# Fairness in multi-agent systems

# Summary

Within the field of artificial intelligence, the research area of multi-agent systems investigates societies of autonomous entities, called *agents*, that need to cooperate or compete in order to achieve a certain goal (Jennings et al., 1998; Shoham et al., 2007). Humans may be part of these societies. Example applications include resource distribution, auctions, and load balancing. In many of these applications, we observe elements of so-called *social dilemmas*, in which taking into account fairness and social welfare is necessary. In some dilemmas, humans are known to care strongly for fairness and social welfare; in others, caring for fairness and social welfare is essential for agents to achieve a satisfactory solution.

In this thesis, we show how agents may be stimulated to care for the fairness of their actions. The human-inspired mechanisms of altruistic punishment and withholding action are central to our approach. Chapter 1 therefore presents the following problem statement.

**PS** *How can we obtain human-inspired fairness in multi-agent systems?*

The remainder of Chapter 1 provides an overview of five research questions resulting from this problem statement. These questions are given below.

**RQ1** *How are humans using fairness in their decisions?*

**RQ2** *What are the foundations of human-inspired computational fairness?*

**RQ3** *How can human-inspired fairness be modeled computationally, taking into account the foundations of human fairness?*

**RQ4** *What are the analytical properties of the computational human-inspired fairness models developed in this research?*

**RQ5** *What are the (empirical) benefits of incorporating explicitly human-inspired fairness in adaptive multi-agent systems?*

Thereafter, Chapter 1 discusses the research methodology followed. Next, in Chapter 2, we review the fundamental background knowledge required for research in multi-agent systems in general, i.e., game theory and multi-agent reinforcement learning. Chapter 2 also elaborately explains the social dilemmas we investigate throughout the thesis, i.e., the *Ultimatum Game*, the *Nash Bargaining Game*, and the *Public Goods Game*.

The five research questions are addressed in Chapters 3 to 6. RQ2 is addressed before the other four research questions, i.e., in Chapter 3. Essentially, we there present two foundations, i.e., (1) a set of three *requirements* that need to be met by human-inspired computational fairness models, and (2) a *template model* based on these requirements. We require that any computational model should be (R1) rooted in a game-theoretic background (as game theory is a well-established manner of describing interactions between multiple

agents), (R2) computationally applicable in an adaptive multi-agent system (i.e., we require tractable solution concepts), as well as (R3) inspired by humans. With respect to the last requirement, we state that agents must be able to answer three questions, i.e., (R3-Q1) to what extent an interaction is fair, (R3-Q2) whether one or more of their peers need(s) to be punished, and (R3-Q3) whether it is desirable to withhold action, i.e., not to participate in a certain interaction. We present a template model, based on the well-known concept of a utility function, and show how this model may be instantiated in such a way that our requirements are met.

The following three chapters, i.e., Chapters 4 to 6, form the core of the thesis, and follow a similar structure. In each chapter, we discuss a specific computational model of human-inspired fairness, based on the foundations presented in Chapter 3. For each model, we address RQ1 by discussing a specific descriptive model of human fairness. We then create a computational model of fairness, incorporating this specific descriptive model (RQ3), analyze the computational model (RQ4), and use the model in an adaptive multi-agent system that is learning to find good solutions to social dilemmas (RQ5).

Chapter 4 presents a computational model based on a descriptive model of *inequity aversion*, as developed by Fehr and Schmidt (1999). Inequity aversion entails that human decisions are influenced by differences in observed rewards. The descriptive model of Fehr and Schmidt (1999) is able to explain a great deal of (irrational) human decisions in interactions where limited resources need to be shared. Even though this is the case, the model has not yet convincingly found its way into multi-agent systems. We address this issue by developing a computational model of fairness, based on inequity aversion. We show that our computational model allows (a small number of) agents to reach satisfactory, human-inspired solutions to the social dilemmas under study.

