd kiezen voor niet samenwerken.

CHAMPION en EATHERLY

CHAMPION en EATHERLY zijn één van de meest complexe bots in het toernooi. EATHERLY houdt het aantal keer dat de tegenstander niet samenwerkte in de gaten. Als de tegenstander dan een keer niet samenwerkt, bekijkt EATHERLY het percentage dat de tegenstander niet heeft samengewerkt, en EATHERLY kiest met dat percentage of hijzelf ook niet samenwerkt. CHAMPION werkt hetzelfde, alleen in het begin van de ronde kiest CHAMPION altijd een paar zetten voor samenwerken, gevolgd door een paar zetten gedrag kopiëren, en vervolgens doet CHAMPION hetzelfde als EATHERLY, alleen als de tegenstander meer dan 60% van de tijd kiest voor samenwerken, zal zelfs CHAMPION samenwerken kiezen als de tegenstander een keer niet samenwerkt.

FRIEDMAN

Het FRIEDMAN algoritme is een algoritme dat berust op het principe van een kill-switch. Dit algoritme zal als eerste zet altijd kiezen voor samenwerking. Hierna blijft het FRIEDMAN algoritme kiezen voor samenwerking, mits de tegenstander ook blijft kiezen voor samenwerking. Als de tegenstander van FRIEDMAN namelijk maar één keer kiest voor niet samenwerken, zal FRIEDMAN nooit meer kiezen voor samenwerken. Op dat moment wordt de kill-switch van dit algoritme ingeschakeld. FRIEDMAN kan hierdoor erg divers gedrag vertonen. Als de tegenstander van FRIEDMAN bijvoorbeeld de eerdergenoemde ALL_C is, dan zal FRIEDMAN precies hetzelfde gedrag vertonen als ALL_C. Maar als ALL_D bijvoorbeeld de tegenstander wordt, dan zal FRIEDMAN dus maar één keer samenwerken, en verder niet.

GENEROUS_TIT_FOR_TAT

GENEROUS_TIT_FOR_TAT is een algoritme dat altijd begint met het kiezen voor samenwerken in de eerste zet. Na deze eerste zet wordt het algoritme een kopieer algoritme. De tweede zet van GENEROUS_TIT_FOR_TAT is identiek aan de vorige zet van de tegenstander. Echter heeft GENEROUS_TIT_FOR_TAT een unieke eigenschap die ervoor zorgt dat niet samenwerken van de tegenstander niet altijd wordt gekopieerd. Een kans tussen de 0 en 1 wordt meegegeven aan GENEROUS_TIT_FOR_TAT. Deze *p_generous* geeft de kans aan dat GENEROUS_TIT_FOR_TAT toch zal kiezen voor samenwerken als de tegenstander dat niet heeft gedaan in de vorige zet. In dit onderzoek wordt de *p_generous* van GENEROUS_TIT_FOR_TAT weergegeven als een getal achter de naam van het algoritme. Zo wordt er in dit onderzoek gebruikt gemaakt van de kansen 0,1 en 0,3.

JOSS

Het JOSS algoritme is in principe het complement van het eerder genoemde GENEROUS_TIT_FOR_TAT algoritme. JOSS zal in de eerste zet ook altijd kiezen voor samenwerken, maar daarna zal JOSS met een meegegeven kans kiezen voor niet samenwerken als de tegenstander heeft gekozen voor wel samenwerken. Hiermee is dus te zien dat JOSS het omgekeerde is van GENEROUS_TIT_FOR_TAT. JOSS zal altijd het gedrag van de tegenstander kopiëren als de tegenstander kiest voor niet samenwerken, maar als de tegenstander kiest voor wel samenwerken, dan zal JOSS met een kans *p_sneaky* toch kiezen voor niet samenwerken. Deze *p_sneaky* is een kans met de waarde tussen de 0 en 1. In dit onderzoek gebruiken we de kansen 0,1 en 0,3.

MAJORITY

MAJORITY is een algoritme dat ook altijd zal beginnen met samenwerken. Deze eerste zet van samenwerken wordt dan doorgezet, mits de tegenstander vaker heeft gekozen voor samenwerken dan niet samenwerken. Deze regel wordt bij elke zet in het spel opnieuw nagekeken, dus MAJORITY kan in een spel blijven veranderen van strategie. De bepaalde strategie ligt compleet in de handen van de tegenstander. MAJORITY heeft twee versies van het algoritme die worden gebruikt in dit onderzoek. Een HARD en SOFT versie. Het verschil tussen MAJORITY_HARD en MAJORITY_SOFT komt alleen naar voren als de tegenstander een gelijk aantal keer voor samenwerken en niet samenwerken heeft gekozen. MAJORITY_HARD zal bij een gelijk aantal kiezen voor niet samenwerken, en MAJORITY_SOFT zal kiezen voor wel samenwerken. Dit kleine verschil kan in veel situaties erg belangrijk zijn. Zo kan een tegenstander van MAJORITY in de eerste beurt samenwerken kiezen, en in de tweede beurt niet samenwerken. MAJORITY_HARD zou dan voor de volgende beurt kiezen voor niet samenwerken. Dit kan MAJORITY_HARD bij de tegenstander in een slecht daglicht zetten.

