UNIVERSITY OF
KWAZULU-NATAL

INYUVESI
YAKWAZULU-NATALI

A CO-EVOLUTIONARY APPROACH TO DATA-DRIVEN AGENT-BASED MODELLING: SIMULATING THE VIRTUAL INTERACTION APPLICATION EXPERIMENTS

Submitted in fulfilment of the requirement for the degree
Doctor of Philosophy (Psychology)

By

Kevin Chizoba Igwe

Social Psychology Discipline

School of Applied Human Sciences

College of Humanities
University of KwaZulu-Natal

Supervisor:

Prof. Kevin Durrheim (PhD)

Jan 2023

DECLARATION

I Kevin Chizoba Igwe certify that the work in this thesis entitled "A co-evolutionary approach to data-driven agent-based modelling: Simulating the Virtual Interaction APPLication experiments" has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree to any other university or institution other than the University of KwaZulu-Natal.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged.

In addition, I certify that all information sources and literature used in the thesis are cited according to the requirements of the American Psychological Association (APA), Seventh Edition (2019) reference style.

Kevin Chizoba Igwe

(Student No: 212553209)

Jan 2023

DEDICATION

To my wife, Jane; my daughter, Tobe; & my son Kene

ACKNOWLEDGEMENTS

My gratitude to God for His steadfast love.

I am grateful and thankful to my supervisor, Prof. Kevin Durrheim (PhD) for his mentorship, care and advice, which he never denied me.

I am thankful to Shameek Takoorparsadh, Larry Tooke and all VIAPPLers for their contributions. To my friends and all who have, at least once thought of my success even without my knowledge, thank you.

Lastly, I am grateful to my family, especially my wife, daughter and son – Jane, Tobechi and Kenechi – for their cheerful heart.

ABSTRACT

The *dynamics* of social interactions are barely captured by the traditional methods of research in social psychology, vis-à-vis, interviews, surveyed data and experiments. To capture the dynamics of social interactions, researchers adopt computer-mediated experiments and agent-based simulations (ABSs). These methods have been efficiently applied to game theories.

While strategic games such as the prisoner's dilemma and GO have optimal outcomes, interactive social exchanges can have obscure and multiple conflicting objectives (fairness, selfishness, group bias) whose relative importance evolves in interaction. Discovering and understanding the mechanisms underlying these objectives become even more difficult when there is little or no information about the interacting individual(s). This study describes this as an *information-scarce* interactive social exchange context. This study, therefore, forms part of a larger initiative on developing efficient simulations of social interaction in an information-scarce interactive social exchange context.

First, this dissertation develops a context for and justifies the importance of simulation in an information-scarce interactive social exchange context (Chapter 2). It then performs a literature review of the studies that have developed a computational model and simulation in this context (Chapter 3). Next, the dissertation develops a co-evolutionary data-driven model and simulates exchange behaviour in an information-scarce context (Chapter 4). To benchmark the data-driven model, this dissertation develops a rule-based model. Furthermore, it creates agents that use the rule-based model, integrates them into Virtual Interaction APPLication (VIAPPL) and tests their usefulness in predicting and influencing exchange decisions. Precisely, it measures the agent's ability in reducing in-group bias during interaction in an information-scarce context (Chapter 5). Likewise, it creates machine learning (adaptive) agents that use the data-drivel model, and tests them in a similar experimental context. These chapters were written independently; thus, their objectives, methods and results are discussed in each chapter. Finally, the study presents a general conclusion (Chapter 6).

ACRONYMS

AAT	Aspiration Adaptation Theory
ABM	Agent-Based Model
ABS	Agent-Based Simulation
AI	Artificial Intelligence
ANN	Artificial Neural Network
BGR	Bounded Generalised Reciprocity
DL	Deep Learning
ESEP	Essential Social Exchange Predicates
GPS	Global Positioning System
HMM	Hidden Markov Model
KIDS	Keep it Descriptive, Stupid
KISS	Keep it Simple, Stupid
ML	Machine Learning
PIF	Pay-it-Forward
RR	Rewarding Reciprocity
SET	Social Exchange Theory
SIT	Social Identity Theory
UG	Unilateral Giving
VIAPPL	Virtual Interaction APPLication

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

CHAPTER 1. GENERAL INTRODUCTION

"It's not where you take things from - it's where you take them to."
– Jean-Luc Godard

1.1 The Purpose of the Dissertation

Despite the dynamics of social interaction being a pivotal research interest in social psychology, the traditional research methods – experiments, survey data and interviews, barely capture the essence of the processes and evolving outcomes (Mann, 2016; Vallacher et al., 2015). In light of this problem, researchers adopt computer-mediated experiments and agent-based simulations to study the dynamics of social interaction. These methods have been successfully applied to strategic social interaction games.

Whereas strategic games such as the Prisoner's Dilemma and GO (or Weiqi) have optimal outcomes, interactive social exchanges can have obscure and multiple conflicting objectives (fairness, selfishness, in-group favouritism) whose relative importance evolves through interaction. These conflicting objectives in interactive social exchanges have effects on the emergent exchange behaviour as the emergent systematic conditions begin to shape motives and drive interaction. Discovering, identifying and understanding the mechanisms underlying these objectives become even more difficult when there is little or no information surrounding the interacting individual. This dissertation describes this as an *information-scarce* interactive social exchange context.

Recent technological advancements facilitate advanced approaches to investigating and understanding the dynamics of social interaction. The advancements provide opportunities for understanding and modelling dynamic interactions via adaptive agents (March, 2021).

This study, therefore, forms part of a larger initiative on developing efficient simulations of social interaction in an information-scarce interactive social exchange context. It leverages technological advancements by combining Artificial Intelligence (AI) methods and social psychology theories to model agents' decision-making in interactive social exchanges. As such, it applies the model to investigate ways of setting social norms to reduce intergroup bias.

1.2 Aims and Objectives of the Dissertation

The aim of this dissertation is informed by the fact that, in social interactions, interdependencies such as reciprocation provide the foundations of emergence (a phenomenon relative to the complexity that is originally neither present in the system design nor the rules that govern behaviour within the system, see 2.4.2), normative behaviours and social relationships. The interdependencies also control the availability of information and the quality

– the usefulness – of the available information within an exchange system. Thus, this dissertation builds on the premise that the ability to influence emergent behaviours or set norms within a social exchange context relies on the ability to understand and influence the interdependencies between individuals and groups within that context. For example, trust is instrumental in determining which company to invest in and which stock to buy in the stock market. The likelihood of investing in a company will increase or decrease depending on how much the investor trusts the company.

Against this backdrop, the aim of this dissertation is twofold: 1) to develop adaptive agents that can act in an information-scarce interactive social exchange context, and 2) to use these agents to engineer social outcomes, namely, weaken in-group favouritism – a problem that pervades society. To realise this twofold aim, the objectives discussed below must be achieved. The aim and objectives of each chapter are defined to help realise these objectives

The first objective of this dissertation (Objective 1) is to advance: (i) the understanding of interactive social exchange as a system of pure generalised exchange, (ii) the importance of studying the system in an information-scarce environment, and (iii) the role of interdependencies in the emergence of behaviours and complexity of interactive exchange systems.

These emergent behaviours can be influenced to mutually benefit the interacting individuals. However, this is a non-trivial task that requires the understanding and prediction of exchange decisions during interactive social exchanges. Thus, the second objective (Objective 2) is to develop a model for decision-making in an interactive social exchange, based on the co-evolution of two algorithms: an artificial neural network (ANN) – a biologically inspired algorithm, and a hidden Markov model (HMM) – a statistical modelling tool.

The model must be able to predict exchange decisions with an acceptable level of accuracy. Although the term 'acceptable level' is subjective, the model must be evaluated using established evaluation metrics to determine its accuracy. Thus, the third objective (Objective 3) is to evaluate the model on the data collected from Visual Interaction Application (VIAPPL) experiments, where Visual Interaction Application – a computer-mediated environment – has been customised to represent an information-scarce environment.

While vetting a model provides useful insight into the model's performance, there is a need to investigate the claim that emergent behaviours can be influenced by manipulating the interdependencies that control the emergence. In light of this need, the fourth objective (Objective 4) is to develop and integrate agents that can use the model (developed to realise Objective 2) to act in an interactive social exchange environment.

Can the agents (in Objective 4) successfully influence interdependencies and set norms that are beneficial to groups and individuals during interactive social exchanges? To answer this question, the last objective (Objective 5) is to conduct experiments comprised of humans and artificial agents, in order to understand and influence human exchange decisions toward reducing in-group favouritism during interactive social exchanges.

This dissertation, therefore, draws from three perspectives: theory, data, and co-evolution, to contribute towards efficient simulations of interactive social exchange and its application. Each chapter is written as a stand-alone study, with its specific objectives, methods and results discussed in the chapter.

The rest of the introduction is structured as follows: The next section gives the background of the dissertation. After that, the methodological approach is presented, while the last section presents an overview of the chapters.

1.3 Background

Within society, individuals autonomously interact with one another regularly. The context and conditions make these interactions specific and unique. These interactions are driven by subjective motivations and perceptions of the individuals from their environment, which in turn, forms feedback to shape perceptions and interdependencies within the context. Thus, social interaction may be defined by the actions and reactions of individuals to other individuals and their environment.

Individuals may interact once without the possibility of interacting again. This is called one-shot interaction. Individuals involved in one-shot interaction often do not see the need to understand the future motives and objectives of others but behave opportunistically/selfishly. One-shot interactive social exchange is often demonstrated using two-player public goods games. A game can be defined by a setting in which two or more players can compete against one another by choosing some strategies which in turn affect the actions of the other players. Examples of such games include the Dictator game (Bardsley, 2008; Larney et al., 2019) and the Prisoners' Dilemma (Capraro et al., 2014; Frank et al., 1993).

However, most social and psychological phenomena occur as a result of repeated interaction between homogeneous and/or heterogeneous individuals and their environment over time (Smith & Conrey, 2007). Although individuals often make decisions in isolation, these decisions are the response to the feedback from their environments and those with whom they have previously interacted. The feedback-response chains and interdependencies between these individuals are the defining elements of emergence – an unintended evolving outcome – at a higher level. In fact, interacting individuals may be affected by network interactions in complex ways, not anticipated by the individuals, and not readily described in any rule (Macy & Flache, 1995).

Emergence is a paradox of life that can occur in interpersonal and/or intergroup relationships. For instance, consider individuals of different races and colours living in a given geographical area. After several years, it is often noticed that each individual tends to relocate closer to other individuals of the same or similar race and colour. Thus, segregation occurs even when nobody individually prefers segregation (see Schelling's 1971 seminal experimental work on segregation). Such emergent phenomena make social interactions very complex and difficult to study because most cannot be described formally: it is almost always the case that there is no set of equations that can be formulated to describe a social interaction (Pavard & Dugdale, 2002).

As with segregation, other studies show that intergroup bias often emerges during social interactions, and specially in contexts of intergroup competition (Balliet et al., 2014; Tajfel & Turner, 1979). Studies and investigations surrounding intergroup discrimination can be characterised into two distinct categories: i) those that tend towards understanding and explaining the cause of discrimination and bias (Fibbi et al., 2021; Walker, 2019), and ii) those that focus on devising strategies to reduce discrimination and bias (Brochu et al., 2020). Some of the research and experiments focusing on the latter have applied modelling and agent-based techniques to evaluate different strategies for reducing bias. This has led to the increase in the use of artificial agents as confederates (Collins et al., 2016; Krafft et al., 2017) in psychology experiments, which has resulted in the need for creating agents that can interact with humans adeptly.