In Chapter 5, we discuss that human behavior is not only influenced by (differences in) observed rewards, but also by additional information humans may have or gather about the others participating in an interaction. Existing research proposes reputation-based approaches to model this phenomenon. We argue that an important element is missing from reputation-based models, i.e., that there may be additional information that is immediately present, e.g., bargaining powers, stereotypes, or priorities. We present a descriptive model named *priority awareness* to address both additional information that is immediately present, as well as reputation. In an approach similar to that of Chapter 4, we show how a computational model of fairness may be based on priority awareness, and how this influences outcomes to interactions between agents.

In Chapter 6, we increase the scale of our work from at most a few dozen to a few thousand agents. In the last ten years, a great deal of research has been devoted to *social networks*, which have been shown to have a decisive impact on the manner in which humans (as well as artificial agents) change their behavior as a result of interactions with others, based on neighbor relations in a certain network structure. Existing work examining how networked agents change their behavior in social-dilemma-like interactions has thus far been limited to social dilemmas with only a discrete, small number of possible actions (e.g., two). Since

we are interested in addressing more realistic social dilemmas, we investigate how network structure influences agents' behavior in social dilemmas with a continuum of actions. We show that a number of mechanisms promoting desirable behavior in discrete dilemmas also work in continuous dilemmas (i.e., most prominently the possibility of agents to withhold action by breaking the link between them and an undesirable neighbor), while a number of other mechanisms do not provide additional benefits (e.g., reputation).

We conclude in Chapter 7 by answering our research questions, summarizing our findings, answering the problem statement, and looking at opportunities for future work. We show that human-inspired fairness in multi-agent systems may be obtained by means of a four-step process, i.e., (1) experiments with human subjects, (2) modeling human fairness in descriptive models, (3) establishing the foundations of computational models of human-inspired fairness, and (4) translating descriptive models to computational models respecting the foundations. Using the three computational models presented in this thesis, we may conclude that agents are able to find desirable solutions to social dilemmas, even in the presence of agents that do not care about fairness and try to undermine a desirable, fair solution.

# Samenvatting

Binnen de kunstmatige intelligentie bevindt zich een onderzoeksgebied genaamd 'multi-agent systemen'. In dit gebied wordt onderzoek gedaan naar collectieven van autonome entiteiten, die men *agenten* noemt. Deze agenten moeten samenwerken of met elkaar wedijveren om een bepaald doel te bereiken (Jennings et al., 1998; Shoham et al., 2007). Het is mogelijk dat mensen deelnemen aan zulke collectieven. Voorbeeld-toepassingen zijn onder meer gelegen in het verdelen van hulpbronnen en beloningen. In veel toepassingen vinden we elementen terug van de zogenoemde *sociale dilemma's*. In dit soort dilemma's moet overwogen worden om eerlijkheid en sociaal welbevinden in acht te nemen. In sommige gevallen is dit wenselijk omdat mensen hier zeer veel belang aan hechten; in andere gevallen is het strikt noodzakelijk, omdat er anders geen bevredigende oplossing wordt gevonden.

In dit proefschrift tonen we hoe agenten gestimuleerd kunnen worden om te geven om de eerlijkheid van hun acties. De mechanismen van altruïstisch bestraffen en afzien van actie, die geïnspireerd zijn door mensen, nemen een centrale plaats in. Hoofdstuk 1 presenteert een probleemstelling als volgt.

**PS** *Hoe kan op mensen geïnspireerde eerlijkheid verkregen worden in multi-agent systemen?*

De rest van Hoofdstuk 1 geeft een overzicht van vijf onderzoeksvragen die voortvloeien uit deze probleemstelling. Deze vragen worden hieronder weergegeven.[1]

**RQ1** *Hoe gebruiken mensen eerlijkheid in hun beslissingen?*

**RQ2** *Wat zijn de fundamenten van door mensen geïnspireerde, computationele eerlijkheid?*

**RQ3** *Hoe kan door mensen geïnspireerde eerlijkheid computationeel gemodelleerd worden, met inachtname van de fundamenten van menselijke eerlijkheid?*