PAVLOV

PAVLOV is een algoritme dat berust op het principe dat een winnende strategie nooit veranderd moet worden. PAVLOV begint altijd met kiezen voor samenwerken. Daarna zal PAVLOV bekijken of de zet van de tegenstander ook samenwerken is of niet. Als de zet van zowel PAVLOV en de tegenstander niet hetzelfde is, zal PAVLOV kiezen voor niet samenwerken. Als de twee spelers wel dezelfde zet hebben gedaan, zal PAVLOV kiezen voor samenwerken.

RANDOM

Het RANDOM algoritme is een algoritme dat niet kijkt naar de tegenstander. Het RANDOM algoritme zal met een meegegeven kans $p\_cooperate$ kiezen voor samenwerking. In dit onderzoek worden er drie versies van RANDOM gebruikt met coöperatiekansen van 0,2; 0,5 en 0,8.

SUSPICIOUS_TIT_FOR_TAT

SUSPICIOUS_TIT_FOR_TAT is een kopieer algoritme net zoals het eerdergenoemde GENEROUS_TIT_FOR_TAT. Echter bevat SUSPICIOUS_TIT_FOR_TAT geen element van kans in de zetkeuze. SUSPICIOUS_TIT_FOR_TAT kiest als eerste zet altijd voor niet samenwerken. Hierna zal SUSPICIOUS_TIT_FOR_TAT het gedrag van de tegenstander kopiëren.

TESTER

TESTER is een algoritme dat altijd begint met niet samenwerken. Hierna kijkt TESTER wat de tegenstander doet. Als de tegenstander samenwerkt zal TESTER dat ook doen voor de volgende twee zetten. Hierna wisselt TESTER tussen wel en niet samenwerken. Als de tegenstander echter niet samenwerkt na de eerste zet, dan zal TESTER een vriendelijk gebaar maken en de volgende beurt wel samenwerken, en vervolgens het gedrag van de tegenstander kopiëren. Op het eerste gezicht lijkt TESTER een bot met goede bedoelingen. Maar de eerste zet van TESTER zorgt ervoor dat het goede gedrag van dit algoritme bijna niet tot uiting komt. Door een eerste zet van niet samenwerken zullen veel tegenstanders ook niet samenwerken, en hierdoor zal al snel een situatie ontstaan waarin niet samenwerken voor beide spelers de meest gespeelde zet is.

TIT_FOR_TAT

TIT_FOR_TAT is het algoritme dat de meeste variaties heeft in dit onderzoek. Zelf is TIT_FOR_TAT het basis kopieer algoritme zonder meegegeven kansen of andere eigenschappen. TIT_FOR_TAT zal in de eerste zet kiezen voor samenwerken, en daarna zal TIT_FOR_TAT de vorige zet van de tegenstander kopiëren en zelf spelen. TIT_FOR_TAT heeft twee variaties die worden gebruikt in dit onderzoek. Deze variaties zijn TIT_FOR_TWO_TATS en TWO_TITS_FOR_TATS.

TIT_FOR_TWO_TATS is een algoritme dat afgeleid is uit TIT_FOR_TAT. Het verschil tussen deze twee algoritmen is dat TIT_FOR_TWO_TATS pas kiest voor niet samenwerken als de tegenstander twee maal achter elkaar heeft gekozen voor niet samenwerken.

TWO_TITS_FOR_TATS vertoont een minder morele versie van dit gedrag dan de eerdergenoemde algoritmen. TWO_TITS_FOR_TATS zal twee keer achter elkaar kiezen voor niet samenwerken als de tegenstander maar één keer kiest voor niet samenwerken.

In de methodesectie hierboven is gekeken naar de manier waarop we de moraliteitsmetriek in het Snowdrift spel gaan onderzoeken en bekijken. Zo is er beschreven hoe de berekening van de coöperatiewaarden tot stand komt, en hoe de coöperatiewaarde van het gevangenendilemma wordt omgezet naar de anti-coördinatiewaarde van het Snowdrift spel. Verder is er gekeken naar de middelen die nodig zijn om de anti-coördinatiewaarde van het Snowdrift spel te kunnen vinden. Dit wordt gedaan door middel van een geïtereerd Snowdrift spel met een aantal algoritmen. Deze algoritmen, met al hun variaties, zijn besproken en toegelicht. Omdat nu bekend is langs welke weg de moraliteitsmetriek in het Snowdrift spel wordt onderzocht, en ook met welke middelen, kan het experiment worden uitgevoerd.

3.2 Experiment

Om het geïtereerde Snowdrift spel te spelen, worden eerdergenoemde algoritmen tegen elkaar opgesteld. De lengtes van al deze interacties worden vastgesteld met een bepaalde kans w. Met deze kans w wordt bepaald hoelang de interacties tussen de bots duren. In dit experiment wordt de kans $w = 0,995$ gebruikt. Dus de interacties hebben per zet in het Snowdrift spel een kans van 0,995 om door te gaan naar een volgende interactie met dezelfde bot. In het opzetten van het toernooi wordt ook de parameter voor het aantal ontmoetingen meegegeven. Deze parameter geeft aan hoe vaak bots met alle andere bots spelen. In dit experiment staat deze parameter op 5. Voor elke ontmoeting in het toernooi wordt de lengte van de interacties opnieuw berekend. Zo ontstaan er ontmoetingen met mogelijk zeer uiteenlopende lengtes.