The need to create agents that can interact adeptly with humans has made agents' decision-making very important, not only in psychology but also in many other fields of study (Bourgin et al., 2019; Fast & Schroeder, 2020; Zheng et al., 2020). For example, social researchers aim to understand and predict human behaviour by observing the influence of artificial agents on individuals' behaviour during interactions with others (Fast & Schroeder, 2020), thus taking the view of social interactions as complex systems. The study of social interactions as complex systems provides a means of capturing non-linear structures that are formed as social processes evolve and understanding how interaction at a lower level, such as an exchange of values between two individuals in a society (micro-interactions), leads to emergent behaviours at a system level – the society (macro-level).

Humans interact in a complex way – via different means such as exchange of goods, making and receiving calls and checking emails – and by virtue of these interactions, they create a dynamic interaction network (Phelps, 2013). Ostrom (1988) has recognised that these networks often emerge in a bottom-up fashion and postulates that they can be studied using computer simulations such as agent-based models. With the recent upsurge in computing power and technology, researchers not only use agent-based models for observations of the consequences of theoretical assumptions (Smith & Conrey, 2007), but have started combining simulation and artificial intelligence methods (Lamperti et al., 2018), which includes deep learning (van der Hoog, 2017), to build more powerful models of complex systems.

However, models that use methods such as deep learning require the collection of a huge quantity of data about the subject. Such a quantity of data is difficult to collect in most psychology domains and therefore rarely available. When they are available, the available data are so noisy that expert knowledge and domain theories are required to make sense of the data – this dissertation refers to such domains as information-scarce environments. Developing adaptive agents that can act in an information-scarce interactive social exchange context presents a unique challenge and opportunity for which the studies in this dissertation are conducted. For example, adaptive agents may be used as confederates for the investigation of manners by which norms for in-group favouritism may be weakened.

1.4 In-group Favouritism

The question of whether or not people prioritise helping in-group members over out-group members has received much attention (for example, Balliet et al., 2014; Dovidio, Gaertner, et al., 2017; Durrheim et al., 2016; Fu et al., 2012; Stürmer & Siem, 2017). This so-called in-group favouritism remains an ongoing debate among researchers interested in intergroup studies (Balliet et al., 2014). While most of the studies (Abbink & Harris, 2019; Durrheim et al., 2016; Fu et al., 2012) on in-group favouritism show that people are more likely to help in-group members over out-group members, some previous studies (for example, Wispe & Freshley, 1971) have provided evidence that people provide help equally to both in-group and out-group members. A few have suggested that people sometimes help out-group members instead of in-group members (Dovidio & Gaertner, 1981).

According to Balliet et al. (2014), despite theoretical foundations stipulating that individuals are more willing to incur costs to benefit in-group members when compared to out-group members, there is much inconsistency in support of this view. Through meta-analysis, Balliet et al. (2014) found that co-operative decision-making is more prominent in the in-group than in the out-group. Furthermore, it was found that there remains a slight bias for in-group favouritism, through categorisation, despite no mutual interdependence between group members (e.g. Dictator games) (Balliet et al., 2014). The bias for in-group favouritism is increased in the presence of interdependence (e.g. social dilemmas) (Balliet et al., 2014).

In-group favouritism is present more prominently in exchange systems where there is common knowledge of group membership, and it is cemented during simultaneous exchanges (as opposed to sequential exchanges) (Balliet et al., 2014). However, in-group favouritism is not without consequences, as it often results in intergroup discrimination (Balliet et al., 2014).

While in-group favouritism is common, its implementation is dynamic and flexible (Fu et al., 2012). Abbink and Harris (2019) show that in-group favouritism leads to discrimination against the out-group when there are situationally primed threats present in the exchange system. Abbink and Harris (2019) show evidence of a strong presence of out-group discrimination in their experiments, which is largely absent from previous research (such as Balliet et al., 2014). However, this may be due to their group sizes being around 20 individuals. These large numbers favour the derogation of the out-group as opposed to instances where the recipient is only one person (eg. standard Dictator, Public Goods and Prisoner's Dilemma games) (Abbink & Harris, 2019).

One of the ways in which a positive self-concept is developed and maintained is in-group favouritism. Individuals identify with a specific group, as their interests align with the interests of the group. Zuo et al. (2018) found that in the context of no internal competition, individual behaviours served to promote in-group favouritism. Greater instances of in-group favouritism were correlated with higher feelings of identification (Zuo et al., 2018). Social identity theory (SIT) maintains the assumption that individuals more readily view their group in a positive light and often expect positive mutual interdependence among in-group members (Zuo et al., 2018). Despite this, when exposed to an intragroup competitive system, in-group members find it significantly more difficult to benefit from the group – identification and motivation become contradictory which leads to the collapse of in-group favouritism (Zuo et al., 2018).

According to Stangor et al. (2022), in-group favouritism occurs for a plethora of reasons. In its most simplistic form, it occurs as a result of social categorisation, as the presence of an in-group and out-group aids in environmental structure simplification. Subsequently, individuals who possess a strong desire for the simplification of their environment show higher levels of in-group favouritism (Stangor et al., 2022). Furthermore, such simplifications enable in-group favouritism to manifest as a result of group membership, especially in instances when there are clear differentiations between groups (Stangor et al., 2022). The familiarity and security associated with the in-group enable the sustainment and betterment of the individual – in a positively distinctive manner (Stangor et al., 2022). In some cases, in-group members may hold not so positive outcomes at the individual level, but still maintain positive feelings towards group membership (Stangor et al., 2022).

Even though it is a general behavioural leaning to show in-group favouritism, there are many examples of in-group favouritism not occurring prominently. One such situation is when members of the in-group are outright inferior to the out-group. Members of the low-status group often display less in-group favouritism when compared to members of the high-status group (Stangor et al., 2022). Another situation where in-group favouritism is not prominent occurs when in-group members have been evaluated negatively. In other words, when an individual’s group behaves in a manner that threatens the positive image of the group, they are effectively devalued (Stangor et al., 2022).

Chae et al. (2022) successfully predicted that in-group favouritism occurs more prominently when limited resources affect the well-being of fellow in-group members. As such, Chae et al. (2022) posit that in-group favouritism plays an instrumental role in human behaviour and interaction. However, in-group favouritism creates fundamental problems for intergroup relations (Chae et al., 2022). These problems are capable of escalating intergroup competition. While in-group favouritism increases care towards in-group members and increases the survival odds of the group, it simultaneously increases the chances of intergroup conflict (Chae et al., 2022). The latter is evidence of the overarching need to reduce or weaken in-group favouritism, a task which this dissertation means to achieve.

1.5 Methodological Approach

The work conducted in this dissertation is interdisciplinary as it cuts across computer science and social psychology. This has methodological implications. First, the dissertation will employ a computer science methodology, namely, a proof by demonstration defined by Johnson (2006) and recently implemented by Yiu (Yiu, 2021). This involves three iterative phases: developing a system, testing the performance, and iteratively refining the system. The iterative refinement stops if the desired result is obtained or the result is not improving further with changes in the system parameters.

Second, experiments will be conducted in the Virtual Interaction APPLication (VIAPPL, see www.viappl.org) platform. Visual Interaction Application is a computer-mediated experimental platform designed to study social interaction and group processes (Durrheim et al., 2016) Participants are represented as avatars and are referred to as players. Participants interact in a controlled environment over several rounds by performing social exchanges, such as the exchange of tokens that represent money. Visual Interaction Application incorporates mechanisms to record interactions, which can be downloaded and analysed. It allows variables such as group size, group status and the number of groups to be manipulated. It also allows various experimental designs, such as pure generalised exchange and direct exchange, to be created and provides various means to allocate participants into groups. Thus, Visual Interaction Application is a fully customisable platform. For example, the initial number of tokens assigned to each player can be customised to vary between groups or among individuals. The number of players per group can be varied. Also, rather than a circular shape, nodes may be in another shape such as a square or triangle. They may be located at any position on the screen. Also, groups may be distinguished by the shape rather than the colour of their nodes

An information-scarce environment will be created by employing the minimal group paradigm settings (Diehl, 1990), where group identity is the only information available to players before a game. Participants sit and interact via Visual Interaction Application installed on computers connected over a network. Hence, Visual Interaction Application does not allow face-to-face interaction. Details of the Visual Interaction Application environment specific to each study will be provided in the chapter that discusses the study.

1.6 Overview of the Chapters

This dissertation sets to achieve its aim and objectives in four chapters – two review chapters, an implementation chapter and an empirical chapter – while Chapter 1 and Chapter 6 present the introduction and conclusion of the dissertation, respectively.

1.6.1 Chapter 2

Considering the first objective of this dissertation, Chapter 2 takes the perspectives of social exchange to advance the body of knowledge in social interaction. It identifies emergence and information scarcity as key factors contributing to the complexities of interactive social exchange. Emergence is controlled by interdependencies between individuals and between groups. Furthermore, information-scarcity results from complex and obscure motives, not clearly shown by the individuals during interactions. Thus, the chapter identifies the context of behaviour, emergence, and obscure and complex motives as key features for understanding social exchanges.

Additionally, the chapter posits that interdependencies such as reciprocity control emergence and are the key to influencing social outcomes. It supports the idea that norms and behaviours are influenced by these interdependencies. Lastly, the chapter provides suggestions and recommendations on how these features can be beneficial to psychologists, and also provides insight into their benefits for simulations of interactive social exchanges.

1.6.2 Chapter 3

Taking the suggestions provided in Chapter 2, Chapter 3 explores the literature on simulations of interactive social exchanges. The chapter argues that interactive social exchanges are stochastic and, thus, simulating decision-making within the exchange environment requires the integration of methods capable of: (i) discovering motives of the interacting individuals, (ii) incorporating emergent behaviours into the model, and (iii) incorporating the context of behaviour. The chapter suggests that the integration of methods inspired by machine learning and a social psychology theoretical framework will provide a better means of incorporating this triad and understanding the interdependencies underlining them.

1.6.3 Chapter 4

In line with the suggestion in Chapter 3, Chapter 4 aims to create a model that can be used by artificial agents to act in an interactive social exchange context. Chapter 4 develops a novel machine learning model – a co-evolutionary model that integrates: (i) a clustering algorithm for discovering the emergent behaviours and obscure motives during interactive social exchanges, (ii) artificial neural networks for learning and classifying exchange strategies, and (iii) a hidden Markov model for predicting exchange decisions. The model, evaluated using previous social exchange data collected in the Visual Interaction Application environment, was able to predict human exchange decisions – reciprocity, in-group or out-group allocation and allocation to high-status or low-status individuals – during interactive social exchanges.

1.6.4 Chapter 5

Chapter 5 compares the performance of agents using the co-evolutionary model with those using a rule-based model for reducing in-group favouritism. The aim was to evaluate out-group altruism for reducing in-group favouritism. Two studies – Study 1 and Study 2 – are conducted. In Study 1, agents using the rule-based model are created to act in an interactive social exchange context. The objective is to set the norm of out-group altruism. These agents are pre-programmed to exhibit out-group altruism. Similarly, (adaptive) agents in Study 2 use the model developed in Chapter 4 to predict exchange decisions. Based on this prediction, agents interact: (i) with out-group members that are predicted to be practising bounded generalised reciprocity, or (ii) with in-group members when the bounded generalised reciprocity (BGR) practice has no previous history of such practice.