**RQ4** *Wat zijn de theoretische eigenschappen van modellen van door mensen geïnspireerde, computationele eerlijkheid?*

**RQ5** *Wat zijn de (empirische) voordelen van het toevoegen van expliciet door mensen geïnspireerde eerlijkheid aan multi-agent systemen?*

Daarna vervolgt Hoofdstuk 1 met een bespreking van onze onderzoeksmethodologie. Vervolgens bespreken we in Hoofdstuk 2 de fundamentele achtergrondkennis die vereist is voor vrijwel elk onderzoek gerelateerd aan multi-agent systemen, te weten speltheorie en 'multi-agent reinforcement learning'. Hoofdstuk 2 legt bovendien uitgebreid de sociale dilemma's uit die we in dit proefschrift steeds gebruiken, namelijk de zogenoemde *Ultimatum Game*, de *Nash Bargaining Game*, en de *Public Goods Game*.

De vijf onderzoeksvragen worden beantwoord in de Hoofdstukken 3 tot en met 6. RQ2 wordt beantwoord in Hoofdstuk 3, voor de andere vier onderzoeksvragen aan bod komen.

[1] We merken op dat de afkorting 'RQ' is afgeleid van de Engelstalige term 'research question'.

In feite bekijken we twee fundamenten, namelijk (1) drie *vereisten* waaraan moet worden voldaan door modellen van computationele, door mensen geïnspireerde eerlijkheid, en (2) een *sjabloonmodel* dat wordt gebaseerd op die drie vereisten. Om precies te zijn vereisen we dat ieder computationeel model (R1) geworteld moet zijn in een speltheoretische achtergrond (omdat speltheorie een goed gefundeerde manier is om interacties tussen meerdere agenten te beschrijven), (R2) computationeel toepasbaar moet zijn in een adaptief multi-agent systeem (dat wil zeggen, de modellen zijn berekenbaar), en tenslotte (R3) geïnspireerd moet zijn door mensen. De laatste vereiste werken we nog verder uit door te stellen dat agenten in staat moeten zijn om drie vragen te beantwoorden, te weten (R3-Q1) in hoeverre een interactie eerlijk is, (R3-Q2) of één of meer van de anderen bestraft dient te worden, en (R3-Q3) of het wenselijk is om af te zien van actie. We presenteren een sjabloonmodel, gebaseerd op het bekende concept van een nutsfunctie, en tonen hoe dit model geïnstantieerd kan worden op een manier die onze vereisten respecteert.

De volgende drie hoofdstukken (Hoofdstuk 4 tot en met 6) vormen de kern van dit proefschrift, en hebben allemaal een gelijkaardige structuur. In ieder van deze hoofdstukken behandelen we een specifiek computationeel model van door mensen geïnspireerde eerlijkheid, gebaseerd op de fundamenten uit Hoofdstuk 3. Voor elk model beantwoorden we RQ1 door aandacht te besteden aan een specifiek descriptief model van menselijke eerlijkheid. We maken vervolgends een computationeel model van door mensen geïnspireerde eerlijkheid dat is gebaseerd op het specifieke descriptieve model (RQ3). Dan analyseren we het computationele model (RQ4) en gebruiken we het in een adaptief multi-agent systeem dat leert om goede oplossingen te vinden voor sociale dilemma's (RQ5).

Hoofdstuk 4 behandelt een computationeel model dat is gebaseerd op een descriptief model van aversie tegen onrecht, zoals ontwikkeld door Fehr and Schmidt (1999). Het model beschrijft dat mensen beïnvloed worden door (in hun ogen) onrechtvaardige verschillen in beloning. Het descriptieve model van Fehr and Schmidt (1999) is in staat om een grote verscheidenheid aan (irrationeel) menselijk gedrag te verklaren in interacties waarbij beperkte hulpbronnen dienen te worden gedeeld. Ondanks dit feit is het model tot nog toe niet zeer overtuigend van nut geweest in multi-agent systemen. We laten zien dat het model wel degelijk nut heeft, door een computationeel model van eerlijkheid te ontwikkelen dat is gebaseerd op aversie tegen onrecht, en dit model succesvol toe te passen. Een (relatief kleine) groep agenten kan, gebruikmakend van het computationeel model, bevredigende en op mensen geïnspireerde oplossingen vinden voor de sociale dilemma's die we bestuderen.