Nadat de lengtes van de interacties en het aantal ontmoetingen zijn vastgesteld, gaan de bots het geïtereerde Snowdrift spel tegen elkaar spelen. Elke bot speelt dus een vastgesteld aantal interacties tegen alle bots. Na elke zet van de bots wordt er een methode gestart die moet bepalen welke volgende zet de bot moet doen. Deze methode kijkt ook naar de opbrengsten van het huidige Snowdrift spel. Dit zijn de opbrengsten die we zien in Tabel 2. De volgende zet van de bots wordt bepaald door de lijst van gespeelde zetten in de interactie met dezelfde bot. Ook spelen de bots interacties tegen elkaar, want ze spelen immers tegen alle bots. Dus het ALL_C algoritme speelt ook een Snowdrift spel

tegen een kopie van zichzelf. In het geval van ALL_C levert dat geen interessante uitkomst op, maar sommige algoritmen vertonen zeer anders gedrag dan de kloon waartegen ze spelen.

Na de ontmoetingen en interacties worden de totale scores en de gespeelde zetten van de bots opgeslagen. Elke interactie van elk bot paar wordt opgeslagen. Deze opgeslagen waarden worden dan gebruikt om eerst de anti-coördinatiewaarde te berekenen, en om vervolgens al deze waarden in de twee anti-coördinatiematrices te stoppen. Met behulp van deze matrices worden net als bij het gevangenendilemma de eigenvectoren gevonden, en met deze eigenvectoren en met de matrices worden dan de eigenjesuswaarde en eigenmoseswaarde gevonden per bot.

Er is nu gekeken naar de manier waarop het geïtereerde Snowdrift spel wordt gespeeld. Het uitgevoerde experiment is beschreven in een chronologische volgorde van handelingen, en er is uitgelegd hoe de interactie tussen de verschillende bots tot stand komt.

4 Resultaten

De experimentsectie heeft beschreven hoe dit experiment zal gaan plaatsvinden. Het experiment is uitgevoerd met de eerdergenoemde parameters om de volgende resultaten te genereren.

De resultaten komen in Tabel 1 te staan. Voor de resultaten is er gekeken naar een geïtereerd Snowdrift toernooi met de volgende opbrengsten, $T = 1$, $R = 0$, $P = -10$, $S = -1$. De kans w dat een interactie tussen al spelende bots doorgaat wordt op 0,995 gezet. Het aantal ontmoetingen van de bots staat op 5 met de lengtes van deze ontmoetingen als volgt gegenereerd door de kans w: [111, 66, 34, 218, 555].

4.1 Eigenjesuswaarde

Tabel 2 laat de eigenjesuswaarde van de verschillende bots zien, afgeleid uit de anti-coördinatiematrix. Er is te zien dat de ALL_C bot bovenaan staat in de Tabel 2 met een waarde van 1,448. Daaronder volgen de CHAMPION en EATHERLY bots met bijna een dezelfde score als de ALL_C bot. Hierna wordt het midden van de tabel voornamelijk in beslag genomen door alle TIT_FOR_TAT bots en hun varianten. Aan de onderkant van de tabel staan de RANDOM_BOTS met alle varianten en de JOSS_BOTS. Als laatste in de tabel staat de ALL_D bot met een waarde van 0.

4.2 Eigenmoseswaarde

Tabel 3 laat een gesorteerde tabel zien van alle bots met hun respectievelijke eigenmoseswaarden. Er is te zien dat de twee MAJORITY_BOT varianten bo-

Bot	**Eigenjesus-waarde**	**Eigenmoses-waarde**
ALL_C	1,448	1,492
RANDOM_0.8	1,037	0,608
GENEROUS_TIT_FOR_TAT_0.3	1,396	1,858
EATHERLY	1,425	1,929
CHAMPION	1,433	1,904
MAJORITY_SOFT	1,366	1,999
TIT_FOR_TWO_TATS	1,407	1,981
GENEROUS_TIT_FOR_TAT_0.1	1,365	1,961
PAVLOV	1,229	1,664
MAJORITY_HARD	1,263	1,992
TIT_FOR_TAT	1,304	1,985
RANDOM_0.5	0,523	-0,453
TESTER	0,827	0,582
TWO_TITS_FOR_TAT	1,201	1,927
FRIEDMAN	1,122	1,873
SUSPICIOUS_TIT_FOR_TAT	0,751	0,362
JOSS_0.1	0,637	0,113
JOSS_0.3	0,301	-0,772
RANDOM_0.2	0,098	-1,277
ALL_D	0,0	-1,737

Tabel 1: Snowdrift toernooi moraliteit metriek resultaten

venaan de tabel staan met een score van net geen 2,0. Na de MAJORITY_BOTS volgen bijna alle TIT_FOR_TAT_BOTS en hun varianten, met bijna allemaal een score hoger dan de 1,9. ALL_C staat deze keer ergens in het midden van onze tabel, in plaats van helemaal bovenaan zoals bij de eigenjesuswaarde. Verder valt op dat de RANDOM_BOTS en de JOSS_BOTS nog in het onderste deel van de tabel staan samen met de ALL_D die weer de laagste waarde heeft. Deze keer met een waarde van -1,737.