Both Study 1 and 2 were conducted in the Visual Interaction Application environment. Among the findings are that agents promote in-group favouritism among the human in-group players, who resisted rather than conformed to the norm of out-group altruism, while treating agents as out-group members. Furthermore, Study 2 shows that the adaptive agents were perceived as being fairer than human, and rated more human than humans.

1.6.5 Chapter 6

Chapter 6 provides the summary of the dissertation with details of the key findings.

CHAPTER 2. DIRECT AND GENERALISED EXCHANGE IN AN INFORMATION-SCARCE INTERACTIVE SOCIAL EXCHANGE CONTEXT: CHALLENGES AND OPPORTUNITIES

Abstract

This chapter aims to identify the importance of investigating direct and generalised exchange as coexisting exchange systems. It focuses on the coexisting exchange systems with little or no information about the interacting individuals, presenting its challenges and opportunities for social scientists and computer scientists. Subsequently, it develops a context for and justifies the importance of simulation in an information-scarce interactive social exchange context.

Firstly, this study describes direct exchange and generalised exchanges, with examples of each type of exchange. Secondly, it identifies various interactions and control mechanisms underlying both these exchanges and describes how one form of exchange affects the other. This study argues that emergence and information scarcity are the two main causes of complexities in interactive social exchange. Emergence deals with the evolution of behaviours in the coexisting systems, while information scarcity results from the multiple and obscure motives not shown by the individuals during interactions. Furthermore, this study posits that emergent behaviours and norms are controlled by influencing the interdependence structures that exist within an exchange system. Finally, it presents the challenges and opportunities for studying the coexisting exchange system and puts them in three categories: context of behaviour, emergence, and obscure motives.

This study concludes by summarising the importance of studying a coexisting system of exchange to both computer scientists and social scientists. It suggests that it will increase the: (i) applicability of state-of-the-art algorithms in computer science for modelling in a social exchange context, (ii) acceptability and applicability of the social exchange models by social psychologists, and (iii) interdisciplinary research in both fields.

2.1 Introduction to Social Exchange Interaction

As defined in Chapter 1, one of the two primary aims of the dissertation was to develop adaptive agents that can act in an information-scarce interactive social exchange context. Therefore, Chapter 2 briefly reviews interactive social exchange, focusing on investigating direct and generalised exchange as coexisting exchange systems.

Many organisations are structured and managed in a way that is intended to facilitate "a positive perception of the group [organisation] and its members" (Willer et al., 2012, p. 125). This encourages supportive relationships among individuals within the organisation, as well as growth and good reputations of the organisation and the individuals. This is also true for groups and societies. Blau (1964) described such a structure as one that facilitates the exchange of resources, thus, taking the perspective of Social Exchange Theory (SET) to analyse the interactions within and between groups.

There are two forms of social exchanges: restricted (direct) and generalised (indirect) (Takahashi, 2000). In restricted or direct exchange, two individuals, say **A** and **B**, exchange resources with each other such that **A**'s reward for giving resources to **B** comes directly from **B**. Conversely, generalised exchange is a social system where three or more individuals interact unilaterally (Simpson et al., 2018). In generalised exchange, the rewards an individual receives are not directly dependent on the individual's effort, but on the system and others with whom the individual interacts (Yoshikawa et al., 2018).

Both direct and generalised exchange can occur between organisations, groups, and individuals, which Social Exchange Theory (SET) (e.g. Blau, 1964) referred to as actors. SET (e.g., Blau, 2017; Blau, 1964) hypothesised that human behaviours could be described as the exchange of resources among these actors. Such resources, which may be concrete or abstract, include but are not limited to approval, recognition, rewards (Yoshikawa et al., 2018), love, services, money, goods, status (Cook et al., 2013), and information, whose value and quality the actors seek to balance. The value of these resources is subjective, and an individual may value more than one resource at a time. Thus, the quality of social exchange is also subjective.

A typical example of direct exchange is buying of goods such that an individual, say **A**, gives money to **B** in exchange for goods after negotiating the price of the goods. This form of direct exchange is known as negotiated exchange. Thus, negotiated exchange occurs when both **A** and **B** exchange resources simultaneously after negotiation about the exchange. An alternative direct exchange is reciprocal exchange, which involves a unilateral giving where **B** rewards **A** for a favour **A** had previously done for **B**.

In reciprocal exchange, actors "perform individual acts that benefit another, such as giving assistance or advice, without negotiation and without knowing whether or when or to what extent the other will reciprocate" (Molm et al., 2007, p. 209). It is worth noting that direct exchange is dyadic. This implies that direct exchange deals with interactions between two actors, unlike generalised exchange, which deals with interactions between three or more actors.

A typical example of a generalised exchange is voting someone into power with the expectation that the political system will provide good governance. Another example is the donation of blood by an individual, where the donor gives blood to a blood bank, without any direct expectation of reward from the blood bank. Sohn and Leckenby (2007) termed such donation as group-generalised exchange. As pointed out by Bearman (1997), group-generalised exchange can be reduced to dyadic exchange between an individual and a group as a whole. Other forms of generalised exchange are chain-generalised exchange and pure-generalised exchange. In chain-generalised exchange, resources flow in one direction. Assume a system with three actors **A, B** and **C**. Chain-generalised exchange exists if **A** gives resources to **B**, **B** gives to **C,** and **C** gives to **A**. In pure-generalised exchange, "there is no fixed structure of giving; that is, **A** might give to **B** on one occasion and to **C** on a different occasion" (Molm et al., 2007, p. 208). Revisiting 'the donation of blood' in the previous example, a donor, say **A,** might donate blood to an individual **B** and later to another individual **C** instead of donating to a blood bank**.**

Both direct and generalised exchanges provide theoretical lenses through which many social issues can be explained. The synergy – rooted in the underlying mechanisms – between direct and generalised exchanges is vital in shaping interactions within and between organisations and groups. It creates a plethora of output behaviours and feelings amongst groups and individuals. Scholars such as Ekeh (1974) conceptualised direct exchange as a restricted exchange because it creates a transactional mentality (Lawler, 2001). It falls in line with a 'give-and-take structure, which results in weak solidarity within groups and between individuals (Lawler, 2001) but can produce relational cues for emotions among individuals (Lawler, 2001). Some scholars posit that direct exchange can generate more cohesion than generalised exchange (Willer et al., 2012). This supports the idea that direct exchange within interactions fosters social cohesion, which is essential (Marmarosh & Sproul, 2021) for the survival of groups and organisations.

Conversely, generalised exchange has been posited to foster greater feelings of solidarity than direct exchange (Willer et al., 2012). Note that cohesion refers to the bond between actors, whilst solidarity is stipulated to manifest when cohesion catalyses cooperative behaviours towards achieving a common goal. Direct exchange behaviours are promoted when group members seek to maintain their cohesion within the group. These behaviours are not as characteristic of an altruistic nature, nor as effective in creating an emotional impact within the in-group. By contrast, generalised exchange behaviours are promoted when solidarity is required; this solidarity is often a result of the identification of a universally experienced dilemma within the group. Due to the generalised exchange lacking any negotiation, trust is better forged through exchanges, cementing an emotional bond between actors. In addition, structures characterised by generalised exchange systems are more capable of providing group-based attributions (Lawler, 2001). For example, Parsell and Clarke (2020) provided a rationale for a generalised exchange "that enables people to have the opportunities to give back" (p. 15). The study viewed charity as an act of indirect reciprocity resulting from generalised exchange, where most people, especially the less privileged, give back to society. Kollock (1999) argued that General Public License (open source) is a manifestation of the generalised exchange concept that allows programmers to contribute modifications, with the belief that everyone will have access to them. This also gives them automatic access to the contributions of many others who contributed to the modification.

2.2 Controls and Interactions in a Social Exchange

2.2.1 The Role of Interdependencies

Behaviours within a social exchange system are primarily controlled by interdependencies that exist between individual actors, and between groups, in social exchange systems, such as reciprocation (De Dreu et al., 2020). According to De Dreu et al. (2020), there must be a focus on the interdependence structures which exist between individual actors, and between groups, because these structures, and the relationships between them, provide key insights into the dynamics of social interaction.

De Dreu et al. (2020) postulate that all groups are internally interdependent. However, these interdependencies can be: (i) independent of other groups, which enables a more mutually beneficial coexistence, and the emergence of group-specific norms and practices (such as in-group favouritism, and cohesion), (ii) positively interdependent, which enables the emergence of positive generalised reciprocity and across-group cooperation, and lastly (iii) negatively interdependent, which increases the odds of competitiveness or even conflict emerging between groups.

These interdependencies are affected by the number of actors (dyadic or multi-actor), and the possible number of times the actors may interact. Actors may interact once without the possibility of interacting again. This is termed one-shot interaction. Conversely, the actors may interact more than once. This is termed multi-round social interaction.

In a one-shot social interaction, people often do not see the need to understand the motive and objectives of others but behave opportunistically/selfishly. One-shot social interaction is often demonstrated using two-player public goods games. A game can be defined as a setting in which two or more players can compete against one another by choosing various exchange strategies that influence the actions of the other players. Examples of such games include the Dictator game (Bardsley, 2008; Larney et al., 2019) and the Prisoner's Dilemma (Capraro et al., 2014; Frank et al., 1993). Although one-shot interaction is often demonstrated by researchers interested in cooperation, it aids in reasoning about the complexities of how interdependence results in the emergence of interactional patterns introduced by multi-round (repeated) social exchange.

A multi-round social exchange requires that individuals repeatedly interact a specified number of times, thus introducing more complexity than one-shot social exchange. These complexities create considerable trouble for understanding the motives and objectives of various behaviours, as they inhibit more concrete conclusions to be drawn from data. In a multi-round social exchange, individuals' motives and objectives may be shaped by the emergent phenomena and the assumed motive of other interacting individuals. Thus, the complexity of multi-round social exchange is deepened by the actor's model of others' motives and what others think of them. A multi-round social interaction can be kept relatively simple by limiting the number of actors to two (Hula et al., 2015; Ray et al., 2009; Xiang et al., 2012).

As one may anticipate, where both the direct and generalised exchanges exist, the complexity of direct exchanges is inherently embedded in that of generalised exchange, specifically the pure-generalised exchange. Pure generalised exchange is more complex than direct exchange, because the former embodies the interdependencies that exist in both direct and generalised exchange. This study shows the coexistence of direct and generalised exchanges (see Figure

2.1). Although the structure of direct exchange and the underlying mechanism – negotiated and reciprocal exchanges – is straightforward (see Figure 2.1), it is noteworthy that reciprocal exchange requires at least two rounds of interaction to exist.