In Hoofdstuk 5 bespreken we dat menselijk gedrag niet alleen wordt beïnvloed door (verschillen tussen) waargenomen beloning, maar ook door additionele informatie die mensen kunnen hebben of verkrijgen over de anderen die aan een interactie deelnemen. Bestaand onderzoek stelt voor om dit fenomeen te modelleren door middel van reputatie. We observeren dat er een belangrijk element mist in de bestaande modellen van reputatie, namelijk dat de relevante additionele informatie geregeld onmiddellijk beschikbaar is; we denken daarbij aan bijvoorbeeld onderhandelingspositie, stereotypes, of prioriteiten. We introduceren een descriptief model gebaseerd op prioriteiten, dat zowel onmiddellijk beschikbare

als ook door reputatie verkregen additionele informatie kan modelleren. Door middel van een aanpak die verder sterk lijkt op die van Hoofdstuk 4 laten we zien hoe we een computationeel model kunnen baseren op het idee van prioriteiten, en hoe deze prioriteiten invloed hebben op de uitkomst van interacties tussen agenten.

In Hoofdstuk 6 stappen we over van systemen met maximaal enkele tientallen agenten naar systemen met enkele duizenden agenten. De laatste tien jaar is een opmerkelijk grote hoeveelheid onderzoek gericht geweest op *sociale netwerken*, waarvan is aangetoond dat ze een grote invloed hebben op de manier waarop mensen (en ook kunstmatige agenten) hun gedrag veranderen als gevolg van interacties met anderen, en op basis van buur-relaties zoals geordend in een bepaalde netwerkstructuur. Bestaand werk dat heeft onderzocht hoe 'genetwerkte' agenten hun gedrag veranderen in sociale dilemmas, is vooralsnog beperkt geweest tot sociale dilemma's met een discreet, klein aantal mogelijke acties (bijvoorbeeld twee). Wij zijn in ons werk geïnteresseerd in meer realistische sociale dilemma's. Daarom onderzoeken we hoe netwerkstructuur het gedrag van agenten beïnvloedt als agenten interacteren in sociale dilemma's met een continuum van acties. We laten zien dat een aantal mechanismen die aantoonbaar werken om gewenst gedrag te bevorderen in discrete dilemma's, ook werken in continue dilemma's (het beste voorbeeld hier is de mogelijkheid voor agenten om af te zien van actie door de verbinding te verbreken tussen henzelf en een ongewenste buur in het netwerk). Er zijn echter ook een aantal mechanismen die geen merkbaar positief effect hebben in continue dilemma's (een voorbeeld is reputatie).

Het proefschrift wordt afgerond in Hoofdstuk 7, waarin we onze onderzoeksvragen beantwoorden, onze bevindingen samenvatten, de probleemstelling beantwoorden, en kijken naar mogelijkheden voor toekomstig werk. We laten zien dat door mensen geïnspireerde eerlijkheid in multi-agent systemen verkregen kan worden door middel van vier stappen. Ten eerste worden er experimenten met mensen uitgevoerd. Ten tweede modelleren we menselijke eerlijkheid in descriptieve modellen. Ten derde bepalen we de fundamenten van computationele modellen van door mensen geïnspireerde eerlijkheid. Ten vierde vertalen we descriptieve modellen naar computationele modellen die worden gebaseerd op de fundamenten. Gebruik makend van de drie computationele modellen in dit proefschrift, mogen we concluderen dat agenten in staat zijn om goede, gewenste oplossingen te vinden in sociale dilemma's, zelfs als er agenten aanwezig zijn die niet geïnteresseerd zijn in eerlijkheid, en die daarom proberen een gewenste, eerlijke oplossing te ondergraven.