Bot	Eigenjesus-waarde
ALL_C	1,448
CHAMPION	1,433
EATHERLY	1,425
TIT_FOR_TWO_TATS	1,407
GENEROUS_TIT_FOR_TAT_0.3	1,396
MAJORITY_SOFT	1,366
GENEROUS_TIT_FOR_TAT_0.1	1,365
TIT_FOR_TAT	1,304
MAJORITY_HARD	1,263
PAVLOV	1,229
TWO_TITS_FOR_TAT	1,201
FRIEDMAN	1,122
RANDOM_0.8	1,037
TESTER	0,827
SUSPICIOUS_TIT_FOR_TAT	0,751
JOSS_0.1	0,637
RANDOM_0.5	0,523
JOSS_0.3	0,301
RANDOM_0.2	0,098
ALL_D	0,0

Tabel 2: Lijst met bots gesorteerd op eigenjesus-waarde

5 Discussie

In de discussie gaan we de onderlingen resultaten van het Snowdrift toernooi toelichten en vergelijken. Voor deze vergelijking maken we gebruik van de data van Singer-Clark (2014).

5.1 Eigenjesuswaarden toegelicht

De resultaten uit Tabel 2 worden weer bekeken. Aangezien de anti-coördinatiematrix nadruk legt op het samenwerken ofwel afwijken van de bots in het Snowdrift

Bot	**Eigenmoses-waarde**
MAJORITY_SOFT	1,999
MAJORITY_HARD	1,992
TIT_FOR_TAT	1,985
TIT_FOR_TWO_TATS	1,981
GENEROUS_TIT_FOR_TAT_0.1	1,961
EATHERLY	1,929
TWO_TITS_FOR_TAT	1,927
CHAMPION	1,904
FRIEDMAN	1,873
GENEROUS_TIT_FOR_TAT_0.3	1,858
PAVLOV	1,664
ALL_C	1,492
RANDOM_0.8	0,608
TESTER	0,582
SUSPICIOUS_TIT_FOR_TAT	0,362
JOSS_0.1	0,113
RANDOM_0.5	-0,453
JOSS_0.3	-0,772
RANDOM_0.2	-1,277
ALL_D	-1,737

Tabel 3: Snowdrift toernooi moraliteit metriek resultaten

spel, is er te zien dat ALL_C bovenaan de tabel staat. Hieruit volgt dan ook dat een bot die nooit kiest voor het samenwerken, een score van 0 krijgt. Deze ALL_D bot zal altijd niet samenwerken en is daarom het ideale voorbeeld van een immorele bot in het Snowdrift spel, bekeken van de eigenjesus metriek.

Opvallend is dat drie RANDOM_BOTS in de onderste helft van de tabel staan met respectievelijke waarden van 1,037, 0,523 en 0,098 voor RANDOM_0.8, RANDOM_0.5 en RANDOM_0.2. DE RANDOM_BOTS zijn bots die een kans meekrijgen om voor samenwerking te kiezen. Deze kans is het getal achter de string RANDOM. Intuïtief zou de eigenjesuswaarde van de RANDOM_0.8 dan

op ongeveer 80% van de maximale eigenjesuswaarde van ALL_C moeten liggen. Dit hoeft echter niet noodzakelijk het geval te zijn, wat in de resultaten ook blijkt. Zo kan RANDOM_0.8 per toeval hebben samengewerkt met bots die altijd niet samenwerken zoals ALL_D. Deze niet samenwerkende bots voegen dan weinig tot niets toe aan de totale eigenjesuswaarde van RANDOM_0.8. Als dit het geval zou zijn, dan zal de waarde van RANDOM_0.8 al snel minder zijn dan 80% van ALL_C. De lage score van de RANDOM_BOTS zou ook nog een andere oorzaak kunnen hebben. Aangezien RANDOM_0.8 ongeveer 20% van de tijd niet samenwerkt, kan RANDOM_0.8 per toeval lage punten hebben binnengehaald als de tegenspeler wel altijd samenwerkt. Deze gegevens zorgen ervoor dat de RANDOM_BOTS kunnen verschillen van plek op de ranglijst voor elk uitgevoerde Snowdrift toernooi. Er is echter nog een andere bot aan de onderkant van onze tabel die iets minder willekeurigheid vertoont.

De JOSS_BOTS en de SUSPICIOUS_TIT_FOR_TAT bot bevinden zich ook in het onderste gedeelte van de tabel, met een redelijk marge tot de rest boven hen. De plek van deze bots is iets minder willekeurig dan de plek van de RANDOM_BOTS. SUSPICIOUS_TIT_FOR_TAT staat het hoogst van deze drie bots. SUSPICIOUS_TIT_FOR_TAT is een bot die altijd begint met niet samenwerken. Daarna kopieert deze bot het gedrag van de vorige zet van de tegenstander. SUSPICIOUS_TIT_FOR_TAT staat laag in de tabel door de eerste zet van niet samenwerken. Dit kan er bij sommige tegenstanders voor zorgen dat er altijd gekozen wordt voor niet samenwerken, en hierdoor zal de eigenjesuswaarde van SUSPICIOUS_TIT_FOR_TAT weinig oplopen. Dit wordt bevestigd door de eigenjesuswaarde van TIT_FOR_TAT. Deze bots verschillen alleen in de eerste zet, en zoals te zien is in Tabel 2 ligt de waarde van TIT_FOR_TAT ongeveer 75% hoger dan de eigenjesuswaarde van SUSPICIOUS_TIT_FOR_TAT. De enige manier om een hogere score dan de JOSS_BOTS te verkrijgen, is als de tegenstander een bot is die altijd kiest voor meewerken, zoals ALL_C.