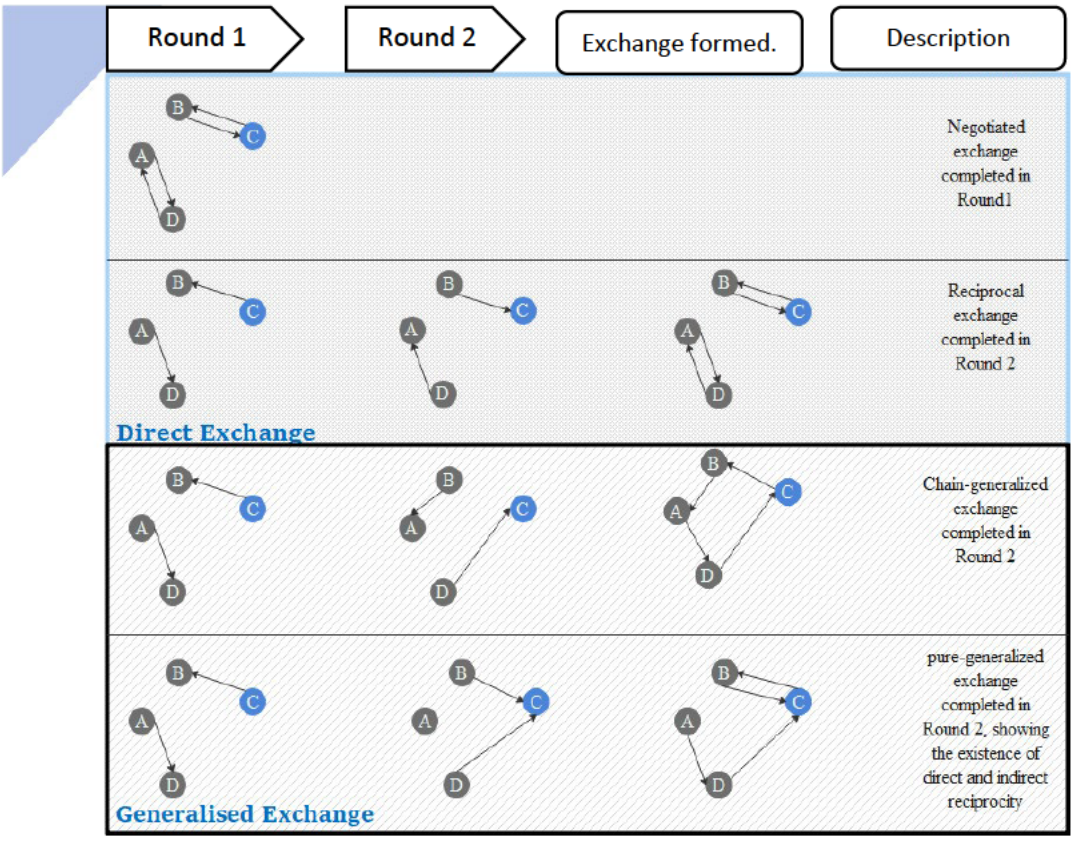

Round 1
Round 2
Exchange formed.
Description
B
C
A
D
Negotiated exchange completed in Round1
Reciprocal exchange completed in Round 2
Direct Exchange
Chain-generalized exchange completed in Round 2
pure-generalized exchange completed in Round 2, showing the existence of direct and indirect reciprocity
Generalised Exchange

Figure 2.1. *Formation of different exchange structures from four actors (avatar A, B, C, and D) within two groups represented by the colour of the avatars. The first two rows show direct exchanges, while the 3rd and 4th rows show generalised exchanges. Complete exchange structure is formed in a single round (round 1) of a negotiated exchange (1st row). Reciprocal (2nd row), chain-generalised (3rd row), and pure-generalised (4th row) exchanges require at least two rounds of interaction for a complete exchange structure to be formed. Interestingly, the last row shows that direct reciprocity (which may be sustained over many rounds) can emerge within pure-generalised exchange (4th row), thus resulting in direct exchange within pure-generalised exchange.*

The next section focuses on explaining the underlying generalised exchange mechanisms, in order to fully understand the complexities imposed by interdependencies in a generalised exchange context.

2.2.2 Underlying Generalised Exchange Mechanisms

Exchange theory provides a theoretical perspective to explain social interactions (Blau, 1964; Homans, 1961, 1974). Homans (1961) took a reductionist approach, arguing that social interaction and its emergent phenomenon could be fully explained at the individual level. This assumption considered neither the environmental influence nor interdependencies in generalised exchange; instead, it was focused on dyadic interaction.

Homans (1961) associated each action with a value. Consequently, Homans simply assumed the cost of an action to be the value of the forgone alternative (Cook et al., 2013) and reward to be the value of the chosen alternative. The problem with this assumption is that the actors may value the same reward or resource differently in social exchange. For example, **A** may prefer friendship to money, while **B** prefers money to friendship.

Building on Homans (1961) work, Blau (1964) viewed social exchange and its emergent phenomenon as a result of the interplay between groups and between individuals within a group. Blau's approach focused primarily on the forms of association and emergent social structures that originated from this kind of social interaction. As pointed out by Cook et al. (2013), these structures of interdependencies play a greater role than that of the dyad. Thus, Blau (1964) view created a platform for Yoshikawa et al. (2018), who proposed regulating rules for generalised exchange.

Yoshikawa et al. (2018) attribute the emergent structures to simple acts of giving, which the authors labelled as rules. The first, *pay-it-forward* (PIF), is an act of giving because the actor received (Nowak & Sigmund, 1998) from someone in the social group. The second, *rewarding reciprocity* (RR), is an act of giving because the intended receiver gave to someone else (Takahashi, 2000). The third, *unilateral giving* (UG), is an act of giving with an expectation to receive (Yamagishi et al., 1999) from someone in the social group. Although the author noted that the focus on rules was to address the *black box* (non-explainability) in social exchange research, this study argues that simplifying the complexity of generalised exchange within these three rules ignores some important factors necessary to explain the emergent behaviours. For example, assume **A**, **B**, **C,** and **D** are actors in a social group. In Rewarding Reciprocity, an actor **A** may reward actor **B** and not **C**, where **B** and **C** gave equal resources to actor **D**. One of the reasons for **A**'s choice of reward could be that **A** rewards the rich and not the poor, in order to be famous (seeking power); or **A** is an in-group member and **B** is an out-group (identity). It could also be that **A** rewards the poor and not the rich in the act of being fair, as seen in (Hauser et al., 2019).

In addition, the term reward is subjective and thus dependent on the objectives of an individual involved in an exchange. In a social exchange context, an objective of individual **A** may be to make more friends while individual **B** wants to make more profit or be famous. An individual may also have multiple, obscure and complex objectives in relation to the experienced interdependencies within an exchange structure. This is due to interdependencies influencing the individual's take on the meaning of the rules of the game.

Taking on the effects of interdependencies within and between groups, recent work (De Dreu et al., 2022; De Dreu et al., 2020) shows that interdependence (such as reputation and reciprocity) structures could foster intergroup conflict in a social exchange context. For example, individuals engage in repeated interactions with the in-group members when they take a positive reputation or in-group reciprocity as the rule to become a reliable in-group member; this, in turn, creates group boundaries (De Dreu et al., 2022). Indeed, this view of interdependence structures can also create parochial competition (De Dreu et al., 2020), where in-group members are "willing to self-sacrifice for the protection and prosperity of the in-group" (De Dreu et al., 2022, p. 114).

2.3 Complexities in a Pure Generalised Exchange Context

Interdependencies in social exchanges create complexities via two phenomena: information scarcity and emergence. Both these phenomena need to be demystified to enhance modelling and simulating social exchange. Whereas emergence results from downward and upward causation, growth and appearance, and evolution of structures (Corning, 2002) in a social exchange system, information scarcity arises as a result of multiple (multiple goals), obscure (not shown) and complex (evolving) objectives of the individuals.

In order to fully grasp social exchange, it is critical to understand the reality of a situation. Moreover, it is important to understand the behaviours that are exhibited by actors, through determining the interdependencies that influence the structure of their objectives (Kelley et al., 2003). Identifying the objectives of emergent behaviour is critical to understanding why behaviour occurs.

Objectives may be multiple in nature, which means that they are set to accomplish multiple goals. This creates many hurdles in being able to identify emergent behavioural patterns and why they occur. The problem of modelling and simulating social exchange does not only lie in the multiplicity of the objectives of the individuals, but also in the obscurity and complexity of the objectives. For example, individuals may only cooperate to benefit themselves and may keep this motive hidden so as not to create in-group conflict. Thus, obscure objectives refer to those objectives which individuals do not clearly show during interactions. These objectives also often change during interaction and become problematic for modelling an effective social system.

There have been many ways of tackling multiple objectives (see Gunantara, 2018 for a recent review). For example, different weights can be assigned to each objective based on how important the objectives are to the individual. However, this is not the focus of this review. In addition, given the obscure and complex nature of the objectives, it may be difficult to know with certainty which objective is more important to the individual. Conversely, it is difficult, if not impossible, to determine the fitness (i.e. a measure of efficacy) of each individual in achieving its objective. This is because the fitness of an individual is related to its objective, indirectly or directly. Put differently, it will be non-trivial to measure how effective an individual is at achieving its objective if the individual's objective is not known. For example, an individual may cooperate or act collectively to gain approvals from high-status in-group members but may hide this motive. This study problematises this (see Chapter 4) and suggests ways in which artificial intelligence can be utilised to model individuals and the emerged system in social exchange.

2.3.1 Information-Scarce Interactive Social Context

As discussed, multiple, obscure and complex objectives give rise to information scarcity. What implication does this have for modelling and simulating social exchanges? How much information is required to model actors' exchange behaviour in a social exchange context? Evidence from literature shows that the accuracy of a social exchange model depends heavily on how well the model captures both the exchange structure and the participants' interpretation of their context. Morgan et al. (2021) pointed out that "both the structure of exchange relationships, and the cultural logics that govern them influence the benefits that exchange partners contribute and receive" (p. 1).

Morgan et al. (2021) stipulate that the impact of cognitive information is critical in understanding social behaviour. Morgan et al. (2021) highlight that their model replicated negotiated exchange but not reciprocal exchange. This is because reciprocal exchange has more variability and non-uniformity in nature when compared to negotiated exchange. In other words, reciprocal behaviour encompasses more dynamic and interdependent objectives during exchange. The presence of these complex objectives creates hassles for understanding the motivation behind behaviour. This report highlights a new challenge for exchange scholars interested in modelling exchange behaviours where negotiation is rare or not feasible among participants, who alternatively rely on reciprocal exchange.

This study focused on information-scarce environment for two reasons. (i) The data that is accessible to the researcher does not contain all the required information to building complex model capable of predicting human behaviour in a complex social exchange environment. (ii) Situations exist where the only available information is the structure of the exchange relationship and the history of exchange, in the form of direct and indirect reciprocal exchanges. Furthermore, the presence of objectives that are multiple, complex and obscure in nature creates variation which adds to information scarcity within the exchange environment. The multiple, complex and obscure objectives present a situation where information that are available are not adequate for the intended level of complexity required to be modelled. It is a challenging task for modellers to incorporate such complexities into their models while presented such information. This suggests that (i.) psychologists should do experiments in rich social settings capturing qualitative exchanges, and (ii.) this study is faced with the challenges of building complex model that can learn from data that does not contain all the required information. Morgan et al. (2021) demonstrated that it becomes difficult to model exchanges without cognitive information surrounding the exchange relationship structure and objectives in such an information-scarce environment. For example, when shopping, one may find an item one wishes to buy. However, their objectives for purchasing such an item may be multiple, complex and obscure. Motivations may range from buying a gift for a loved one, or simply attempting to save money on a certain item – or even both. Moreover, the rationale for buying a gift may be obscure and only known to the individual. Understanding the cognitive information within the system will reduce variation in the model; however, this is no easy feat (Morgan et al., 2021).

2.4 Challenges and Opportunities

The main challenges to understanding, modelling and simulating social exchange remain the difficulty in (i) understanding the motives behind an individual's action(s) when little or no cognitive information is available, and (ii) understanding how interdependencies influence the behaviours that emerge. Despite Yoshikawa et al. (2018) proposal that pay-it-forward, rewarding reciprocity and unilateral giving are the regulating rules for interaction in a generalised exchange context, generalised exchange is not exempted from this main challenge. This is because the action may be because of factors such as: (i) previous interaction, (ii) environmental influence, or (iii) various hidden intentions not captured by those rules but which emerge during interactions over time. This study suggests that the context of behaviour, emergence of behaviour, and the motives of the interacting individuals are essential considerations for tackling these challenges and understanding the social exchange system.

2.4.1 Context of Behaviour

The context of behaviour deals with the co-evolution processes between the state of the system and the individuals within that system. In Bedau's (1997) concept, the context of behaviour is the manner in which interdependencies influence interactions between micro-states and the macro-state. The previous section established that these processes lead to emergent behaviours. Thus, there remains a critical need to understand how interdependencies within a structure are related to its effects (emergent behaviour), in order to further understand both direct and generalised exchange.

According to Gibbs Jr and Van Orden (2001), "dynamical approaches to cognition *also* emphasise that learning always occurs in systems that are environmentally embedded, corporeally embodied, and entrained by feedback" (p. 369). In a social exchange context, this implies the individuals and their states, as well as the environment and the state of the system. For example, the state of an individual involved in a social exchange could be the individual's wealth, social status, and the number of friends the individual has. This could also be described in terms of the subjective context, such as competition. On the other hand, the state of the environment could include the number of groups that exist, the level of inequality in the system and the number of common resources available in the system.

These states have effects on how individuals interact with in-group or/and out-group members. Individuals adjust their goals and strategy as the state of the system changes. For example, the state of a social exchange system may change from parochial cooperation (resulting from shared in-group reciprocity) to parochial competition, which may arise as a result of negative interdependence between groups.

2.4.2 Emergence of Behaviour

Emergence is when "the high-level phenomenon arises from the low-level domain, but truths concerning that phenomenon are unexpected given the principles governing the low-level domain" (Chalmers, 2006). Emergence, a concept whose influence is longstanding in psychology, has been defined from several perspectives (see, Bonabeau & Dessalles, 1997; Sawyer, 2002) and (Kalantari et al., 2020 for a recent review). For example, Bonabeau and Dessalles (1997) view emergence from the perspective of a detector and the phenomenon detected. According to the authors, as pointed out by Deguet et al. (2006), emergence is a phenomenon that occurred when a detector D used a set of tools T to detect a phenomenon at time t+1, such that the relative complexity C of the system S at time t+1 is greater than that at time t. The detector is the observer. If there is no observed change in C, then it is assumed that nothing emerged. This view of emergence is subjective, as it entirely depends on the observer who must decide whether or not a change in the state of the system was observed.

However, emergent behaviour does not occur simply due to cause-and-effect outcomes but rather due to the experiences of the actor within the exchange system. This interdependence is critical in understanding the emergence of social behaviour. The view of emergent behaviour at time t+1 does not allow for the comparison of different observation/detection points to truly identify the point of emergent behaviour.

The active comparison of observation and detection points through an agent makes sense, especially in simulations and agent-based models (e.g. Schelling, 1971), where emergence is assumed to be a phenomenon relative to the complexity that is originally neither present in the system design nor the rules that govern the agents' behaviour.

Sawyer (2002) outlines that the complexity of emergent behaviour cannot be understood simply by defining and describing the components of a model in full. This means that the prediction of behaviour is not possible in advance. Simulations provide a key to unlocking the co-evolutionary nature of emergence.

According to Bedau (1997), “weak’ emergence – in a simulation context – is an evolution that occurs as a result of interactions between various micro-states (individual level) and macro-state (system level), which are governed by the dynamics of the system.

Both the concepts of emergence as posited by (Bonabeau & Dessalles, 1997) and Bedau (1997) have minimal settings for emergence to be claimed: “something appears” – a phenomenon; “it happens within the dynamics of the system” which involves “at least two levels that are distinct” and “it satisfies a criterion that makes it an emergent” (Deguet et al., 2006, p. 28). It is, therefore, not surprising that emergence appeared severally in social exchange literature (Axelrod, 1981; Deneubourg et al., 2002; Titlestad et al., 2019). For example, (e.g., Emlen, 1952; Schelling, 1971) have shown that most systems involving interacting subsystems or elements often produce emergent phenomena typical of social systems that we encounter in our everyday life – from market places to schools and communities.

Therefore, emergence is a fundamental consideration in social interaction. It can take many forms, ranging from structure to behaviour. Indeed, most emergent behaviours create emergent structures, especially in social networks, which is not the focus of this study.

The study at hand considers emergence as the evolution of behaviours in a social exchange context. This evolution is caused by the interdependencies among individuals; between groups and individuals; and the state of the social exchange system. The state of the system has massive impacts on the behaviour of the individuals, over and above the personality of these individuals. This study posits that influencing the interdependence structures in a social exchange system is a way of changing or creating emergent behaviours in the system. This change can be considered positive if it enhances intergroup relations. This agrees with De Dreu and colleagues’ assertion that “positive interdependence between groups can allow cooperative norms or expectations of reciprocity to traverse group boundaries” (De Dreu et al., 2020, p. 762)

Thus, understanding and modelling a social exchange system requires: (i) adequate anticipation and consideration of possible behaviour that may emerge, and (ii) a way of allowing individuals in the system to internalise the emergent phenomenon(a). Consequently, an adequate model of a social exchange system should provide a way to capture the emergent behaviours and feed them back into the system for a more accurate prediction of the system’s behaviour.

Therefore, the modeller must refer back to multiple complex motives, as discussed earlier, to explain the reason for, and the process of, emergence in social exchange and use algorithms flexible enough to capture unintended characteristics that may emerge because of interactions. It will be fair to note that sometimes it is non-trivial to model a system of social interaction where individuals (agents) can internalise an emergent phenomenon. As argued by Li et al. (2006) and pointed out by Kalantari et al. (2020), this is because "the designers and users of a system do not have enough knowledge about why and how they *[the emergent phenomena]* occur in the system" (Kalantari et al., 2020, p. 253). Kalantari et al. (2020), therefore, provide more evidence on the difficulty of modelling in an information-scarce interactive social exchange context.

2.4.3 Obscure Motives

In the words of Gibbs Jr and Van Orden (2001), "our understanding of other people's behaviour rests on assumptions about their intentions". These authors explained that people's intentions cause their actions (Gibbs & Van Orden, 2001). This assertion is not farfetched considering the idea in social exchange that intentions drive interactions. Although Yoshikawa et al. (2018) defined regulating rules, the intentions of the individuals in a social exchange context may not fit within these regulating rules – pay for a good previously done to the individual (from *pay-it-forward*), to reward reputation (from *rewarding reciprocity*) and to seek reward via giving and expectations (from *unilateral giving*). An individual may have multiple obscure objectives and motives, not clearly shown (or even fully understood) by the individual. These objectives may also change during interactions.

For example, individuals may only behave collectively, to further boost their status or other forms of benefit and keep this motive hidden so as not to create in-group conflict. This is problematic for modelling any social system. An adequate model must devise means to infer these objectives based on both the history of the individual and the social context. Such inference must provide feedback to the system to understand the individual's behaviour further, since the obscure motives often play out through the traces of individuals' past behaviours. Contrary to the conventional feedback or reinforcement after receiving a training, this study refers to feedback as the influence of the past behaviour of the individual and the social context in which the individual interacts. Thus, feedback can also be defined as the effect of the state of the system and the experience of the individual on the individuals behaviour.

2.5 Implications

This study has several implications for the two categories of studies, first, those that focus on understanding and modelling social exchanges, and second, those that focus on devising strategies to reduce discrimination and bias during social exchanges (Brochu et al., 2020).

For the first category, understanding the mechanisms of social exchange is the first step in modelling in a social exchange context. A model of social exchange capable of predicting, with some level of accuracy, the next possible state of the system or the subsequent possible behaviour of an individual within the system should have the following mechanisms:

- Individuals in the system should have the capability to internalise their state and the system's state at every point in time.
- Any phenomenon that emerged from the system should have a feedback loop into the system. This can be done by allowing/facilitating individuals to be aware of such emergence. Alternatively, it can be done by exerting the effect of such a phenomenon on the system.
- There should be a way to uncover and incorporate an individual's motives when predicting/modelling the behaviour of the individual in a social exchange context.

Furthering the understanding of the coexistence of direct and generalised exchange has implications for research in computer science and psychology. First, it will enhance the chances of applying state-of-the-art algorithms in computer science for modelling in a social exchange context. Second, incorporating psychology theories will increase the acceptability and applicability of the social exchange models by social psychologists who seek to have insight into the future of a social system. Lastly, it will enhance interdisciplinary research in both fields.

Lastly, for the second category, the utmost interest to this study is that changing or creating emergent behaviours in a social exchange system requires influencing the interdependence structures in the system.

2.6 Conclusion

This dissertation started by presenting an overview of social exchange and its importance to social psychologists. It then presented direct exchange and generalised exchange, and argued for their coexistence and complexities, resulting from interdependence structures that exist in the social exchange system. Later, it presented some well-known formulations of regulating rules for interactions – Pay-it-forward, Rewarding reciprocity and Unilateral giving – in a generalised exchange context. Furthermore, the chapter highlighted some important considerations and showed that the proposed regulating rules are expensive assumptions that may not necessarily hold at all times in a generalised exchange context. Also, it presented some important characteristics of social exchange – context of behaviour, emergence, and complex and obscure motives. It highlighted their concepts and importance, and posited that influencing the interdependence structures of a social exchange will in turn influence or change behaviours that emerge. Lastly, it discussed the implications of the study.

CHAPTER 3. COMBINING METHODS TO ENHANCE MODELLING AND SIMULATION OF INTERACTIVE SOCIAL EXCHANGES

Abstract

What can one consider a good model of human decision-making in an interactive social exchange context, and how can one implement such a model in an information-scarce environment? A plethora of studies has suggested solutions to the first part of the question. However, the second part has received little attention despite its relevance in psychology, where data are not readily available or are expensive to collect.

This chapter aims to review the literature on models and simulations of decision-making during interactive social exchanges. This will be accomplished through the identification and analysis of social exchange computational models. The analysis of these models is based on their ability to integrate and manage the complexities of interactive social exchange, namely, emergent behaviour, the context of behaviour, and obscure motives, during interactive exchanges.

While there remain a plethora of studies that seek theory-driven models and data-driven models for modelling human decision-making, this study argues for using the combination of both methods, with the addition, however, of theory-driven data instead of a traditional theory-driven model, where mathematical equations are formulated based on theoretical knowledge. This study argues that this combination would provide a means to learn from both data and theory. Theory enables an in-depth understanding of how the exchange environment and subjective perceptions of the exchange environment are rooted in exchange behaviour. Furthermore, it presents agent-based models as promising for integrating data into a model. Finally, it argues that integrating theory-driven methods with data-driven methods will enhance the performance of models in an information-scarce environment and discusses challenges that may be faced.

3.1 Introduction

In Chapter 2, the dissertation investigated social interaction, taking the perspective of social exchange. It highlighted important considerations such as emergence, and the multiple and obscure motives of interacting individuals in social exchanges. Based on the investigation, the chapter identified features required for an adequate model of exchange decisions in a social exchange context. This informed the aim of this chapter: to review the literature on models and simulations of decision-making during interactive social exchanges to understand if and how these features have been implemented.