De JOSS_BOTS staan lager dan SUSPICIOUS_TIT_FOR_TAT in de tabel. De JOSS_BOTS zijn ook bots afgeleid uit TIT_FOR_TAT. De JOSS_BOTS zijn bots die als eerste zet wél kiezen voor samenwerken. Te zien is ook dat de JOSS_0.3 bot een lagere waarde heeft dan de JOSS_0.1 bot. Dit ligt in lijn met het gedrag van de bots, want JOSS_0.1 wisselt in 10% van de gevallen van samenwerken naar niet samenwerken, en JOSS_0.3 doet dat in 30% van de gevallen. Dus JOSS_0.3 werkt vaker niet samen dan JOSS_0.1, en dat is duidelijk terug te zien.

Aan de bovenkant van de lijst in Tabel 2 staan GENEROUS_TIT_FOR_TAT_BOTS en TIT_FOR_TWO_TATS als hoogste uit de TIT_FOR_TAT serie. GENERS_TIT_FOR_TAT_BOTS zijn het complement van de JOSS_BOTS. Hierdoor is er te zien dat GENEROUS_TIT_FOR_ TAT_0.3 een hogere eigenjesuswaarde heeft dan GENEROUS_TIT_FOR_TAT_0.1. Want bij GENEROUS_TIT_FOR_TAT_0.3 is er een hogere kans om toch voor samenwerken te kiezen.

TIT_FOR_TWO_TATS staat als hoogste van alle TIT_FOR_TAT bots in Tabel 2. TIT_FOR_TWO_TATS kiest alleen voor niet samenwerken als de tegenstander twee maal achter elkaar voor niet samenwerken kiest. Dit gegeven maakt deze bot het meest moraal van alle bots uit de TIT_FOR_TAT serie.

5.2 Eigenjesuswaarden vergeleken

Nu de resultaten van de eigenjesuswaarde van het Snowdrift toernooi bekeken en bestudeerd zijn, kunnen deze resultaten naast de resultaten gelegd worden van Singer-Clark en zijn geïtereerde gevangenendilemma toernooi. Om zo te onderzoeken of er een verschil bestaat tussen de moraliteit metriek in een sociaal dilemma spel en anti-coördinatie spel.

Bot	**Eigenjesus-Waarde**	**Bot Singer-Clark**	**Eigenjesus Singer-Clark**
ALL_C	1,448	ALL_C	1,377
CHAMPION	1,433	EATHERLY	1,343
EATHERLY	1,425	CHAMPION	1,338
TIT_FOR_TWO_TATS	1,407	MAJORITY_SOFT	1,325
GENEROUS_TIT_FOR_TAT_0.3	1,396	TIT_FOR_TWO_TATS	1,325
MAJORITY_SOFT	1,366	GENEROUS_TIT_FOR_TAT_0.3	1,318
GENEROUS_TIT_FOR_TAT_0.1	1,365	GENEROUS_TIT_FOR_TAT_0.1	1,285
TIT_FOR_TAT	1,304	MAJORITY_HARD	1,229
MAJORITY_HARD	1,263	TIT_FOR_TAT	1,222
PAVLOV	1,229	PAVLOV	1,180
TWO_TITS_FOR_TAT	1,201	TWO_TITS_FOR_TAT	1,105
FRIEDMAN	1,122	RANDOM_0.8	1,103
RANDOM_0.8	1,037	FRIEDMAN	1,048
TESTER	0,827	TESTER	0,887
SUSPICIOUS_TIT_FOR_TAT	0,751	SUSPICIOUS_TIT_FOR_TAT	0,780
JOSS_0.1	0,637	JOSS_0.1	0,745
RANDOM_0.5	0,523	RANDOM_0.5	0,688
JOSS_0.3	0,301	JOSS_0.3	0,416
RANDOM_0.2	0,098	RANDOM_0.2	0,273
ALL_D	0,0	ALL_D	0,0

Tabel 4: Vergelijking eigenjesuswaarden, aangepast uit (Singer-Clark 2014)

In Tabel 4 is een vergelijking te zien tussen de eigenjesuswaarden uit een anti-coördinatie spel (links), en een coördinatie spel (rechts). De resultaten van het coördinatie spel zijn verkregen uit Singer-Clark (2014).

Boven en onderaan beide lijsten staan de twee ALL_BOTS. Deze uitslagen zijn geen verrassing. Vanaf onder bekeken zijn de twee lijsten identiek in volgorde tot de TESTER bot. TESTER bot heeft een waarde van 0,827 in het Snowdrift spel en een score van 0,887 in het gevangenendilemma. Deze waarden geven dus ongeveer dezelfde grens aan voor de onderliggende bots. Een groot verschil tussen de bots met lage waarden is dat de waarden van alle bots onder TESTER lager liggen in het Snowdrift spel dan in het gevangenendilemma. Dit komt doordat in het Snowdrift spel een hogere straf ligt op niet samenwerken dan bij het gevangenendilemma. Daarom staan bij het Snowdrift spel de bots die het vaakst niet samenwerken in een slechter daglicht dan bij het gevangenendilemma. Dit verklaart de lagere waarden.