Modelling has become a generally accepted method in many fields to predict, describe, explain and provide meaningful interpretation to phenomena such as social exchanges. Modelling has evolved from using a simple mathematical equation to more sophisticated statistical methods (Borshchev & Filippov, 2004) and recently to using machine learning as a component of a complex model (see Brearcliffe & Crooks, 2021 for a recent application). In addition, the applications of models have increased, and the purposes have become broader – from prediction to explanation of complex behaviours (Edmonds, 2017). Despite these sophistications and innovations, Taylor et al. (2015, as pointed out by Greasley and Owen, 2018), acknowledged that representation and prediction of human behaviour, more specifically human decision-making, are the most significant and unanswered modelling challenges.

A social exchange model can be used to understand, predict and explain people's opinions and behaviours, and their impact on the emergent social structure, such as stability, internalisation and cohesion, on the interacting individuals and the entire system (see Jenkins et al., 2018). Thus, modelling in a social exchange context will benefit many organisations, institutions, and researchers, including decision-makers, behavioural scientists, economists and psychologists, who seek to answer questions about human behaviour.

Suzuki and O'Doherty (2020) point out that social exchange is a fluid and dynamic process, where individuals must predict the intentions of others before making a favourable decision – a decision optimized to help the individual achieve their objectives. Parallel to this, others also predict the intentions of the individual in question before making a favourable decision. These favourable and non-favourable decisions are non-trivial; each decision within a system might arise for a plethora of reasons. For example, high-status individuals within an unfair status quo may decide to exchange resources with low-status individuals in order to justify their unfair possession of power within the status quo. This would be viewed as a favourable decision among high-status members, as it protects their status by invoking self-interested behaviour among the low-status members. However, low-status group members would view the decision to practice self-interested behaviours within the group as being unfavourable. To investigate these motivations, scholars have built up various theories surrounding social exchange and its mechanisms (Balliet et al., 2014; Blau, 2017; Blau, 1964; De Dreu et al., 2022; De Dreu et al., 2020; Homans, 1961, 1974; Yoshikawa et al., 2018). These theories are each based on their interpretations, assumptions and perceptions of human behaviour.

Theories have stirred suggestions on what motivates decision-making in a social exchange setting; this may include self-interest, altruism, competition, power, a quest for satisfaction, and approval (MacCrimmon & Messick, 1976; Olekalns & Smith, 2021). However, the prediction and correlates of these decision-making motivations are largely unclear and remain a research gap (Konovalov & Ruff, 2021). Researchers have recently moved towards using quantitative frameworks – computational models and simulations – to fill these gaps.

However, humans are social beings and modelling their behaviours is non-trivial. Based on this fact, Helbing and Balietti (2011) describe a social model as a predictive instrument that assumes individuals respond both to their needs and other people's expectations during interaction. Supporting Helbing and Balietti (2011), Balke and Gilbert postulate that:

> Humans are not simple machines, which is why modelling them and their decision making […] is a difficult task […]. It is, therefore, the model designer's task to decide which aspects of the real system to include in the model and which to leave out" (Balke & Gilbert, 2014, p. 36).

Thus, "when modelling humans, the modelled entities should be equipped with just those properties and behavioural patterns of the real humans they are representing that are relevant in the given scenario and no more or less" (Balke & Gilbert, 2014, p. 13).

Whereas numerous models have been implemented for decision-making during social exchanges, this study chooses a few based on their popularity and recency. This chapter considers whether or not these models can handle the complexities of interactive social exchange. For example, can these models take into account emergent behaviours within an exchange context when generating predictions?

This review starts by defining computational models of social exchange as mathematical representations that enable the prediction of human decision-making or the reasons that drive human behaviour during social exchanges. Computational modelling may be defined as an algorithm that describe the mechanisms behind a social exchange. It may also be understood as a series of equations that simulate and predict social exchanges. This study seeks to focus on those computational models, both the theory-driven and the data-driven, whose output represents a decision during a social exchange (predictive models that simulate social exchange).

3.2 From Rules to Theory-Driven Computational Models of Decision-Making in a Social Exchange Context

Decision-making during social exchange has been a research concern for decades. Nilsson (1977) described a rule-based model. The if-then rules by Nilsson (1977) consist of three basic components which are: (i) a set of rules, (ii) a knowledge warehouse where information relevant to the problem domain is stored, and (iii) a rule interpreter which determines the line of action that the actors will perform in a specific context. A general criticism that can be made of rule-based models is that the actors follow predefined rules and 'do not have' any cognitive process. Thus, Agents are unable to understand and react to their contexts unless the rules that agents need to follow to react to each context have been defined. However, it is practically impossible to know beforehand all the possible contexts that may emerge during social interaction. Also, these models are not suitable for dynamic environments such as social exchange settings. However, rule-based models are simple to understand because of the clear "link between the rules and their outcome" (Balke & Gilbert, 2014, p. 16).

The Belief-Desires-Intention (BDI) model implemented by Georgeff et al. (1998) has been used as a computational decision-making model to accomplish complex tasks in dynamic environments (Bordini et al., 2007). The BDI is a sophisticated rule-based model based on the theory of practical reasoning (Dancy, 2018; Perelman, 1979). As the name suggests, an actor using this model has three core components, which are: (i) belief – the actor's knowledge about its environment, (ii) desire – the possible tasks the actor wants to accomplish, and (iii) intention – plans on how to accomplish the tasks (Cohen & Levesque, 1990).

In Belief-Desires-Intention, each actor passes through a deliberation process before taking action; this entails searching through the plans to determine which plan, considering the actor's belief, is suitable to accomplish the task at hand. In that effect, BDI actors possess some cognitive processes.

An actor can change their plan if the actor is unable to accomplish the task at hand using the current plan. A major criticism of this model is that an actor using this model does not learn from the past (Phung et al., 2005); in this manner, it lacks one of the essential characteristics of interactive social exchange models, which is feedback. However, researchers (Broersen et al., 2001; Broersen et al., 2002) have reacted to the criticisms and improved the traditional Belief-Desires-Intention model by introducing more features. These introduced more complexity, such that actors can have emotions, ego and "deliberate about whether or not to follow social rules" (Balke & Gilbert, 2014, p. 9).

Balke and Gilbert (2014) describe both the Belief-Desires-Intention and rule-based models as intentional models, in that actor decision-making is internally linked and motivated by specific intent. Conversely, normative models move away from intentional models by integrating norms and contexts, thereby allowing agents' decisions to be influenced by their state. Several normative system models (e.g. Dignum (2004) have been developed. The differences between these models are mainly in the way they represent the normative concept. The basic components of the normative system models as described by Castelfranchi et al. (1999) are three layers, namely, interactive management, information maintenance, and process control layers. An interactive layer handles the interaction between actors and their environment; the information layer stores information about the actors and their environments, while reasoning and information-processing take place in the process control layer.

Enayat et al. (2020) developed a theory-based computational model of social exchange based on Homans' propositions (Homans, 1961, 1974): success, value, stimuli, deprivation-satiation, aggregation-approval and rationality. As the propositions defined, the model assumed that: (i) a positive reward for action increases the probability of performing that action (success proposition); (ii) a probability of performing an action increases if a similar action has been previously rewarded (stimuli proposition); (iii) the probability of performing an action is proportional to the value of the reward obtained from performing that action previously (value proposition); (iv) The more an action is repeated, the less it is valued (deprivation-satiation proposition); (v) a well-rewarded action, according to the judgement of the actor, is likely to be repeated by the actor (aggregation-approval proposition); and (vi) when presented with alternatives, an actor will choose one with maximum gain (rationality proposition).

Decision-making was modelled as a set of probabilities calculated from the combinations of various propositions during exchanges. The exchange occurs in two forms, namely: *transact* and *explore*. *Transact* refers to exchanges between parties who have exchange histories, where previous transactions are integrated into the exchange. By contrast, *explore* refers to exchanges between parties without exchange history.

The memory retention of past transactions allows for the model to consider cognitive features, which Enayat et al. (2020) reported as one of the reasons for the emergence of the precise structure – different decision groups – shown by their simulated result. However, a key limitation of the model is its inability to consider how the emergent structure or the state of the system affected the interaction among actors. For example, the study reported that the absence of emotional factors, such as feelings and anger, causes less group distinction. The number of groups formed is a system state which could impact the actor's choice during transactions but was not considered during the model formation. Nevertheless, the study supports the hypothesis that dyadic interaction involving emotion promotes cohesion. It further showed that feedback (adjustment of behaviour based on experience) enhances the accuracy of decision-making.

Using a more recognised cognitive model, Morgan et al. (2021) investigated the influence of exchange structures on exchange decision-making. Their main focus was on reciprocal and negotiated exchange. The study implemented a simplified version of ACT-R (Anderson & Lebiere, 2014) for negotiated and reciprocal exchanges, where each actor has partners with whom to exchange tokens. 'Partner' is defined by the position of each actor, which in turn defines the exchange structure. The study allowed a varied number of tokens to be exchanged in a negotiated exchange but a specified number of tokens for the reciprocal exchange. This version of ACT-R encodes the actor's past experiences (i.e. decisions) with their partners as a triad: (i) the partner's name, (ii) the amount of tokens received from the partner in the previous round, and (iii) the round the token was received. These triad are cognitive features involving mental abilities such as memorising and reasoning. The inclusion of these cognitive features enabled the generation of expectations specific to each partner and the number of tokens to be expected during negotiated exchange; however, reciprocal exchange was considered to be a probability (i.e. the likelihood of an actor receiving a token out of reciprocity from a partner).

For negotiated and reciprocal exchange, Morgan and colleagues set up two structures – high-power versus low-power network – originally defined in (Molm et al., 2013), which simulated exchange decisions and compared the simulation result to the real data from the experiment conducted by Molm et al. (2013). Morgan et al. (2021) simulated the decision-making in negotiated exchange but could not in reciprocal exchange. The model showed that cognitive features are vital for negotiated exchange. This result is not surprising, as negotiation often requires remembering the past and comparing it with the present. Furthermore, it promotes the idea that feedback is essential for negotiated exchange but has no predictive assistance on the effect of feedback on reciprocal exchange. Reciprocal exchange presents with more variability when compared to negotiated exchange by this contrast, when simulated. Since decisions in the reciprocal structure are highly stochastic and dynamic (Balliet & Van Lange, 2013; Johnson & Mislin, 2011), data-driven models may enable more concrete conclusions to be drawn.

A general problem with theory-based computational models is that mathematical representations of obscure motives (e.g. feeling, faith, satisfaction) are complex assumptions that may not represent actual motives. A more realistic approach could be a data-driven model. For example, individuals can rate their feelings on a scale, and cluster analysis can be performed to see if a higher or lower value is associated for a particular group or set of individuals with another characteristic, such as money.

While feeding data into computational models has existed for decades, it has become popular with the increasing use of agent-based models and machine learning. In fact, the computational model developed by Morgan et al. (2021) is an agent-based model, with ACT-R as its cognitive architecture. Each actor is an agent in the agent-based modelling terminology. However, the agents followed the rules implicitly defined by the equations of the computational model, thus exhibiting the limitations of rule-based models. A more robust approach is to incorporate a learning module within the computational model via data; thus, a data-driven agent-based model. To understand data-driven agent-based models, this study first describes agent-based models in the following section.

3.3 Agent-Based Models

An agent-based model is a computational model that can be used to model interactions among the individuals in the population and between the individuals and the system. There are two major differences between agent-based models and other computational models, namely, heterogeneity and interaction. Heterogeneity implies that different agents can have different natures – type, name, size, and other characteristics – while interaction, in the context of social exchange, implies that agents can communicate – via exchange of token (Morgan et al., 2021), money, approval, or other social values.

An (2012) postulated that “an agent-based model has a unique power to model individual decision-making while incorporating heterogeneity and interaction/feedback” (An, 2012, p. 27). Railsback and Grimm (2012) postulated that “using agent-based models lets us address problems that concern emergence” (Railsback & Grimm, 2012, p. 10), including behaviours or states that arise as a result of interaction among agents and their environment. Indeed, agent-based models allow one to view social interaction and emergent structures in real-time. It provides a platform for studying systems dynamics and visualising how the characteristics and behaviour of each individual cause a change in the system.

It is evident from literature (Flache & Macy, 2004; List et al., 2008; Macy & Flache, 1995; Schelling, 1971) that agent-based models have been widely used to model social interaction. A question that may come to mind is, considering the capabilities of agent-based models as stated by Railsback and Grimm (2012), can agent-based models deal with all the complexities of interactive social exchanges, including emergence, and incorporate behavioural context and feedback? The answer to the question is highly dependent on what controls the agents’ decision-making.

3.3.1 Agents’ Decision-Making in a Social Context

The agent-based modelling community has explored three basic ways of simulating human decision-making: (i) rational actor models, which assume that people make optimal choices; the goal is to choose the decision that leads to maximum achievable utility; (ii) rule-based models, which assume that people follow a specified rule in making choices; the modeller defines the rule; and (iii) behavioural models, which assume that people sometimes deviate from making optimal choices. Implicitly, rational actor models can be described as rule-based models whose rule is to maximise utility. Rather than defining a simple rule, behavioural models sometimes use probabilities or data from human behaviour or experiment to inform agents’ behaviours.

Several studies (Do et al., 2010; Phelps, 2013) have been conducted using agent-based modelling to model a network of human interactions in a bottom-up fashion, that is, the network emerges from interactions among the agents while providing the context for interaction. However, these models neglect the fact that, in nature, not only group size and reputation, but also other complex human elements such as self-interest, neighbourhood obligation, aspiration, fairness, and other social norms, affect the decisions that humans make (An, 2012).

While most agent-based models describe emergent features of dynamic interaction networks, they are challenged by the fact that human actors may be affected by network interactions in complex ways, not anticipated by rational actor models, and not readily described by rules (Macy & Flache, 1995). The rules themselves may be emergent and change in accordance with the interaction context. These arguments have implications for interactive social exchange because rules undermine emergence – a key feature of social exchange – unless the emergent phenomenon has a way of influencing or changing the rules. It remains an open argument whether it is better to specify rules that govern the agents' behaviours or allow them to be learned over time. In addition, there are arguments about the level of complexity suitable for modelling human behaviours – in this context, decision-making (Axelrod, 1997; Edmonds & Moss, 2004).

3.3.2 Simple versus Complex: Implications for Agent-Based Models of Social Exchange

Carley et al. (1998) and Gilbert (2004) review numerous computational models describing different agent-based frameworks and ways to simulate human decision-making. Most of the models reviewed can be classified as either using the KISS or the KIDS approach. Axelrod (1997) described a simple method called KISS, which is an acronym for 'Keep it simple, stupid', while Edmonds and Moss (2004) proposed a method called KIDS which stands for 'Keep it descriptive, stupid'. The rationale behind 'Keep it simple, stupid' is that "the phenomena that emerge from simulation exercises should be the result of multi-agent interactions and adaptation, and not because of complex assumptions about individual behaviour" (Duffy, 2006, p. 954), while the rationale behind 'Keep it descriptive, stupid' is that models should be as descriptive as possible, "even if it implies greater complexity" (Adam & Gaudou, 2016, p. 207).

The major problem with 'Keep it descriptive, stupid' is that *descriptiveness* also involves specifying all the rules that govern interactions. Specifying these rules requires understanding the objectives of the individuals in the system. However, the objectives are sometimes obscure, complex to understand and not always known beforehand. Assuming the modeller's task is to build a rule-based model of reciprocity, the modeller needs to consider all the conditions that promote reciprocation. One of these conditions may be to *reciprocate if satisfied with what you received*. Satisfaction is subjective, and satisfaction levels change over time. This poses a new challenge to agent-based modellers of social exchange. In addition, these rules are sometimes in competition. How agent-based models deal with competing rules is highly determined by the agent's decision-making process.

Although progress has been made with the evolution of different architectures for modelling human behaviour, models that use these architectures require extensive programming skill and knowledge of data structure. Nevertheless, modellers of interactive social exchange could benefit from such architectures. Challenges arise with construction and prioritising different behavioural rules, and capturing phenomena and rules that emerge during interactions over time. One way of dealing with these challenges is combining models and methods to harness the models' strength while avoiding their loopholes. For example, agent-based models can be combined with machine learning. The former provides the ability to handle interactions and emergence, while the later provides the ability to learn from data.

3.4 Data-Driven Models of Social Exchange

Representing the *properties* and *behavioural* patterns of a complex system is non-trivial. As such, data-driven models (see Kavak et al., 2018) have provided a template for using data from human experiments to inform and calibrate models, thereby focusing on those properties and behavioural patterns of the humans they represent. However, extracting relevant patterns from psychological/behavioural data requires a technique that identifies patterns, especially those that arise from complex interdependencies between individuals.

3.4.1 Combining Machine Learning and Agent-Based Models for Data-Driven Modelling of Interactive Social Exchange

Despite the wide range of real-world events, humans learn how to represent complex structures such as item taxonomies (classification, in machine learning terms), latent patterns (pattern recognition, in machine learning terms) and causal structures, simply through experience (Spicer & Sanborn, 2019). Modelling such human behaviours has recently been the focus of machine learning researchers. Many studies (Andión et al., 2021; Augustijn et al., 2019; Augustijn et al., 2020) have successfully integrated machine learning into agent-based models for modelling human behaviours. Most of these studies are motivated by the availability of data and data sources, and the quest to model humans' remarkable ability to acquire complex representational forms through learning. This study argues for using such combinations, however, with the addition of theory-driven data (i.e. data refined based on the theoretical perspective of the conditions and context during data collection).

3.4.2 Data and Data Sources for Combining Machine Learning and Agent-Based Models

Many data sources have emerged over the last few years as environments for revealing and recording human behavioural patterns (Osoba & Davis, 2019). These data sources include social media platforms such as Twitter, Facebook, public and private video surveillance, voice records and experimental platforms such as laboratories (e.g. www.viappl.org). In addition, behavioural data can be obtained from financial records from financial technology firms and banks. Expectations are that, by diversifying the data sources, there will be an increase in the probability of capturing important behavioural patterns which were previously impossible to detect. Thus, from a modeller's perspective, data availability would considerably increase the capacity for behavioural modelling using machine learning and agent-based models.

Kavak et al. (2018) discussed data usage in agent-based models under four categories: data type, repeated measurements, impact on model, and agent consideration. The first, data type, describes whether the data is quantitative or qualitative, discrete or continuous. The data type to be used is dependent on three factors, namely, the specific case being modelled, the availability of data, and the accessibility of the available data.

Aggregated data are commonly available and support agent consideration at a population level. However, the availability of a large number of data sources and the surge in technology have provided ways of capturing data capable of supporting agent consideration at the individual level. As pointed out by Osoba and Davis (2019), the growth in relevant behavioural data sources brings about a growing dependence on tools and devices for human assistance, for example, the Global Positioning System (GPS) for navigation, and mood management via music streams. Hence, data sources have been cited as a potential catalyst for attaining, if at all possible, what Clark and Chalmers (1998) termed "The Extended Mind" – the existence of cognition and mental deliberation outside the human mind.

3.4.3 Example of Studies Using Data-Driven Agent-Based Models

Several studies have shown that machine learning and agent-based models can be combined to provide insight into a rather complex situation that was previously difficult to understand in interactive social exchange. An approach that combined machine learning and agent-based models, capable of addressing some of the features of the interactive social exchange, is presented as an improved version of the traditional data-driven agent-based model (Kavak et al., 2018).

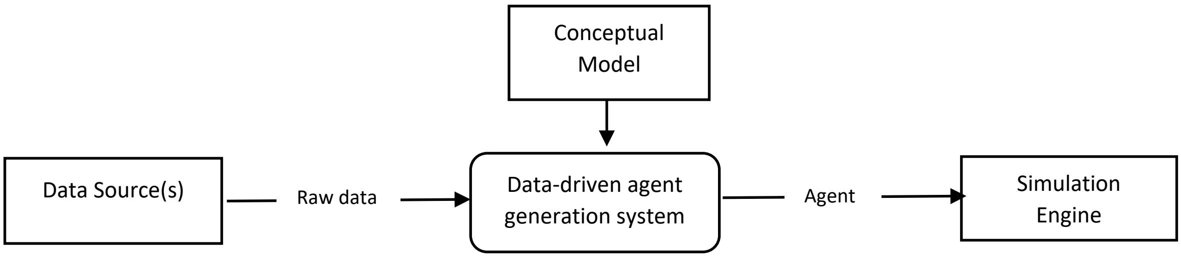


Figure 3.1. *Contextual data flow diagram of Kavak et al. (2018) data-driven approach.*

The model developed by Kavak and colleagues has four basic components, as shown in Figure 3.1. The conceptual model is an abstraction of the traditional agent-based model. It should contain the purpose statement, that is, the goal of the model, the types, the attributes and the behaviour signatures of the agents. It should also include the environment type and variables. Data source(s), as the name implies, describes the source of the data. This component involves pre-processing and separation of data into attribute and behavioural data. For example, the data are prepared for behaviour creation. The tasks include removing noisy entries, managing missing data, transforming data fields, and removing unusable entries. Data-driven agent generation then involves creating the agent's behaviour. An analogy to conceptual and data-driven agent generation system components is the hardware component of a computer and the computer programs that make the hardware functional – a programmer writes a program that *brings life* into the hardware of a computer system. Similarly, a modeller creates a conceptual model (like computer hardware) and generates agents' behaviours (like a computer program) using data. The created agent (the behaviour created via data integrated into the conceptual model) is then used in the simulation engine.

Zhang et al. (2018) developed an agent-based model that used a machine learning technique to provide insight into urban decision-making. The agent-based model simulates traffic volume and user waiting time in a city district such that, given multiple configurations of the city as an input, it used machine learning to predict and optimise the best configuration based on the user's objective: the best route to reduce traffic given a city configuration. Zhang et al. (2018) used a combination of neural network and an agent-based model developed in a GAMA – a platform for developing agent-based models (Grignard et al., 2013). The authors pointed out that the model can "promote collaborations among a broad range of stakeholders to enhance the accessibility and efficiency of public engagement events" (p. 2172); it frees users from excessive deliberations and improves the quality of their urban decisions.

The Zhang et al. (2018) model showed the promise of combining machine learning and agent-based models. However, there are subtle differences between urban decision-making and decision-making in an interactive social exchange context. First is the notion of interdependence. Whereas a city configuration does not depend on another city configuration, the state (e.g. wealth) of an individual depends on the state of the other individuals. Hence, a change in one city configuration does not change the other city configurations, while a change in one individual's state affects the states of the other individuals in a generalised exchange context. Second is that notion of emergence. Whereas all conditions (e.g. traffic and waiting times) are held constant while making the decision (Zhang et al., 2018), the rules and objectives of each individual may change while making a decision in an interactive social exchange context.