Hoger in de lijsten zien we dat de volgorde voor het grootste deel hetzelfde is of bijna hetzelfde met twee omgedraaide bots die wel dicht bij elkaar liggen qua waarde. Het opvallendste verschil tussen deze twee lijsten zijn de hoogte van de waarden. Zo ligt de eigenjesuswaarde van de gehele top 10 van het Snowdrift spel hoger dan de waarden van de top 10 in het gevangenendilemma met dezelfde bots. Dit kan verklaard worden door weer te kijken naar de onderkant van de tabel waar dit ook het geval was, maar dan omgekeerd. Omdat kiezen voor samenwerken in het Snowdrift spel betekent dat de bot zelf het risico loopt om minpunten te krijgen, wordt er een hogere waarde gehecht aan het samenwerken dan in het gevangenendilemma.

Dus uit deze vergelijking kan afgeleid worden dat, wat betreft de eigenjesuswaarden, er in volgorde van de bots weinig verschil aanwezig is. Maar dat het grootste verschil ligt in het feit dat de totale eigenjesuswaarden hoger liggen bij het anti-coördinatie spel. Dit komt doordat er een groter risico is in het niet samenwerken bij het Snowdrift spel dan bij het gevangenendilemma, en omdat kiezen voor samenwerken afgestraft kan worden in het Snowdrift spel met minpunten.

5.3 Eigenmoseswaarden toegelicht

In Tabel 3 zien we dat er verschillen zijn in zowel de rangorde van deze lijst, als de waarden van de eigenmoses, vergeleken met de waarden van eigenjesus. Eigenmoses wordt gezien als de extremere versie van de eigenjesus, want immoreel gedrag en samenwerken met iemand die immoreel verdrag vertoont wordt afgestraft. Dit is terug te zien in de aanwezige min waarden die de onderste vier bots van Tabel 3 hebben. De laag scorende bots in de lijsten van eigenjesus en eigenmoses zijn identiek en liggen in dezelfde volgorde.

Als er nu verder gekeken wordt naar de bovenkant van Tabel 3 komt er al

snel een aparte bot in het vizier. De ALL_C bot staat ongeveer in het midden van de tabel met een eigenmoseswaarden van 1,492. Deze waarde lijkt hoog. Deze waarde ligt immers hoger dan de waarde van de eigenjesus ALL_C, alleen zijn er nu 11 bots die een hogere eigenmoses score hebben. ALL_C bevindt zich in het midden van de tabel omdat ALL_C altijd kiest voor samenwerken. ALL_C lijkt op het eerste gezicht een perfect morele bot, maar dat is niet waar. ALL_C kiest ook voor samenwerken met bots die nooit kiezen voor samenwerken, bijvoorbeeld ALL_D. Omdat we naar de eigenmoseswaarden kijken, levert dat dus minpunten op in de totale eigenmoseswaarde van ALL_C. Het is niet zo dat samenwerken met een immorele bot meteen ervoor zorgt dat ALL_C in de min overblijft, alleen zorgt het ervoor dat bots die beter nadenken en misschien ook wel moreler zijn hoger eindigen.

Verder omhoog in Tabel 3 zien we wat bekende middenmoters zoals PAVLOV en FRIEDMAN, maar ook GENEROUS_TIT_FOR_TAT_0.3. Waar deze bot bij eigenjesuswaarden nog op de vijfde plek stond, staat die nu op de tiende. In het geval van deze bot werkt de goedheid van toch samenwerken als de tegenstander niet samenwerken kiest averechts. Als GENEROUS_TIT_FOR_TAT_0.3 tegen een bot speelt die niet samenwerkt, is het volgens de eigenmoses geoorloofd om ook te kiezen voor niet samenwerken. Maar GENEROUS_TIT_FOR_TAT_0.3 zal dan toch in 30% van de gevallen kiezen voor wel samenwerken. Hierdoor kunnen punten verloren worden, net zoals bij ALL_C.

Na GENEROUS_TIT_FOR_TAT_0.3 staan ook de EATHERLY en CHAMPION en GENEROUS_TIT_FOR_TAT_0.1 op een andere plek vergeleken met Tabel 2. GENEROUS_TIT_FOR_TAT_0.1 heeft minder van die goedheid die GENEROUS_TIT_FOR_TAT_0.3 wel heeft, en staat daarom hoger op de lijst. CHAMPION EN EATHERLY zijn wat gezakt ten opzichte van andere bots.

De twee bots die in Tabel 3 de hoogste eigenmoseswaarde hebben zijn de twee MAJORITY_BOTS. De MAJORITY_BOTS beginnen altijd met samenwerken, en vervolgens zullen ze dat ook blijven doen als de tegenstander vaker heeft samengewerkt dan niet. Dit betekent dat de MAJORITY_BOTS niet per kans kunnen samenwerken met een bot die nooit met andere samenwerkt. Dit maakt de MAJORITY_BOTS tot meest morele bots in de ogen van de eigenmoseswaarden. Wat ook in lijn ligt met de definitie van de eigenmoses. MAJORITY bant immorele bots uit met niet samenwerken, en morele bots worden door de MAJORITY_BOTS beloond met wel samenwerken. Hierdoor stonden deze twee bots ook niet bovenaan bij de eigenjesus. Soms wilde de MAJORITY_BOTS niet samenwerken, en dat betekent dat ze geen punten krijgen of verliezen bij de eigenjesus.

5.4 Eigenmoseswaarden vergeleken

Tabel 5 vergelijkt de eigenmoseswaarden verkregen uit het Snowdrift spel (links), met de eigenmoseswaarden uit het gevangenendilemma toernooi van Singer-

Bot	**Eigenmoses-Waarde**	**Bot Singer-Clark**	**Eigenmoses Singer-Clark**
MAJORITY_SOFT	1,999	TIT_FOR _TWO _TATS	1.831
MAJORITY_HARD	1,992	CHAMPION	1.798
TIT_FOR_TAT	1,985	MAJORITY_SOFT	1.793
TIT_FOR_TWO_TATS	1,981	EATHERLY	1.788
GENEROUS_TIT_FOR_TAT_0.1	1,961	GENEROUS_TIT_FOR_TAT_0.1	1.769
EATHERLY	1,929	TIT_FOR_TAT	1.747
TWO_TITS_FOR_TAT	1,927	MAJORITY_HARD	1.728
CHAMPION	1,904	GENEROUS_TIT_FOR_TAT_0.3	1.705
FRIEDMAN	1,873	TWO_TITS_FOR_TAT	1.593
GENEROUS_TIT_FOR_TAT_0.3	1,858	FRIEDMAN	1.521
PAVLOV	1,664	ALL_C	1.481
ALL_C	1,492	PAVLOV	1.470
RANDOM_0.8	0,608	RANDOM_0.8	0.888
TESTER	0,582	TESTER	0.768
SUSPICIOUS_TIT_FOR_TAT	0,362	SUSPICIOUS_TIT_FOR_TAT	0.506
JOSS_0.1	0,113	JOSS_0.1	0.443
RANDOM_0.5	-0,453	RANDOM_0.5	-0.009
JOSS_0.3	-0,772	JOSS_0.3	-0.448
RANDOM_0.2	-1,277	RANDOM_0.2	-0.897
ALL_D	-1,737	ALL_D	-1.481

Tabel 5: Vergelijking eigenmoseswaarden, aangepast uit (Singer-Clark 2014)

Clark (rechts). Te zien is dat de onderkant van Tabel 5 weer net zoals bij Tabel 4 identiek is qua volgorde van de bots. Opvallend is dat er weer een groot verschil is tussen de eigenmoseswaarden van de identieke bots in beide lijsten. Deze waarneming is met dezelfde rede te verklaren als bij Tabel 4. In het Snowdrift spel wordt zwaarder gestraft voor beide niet samenwerken, en daar staat ook een hogere negatieve eigenmoseswaarden tegenover.

Pas bij de bot ALL_C zijn de eigenmoseswaarden gelijk. Vanaf ALL_C en hoger in beide lijsten zien we al snel dat de eigenmoseswaarden hoger dan 1,9 liggen bij het Snowdrift spel, en hoger dan 1,7 zijn bij het gevangenendilemma. De rangordes die de beide lijsten vertonen zijn deze keer wel verschillend. Echter

is wel te zien dat dezelfde globale structuur van volgorde nog hetzelfde is in de lijsten die worden vergeleken. De bots die bij de beide lijsten in de top staan, staan niet precies op dezelfde plek, maar wel dicht in de buurt van elkaar. Dit is toe te rekenen aan de kleine onderlingen verschillen in beide lijsten. Omdat er een element van kans in de uitslagen van de bots zit, zouden de resultaten van elk Snowdrift toernooi tot op deze graad anders uit kunnen komen. Het is echter belangrijker dat de onderlinge waarden van de bijvoorbeeld RANDOM_BOTS nog wel kloppen.

Het meest opvallende is weer dat de totale eigenmoseswaarden hoger liggen bij het Snowdrift spel. Dit is net zoals bij de onderkant van de tabel het geval door de hogere beloning die wordt toegekend aan kiezen voor samenwerken in het Snowdrift spel.

De algemene waarneming die voortkomt uit de vergelijking van zowel de eigenjesus en eigenmoses is dat de onderlinge volgorde van de scorende bots gering verschilt. Het grote verschil tussen beide toernooien is de totale score van de bots. In het Snowdrift toernooi lagen de waarden dan wel hoger dan wel lager als die van het gevangenendilemma toernooi. Dit is toe te schrijven aan de risico's en mate van zelf opoffering in het Snowdrift spel. Door deze uitwijking van de hogere eigenmoseswaarde en eigenjesuswaarde is er een nieuw feit af te leiden wat betreft het gedrag van de bots. Omdat de negatieve waarden van de eigenmoses lager uitvallen dan die in het gevangenendilemma, is er af te leiden dat de immorele eigenschappen van bots nog meer geaccentueerd worden. De RANDOM_0.5 bot had in het gevangenendilemma een eigenmoseswaarde van -0,009 wat bijna gelijk is aan 0. Hierdoor zou er nog een twijfel kunnen ontstaan over het feit dat RANDOM_0.5 helemaal niet immoreel is. Maar door te kijken naar de eigenmoseswaarden van RANDOM_0.5 in het Snowdrift spel zien we dat deze bot wel degelijk als immoreel kan worden beschouwd in de ogen van de eigenmoses.

6 Conclusie

In dit onderzoek is er gezocht naar een verband tussen moraliteits metrieken in het Snowdrift spel en het gevangenendilemma. Er werd gebruik gemaakt van een toernooi spel om zo de resultaten te verkrijgen. Deze resultaten zijn in dit onderzoek vergeleken met al bestaande resultaten van Singer-Clark (2014). Hieruit bleek dat de morele eigenschappen van de gebruikte bots versterkt werden in het Snowdrift spel. Het Snowdrift spel schetst een nieuwe situatie voor de gebruikte bots om zo hun moraliteit te kunnen onderzoeken, en door de vergelijking met de resultaten van een ander toernooi spel kunnen er stappen gemaakt worden om zo een formele manier te vinden om moraliteit van bepaalde bots of systemen te kunnen omschrijven.

Zoals hierboven ook al gezegd wordt, is er in de toekomst nog tal van onderzoek

mogelijk over dit onderwerp. Zo kan er bijvoorbeeld voortgang worden geboekt in een formele omschrijving van de moraliteit van een bot door gebruik te maken van de eigenmoses- en eigenjesuswaarden. Deze twee metrieken zouden bijvoorbeeld samengebracht kunnen worden tot één uiteindelijke metriek die meer kan vertellen over het morele gedrag van een bot.

Ook mogelijk is de toevoeging van een nieuwe metriek. De bestaande metrieken geven al een goed beeld van moraliteit in een bot. Maar gedrag in de echte wereld kan altijd nog buiten deze twee metrieken vallen. Zo is het gedrag van een minderheid in een samenleving niet goed beschreven door deze metrieken. Als een minderheid alleen samenwerkt met hun eigen volk en verder tegen iedereen immoreel gedrag vertoont omdat ze onderdrukt worden, is het dan werkelijk immoreel om als buitenstaander toch samen te werken met deze groep. Volgens de huidige metrieken wel, en daar zou een nieuw metriek antwoord op kunnen geven.

Een nieuw 2x2 spel spelen met nieuwe bots kan ook meer inzicht geven in het gebruik van de moraliteitsmetrieken. Met dit vervolg onderzoek zouden we nog een stap dichterbij kunnen komen om zo een formele definitie te vinden om simpele bots een graad van moraliteit mee te geven.

Hierboven is al even kort het idee genoemd om een formele definitie te vinden om het gedrag van bots moreel te kunnen omschrijven. Dit is ook het uiteindelijk doel van dit onderzoek. Als deze metrieken worden doorgevoerd naar meer 2x2 spellen of grotere spellen, kunnen we uiteindelijk in simpele KI systemen dit soort metriek toevoegen. Met zo een metriek in een KI systeem is er mogelijk een manier om een bepaalde input te beoordelen of het moreel gedrag is of niet. Dit kan helpen om bepaalde beslissingen af te wegen tegen elkaar om zo te kijken welke keuze moreler is dan de andere. De ogen zijn gericht op het uiteindelijke doel om het morele gedrag van de mens zo goed mogelijk weer te geven in een KI systeem die ons in de toekomst helpt voortgang te boeken in het begrijpen van moreel gedrag in de mens en natuur.

7 Appendix

Software

De gebruikte software voor dit onderzoek is te vinden en te downloaden met de volgende link: `https://github.com/BasKnuppe/Scriptie-Snowdrift-Toernooi`. Deze code is een aangepaste versie van de code van Singer-Clark (2014). Deze originele versie is te vinden op `https://github.com/tscizzle/ IPD_Morality`.

Deze software is gemaakt om toernooien te laten spelen van bepaalde 2x2 spellen met verschillende bots. De scores van deze botspelers worden vervolgens opgeslagen en geanalyseerd. Met deze opgeslagen data worden de moraliteitsmetrieken van de bots weergegeven in een tabel.

Gebruik

Om deze code te runnen is er beschikking nodig over Python versie 3.6. Nadat de code gedownload is, moeten alle Python bestanden ingeladen worden. Om vervolgens de code te runnen moet de klasse *arena.py* worden gestart. *arena.py* geeft een tabel terug met de totale scores van alle bots, en de eigenjesuswaarde en eigenmoseswaarde.

In *the_bots.py* klasse staan alle verschillende bots beschreven met hun meegekregen opbrengst-matrix. Ook wordt de kans w meegegeven aan de bots om zo de lengte van de interacties te bepalen. de beste manier om het berekenen van de moraliteitmetriek aan te passen is door wijzigingen te maken in de $calculate\_cooperation\_stuff()$ methode. Hier kan het aanmaken van de coöperatiematrix worden aangepast zodat de berekeningen van de eigenjesus en eigenmoseswaarden anders verloopt.

Referenties

[1] Singer-Clark, T. *Morality Metrics On Iterated Prisoner's Dilemma Players*, Juni 2014

[2] Guyer, M., Hamburger, H. *A note on the enumeration of all 2 x 2 games*, General Systems, 13, 205-8, 1968

[3] Kilgour, D. M. ,Fraser, N. M. *A taxonomy of all ordinal 2 x 2 games*, Theory and Decision, 24(2), 99-117, 1988

[4] Rogers, Ian. *The Google Pagerank Algorithm and How It Works*, `http://www.sirgroane.net/google-page-rank/`. 16 May 2002.

[5] Singer-Clark T. *Code voor het runnen van IPD toernooien*
`https://github.com/tscizzle/ IPD_Morality`

[6] Github Repository *Aangepaste code van Singer-Clark voor het Snowdrift toernooi*
`https://github.com/BasKnuppe/Scriptie-Snowdrift-Toernooi`