However, Kavak et al. (2018) data-driven agent-based model provides the basics for dealing with most of the complexities of interactive social exchange. However, once the behaviours are created, and the agents using these behavioural rules are introduced in the simulation engine, it is unclear whether or not these agents can change their behavioural rules during interactions over time. The emergence of rules, and the agent's ability to adjust to such emergent rules, are desired features in an interactive social exchange.

3.5 Combining Machine Learning and Agent-Based Models for Data-Driven Modelling of Interactive Social Exchange: A Consideration for Information-Scarce Environments

Although the proliferation of data-driven models resulted from the availability of many data sources, specific behavioural data are still scarce and expensive – and may cause harm – to obtain. In such a context, one way of modelling is to use less data-hungry machine learning algorithms and extract data based on theoretical knowledge.

It is non-trivial to transform psychological data into a form usable for machine learning to generate agents' behavioural rules. This study points out two reasons for this difficulty. The first is that data may be collected under some conditions and norms, including inequality, groups and competition, that are not obvious in the data and may be tricky for machine learning to model if not explicitly defined. Explicitly defining such conditions and norms requires a thorough understanding of psychological theory and the principles that guide such behaviours. Second is that, most times, data in the psychology domain are so noisy that expert knowledge and domain theories are required to make sense of the data.

Other researchers (Adjerid & Kelley, 2018; Dwyer et al., 2018; Jacobucci & Grimm, 2020) have noted similar difficulties. For example, Jacobucci and Grimm (2020) pointed out that "the integration of machine learning algorithms with psychological research requires a great deal of nuance" (p. 809). This is because "issues such as validity do not receive near the same extent of coverage in machine learning, *[a gap]* requiring both the translation of ideas across disciplines and refinement in how the concept of validity can be applied" (p. 810). Researchers who integrate machine learning and psychology must consider a range of theoretical perspectives in order to address issues of validity, and identify and document important psychological concepts.

The integration of theory in artificial intelligence algorithms, including machine learning algorithms, has emerged as a research area called Explainable Artificial Intelligence (XAI) (Wang et al., 2019). Wang et al. (2019) present a theoretically based conceptual framework capable of integrating concepts within human reasoning processes with explainable artificial intelligence techniques. The conception of explainable artificial intelligence aims to assist in the development of user-centric artificial intelligence-based systems. This area is still in its infancy and is yet to be fully adopted by psychology researchers – interested readers to see (Wang et al., 2019).

Rosenfeld et al. (2012) have suggested that a human decision model based on the combination of theories and data-driven methods yields more accurate predictions, especially in complex domains. Rosenfeld and colleagues developed a model informed by Aspiration Adaptation Theory. Aspiration Adaptation Theory posits that individuals make decisions to satisfy one goal at a time; these goals are known as aspirations. The model is able to incorporate previous exchange history in order to evaluate and predict the exchange environment and subsequent emergent exchange behaviour. In this model, individuals are assumed to give ranked importance to negotiation parameters according to their experienced aspiration scale. Decision-making processes are treated as discrete levels of aspirations that must be negotiated. In this manner, the decision to exchange occurs if the level of aspiration defined by the decision is met. Combining data from the individuals' previous choices and decisions based on Aspiration Adaptation Theory, Rosenfeld et al. (2012) showed that knowledge of theory could be used to predict people's choices.

3.6 Translating ‘Keep it descriptive, stupid’ into a Machine Learning Problem: A Way Forward

So far, this study reviewed different models of social exchange and, in the previous section, identified the combination of theories and machine learning as suitable for modelling in a social exchange context. This section suggests ways to actualise this combination.

Adopting the ‘Keep it descriptive, stupid’ idea of ‘descriptiveness’ in an interactive social exchange context suggests that one starts from the available evidence, including expert opinion to model behaviour in this context. This implies that theories and norms that have been proven to guide individual behaviours in a social exchange context must be considered. Research (Yamagishi et al., 1999; Yamagishi & Mifune, 2009) has shown that theories and norms derive behaviours in a social interaction context. For example, Bounded generalised reciprocity (BGR) theory has been used to prove that individuals maintain good relationship with the ingroup to avoid being punished (Yamagishi et al., 1999; Yamagishi & Mifune, 2009). Therefore, the description can be in the form of variables such as fairness and power, formulated from theoretical concepts about social exchanges. Machine learning algorithms, such as clustering algorithms or predictive algorithms, or combinations of these algorithms, could then be used to uncover motives and emergent behaviours from these variables.

3.7 Lessons, Challenges and Further Directions

3.7.1 Interdependency

Machine learning researchers (for example, Santucci et al., 2019) define interdependence in terms of tasks, where a task is subdivided into subsets such that a subset of the task is required to be performed before performing the other subsets. This view of interdependence dominates in the machine learning community, in modelling interdependence as viewed by behavioural scientists (decision, reward or action of an individual dependent on others). The modeller needs a fair understanding of the types of interdependencies between individuals in the system before commencing the modelling task (see De Dreu et al., 2020 for recent work on interdependencies). An integration of perspectives requires an understanding of how the mechanisms of the model influence each other, and the overall model as a whole.

Nevertheless, the availability of data and different machine learning techniques (e.g. clustering algorithms) have enabled researchers to discover interdependencies in data that otherwise are difficult to discover by mere observation. However, theoretical perspectives such as social exchange theory need to be incorporated when dealing with interactive social exchange. This will enable norms and social context to be captured and included as variables that influence the agent's behaviour.

3.7.2 Obscure Objectives

A major challenge is that agents' objectives are not known with certainty. Most models explicitly define the objectives or aspirations of the agents, but this is almost impossible in an interactive social exchange context because the objectives of the individuals evolve. A way of tackling such challenges could be to measure agents' objectives based on the trails of the past interactions. This study posits that this could be achieved by combining unsupervised machine learning algorithms, such as clustering, and supervised machine learning algorithms, such as artificial neural networks. For example, behavioural data obtained from an exchange system could be clustered to discover different objectives and behaviours that exist within the system, while artificial neural networks could be used to classify newly seen behavioural data into one of the existing objectives or as a new objective.

3.7.3 Emergence

The modification of solidified exchange behavioural patterns within the simulation engine of the data-driven agent-based model described by Kavak et al. (2018) is fairly unclear. Consequently, there remains a gap within the social exchange context relating to emergent behavioural patterns within interactive social exchange systems. There is a need to allow agents' behaviour to be modified based on the interaction over time. This will, in turn, affect the interdependence structure, and consequently modify the emergent behaviours. For example, a selfish individual may decide to change his/her behaviour after several interactions with selfless individuals. One way of dealing with the emergent rule is to integrate different models as integral parts of the agents' architecture, to check for new behaviour resulting from agents' interactions over time.

One of the fundamental benefits of agent-based models is that they allow a feedback loop. Also, many agent-based model platforms, such as GAMA (Grignard et al., 2013), allow the emergent structures to be viewed.

3.8 Conclusion

The study started with a brief review of some selected models' ability to integrate and manage the complexities of interactive social exchange, namely, emergent behaviour, the context of behaviour, and complex and obscure motives during interactions. It argued that most non-data-driven computational models make complex assumptions about human behaviour, where the actual behaviour could have been deduced from data. Furthermore, it presented agent-based models, and argued for integrating data and theory in the agent-based model for better decision-making simulation in interactive social exchanges.

CHAPTER 4. A MODEL FOR PREDICTING SOCIAL EXCHANGE DECISION-MAKING: AN INTEGRATION OF THEORY-DRIVEN AND DATA-DRIVEN METHODS

Abstract

Artificial agents that can predict human decisions in social exchange contexts can potentially help to facilitate cooperation and promote prosocial behaviours. Modelling human decision-making is difficult in social exchange contexts where multiple contending motives inform decisions in rapidly evolving situations. We propose a mixed Theory and Data-Driven (TD2) model that is comprised of three modules: (1) a clustering algorithm that identifies strategies in interactive social exchange contexts (2) an artificial neural network that classifies an exchange decision into one of the identified strategies based on empirically defined motives and the observable differences during social exchanges, and (3) a hidden Markov model that predicts situated human decisions based on the strategies applied by humans over time. The TD2 decision-making model was trained and tested using 7,840 exchange data from "minimal group" experimental exchange games in which decisions were motivated by group ties, wealth aspiration, and interpersonal ties. The model was able to classify behaviours with 95% accuracy. Reciprocity, fairness and in-group favouritism were predicted, as separate decisions, with accuracies of 81%, 57% and 71% respectively. The performance of the model improved over time. Future work will evaluate the model in a live experiment involving Human-Agent Cooperation (HAC).

4.1 Introduction

In Chapter 1, a twofold aim was defined: to develop agents that can act in a social exchange context and to use the agents to affect the level of in-group favouritism, which has been discussed by De Dreu et al. (2020) and others in the interdependence context. Towards this aim, Chapter 2 identified emergence and obscure motives of the interacting individuals as the challenges in developing agents that can act in a social exchange context, suggesting that interventions in social exchange are not cause-effect. Consequently, Chapter 3 proposed the combination of machine learning and an agent-based model by adopting the 'descriptiveness' from KIDS – Keep it descriptive, stupid – to overcome these challenges and suggested that an adequate model of social exchange (*vis-à-vis* an agent that can act in a social exchange) requires three key considerations: (i) the understanding of the interdependencies within the exchange system; (ii) the capturing of behaviours, motives and objectives, which evolve in interactive social exchange; and (iii) the ability to capture emergent behaviours and norms via the modification of the interdependence structures within the exchange system. Chapter 4 implements the method proposed in Chapter 3 to develop agent that can act and cooperate in a social exchange context.

In the world where artificial agents have pervaded society, the need to improve Human-Agent Cooperation has increased. By predicting human exchange decisions in interactive social exchanges, artificial agents can potentially help to facilitate cooperation and promote prosocial behaviours. Recent studies (e.g., Domingos et al., 2022) have used strategic games to investigate the effects of agents in social interaction and cooperation. Whereas strategic games such as the Prisoner's Dilemma and GO (or Weiqi) have optimal outcomes, interactive social exchanges like gift giving (1) have multiple obscure motives such as fairness, selfishness and in-group favouritism, (2) whose relative importance evolves through interaction, and (3) in situations that present multiple, overlapping behavioural demands that are poorly differentiated from each other.(Suzuki & O'Doherty, 2020) Think of the multiple obscure motives that underly giving your boss a gift. While decision-making can be easily predicted when motives are known with certainty (Bardsley, 2008; Capraro et al., 2014; Larney et al., 2019), the obscure motives in interactive social exchanges provide indefinite cues for predicting behaviour. We propose a method for predicting decision-making in exchange environments as part of a larger initiative to develop adaptive agents capable of cooperating with humans in social exchange. Specifically, this study develops a model that can infer players' strategies during interactive social exchanges and predict exchange decisions based on these strategies.

Exchange decisions were studied in the context of a simple experimental game (Durrheim et al., 2016) in which members of two 7-player groups were required to allocate a single token to any player in each of 40 rounds. They had to choose *who* to give it to, keeping in mind that a player's token represents the player's wealth. Players' token balances indicate how rich or poor they are relative to others at each round of the game. We define an exchange decision in terms of three identified definite behaviours: (1) reciprocating a gift in the next round, (2) giving to a rich player (seeking power) versus a poor player (fairness) and (3) out-group versus in-group giving. Each exchange decision was defined as the presence or absence of each definite behaviour, represented as three-digit binary combinations in the order defined above. For example, 001 indicates an allocation that was not an act of reciprocation made to a poor out-group player, whereas 111 is a reciprocated allocation to a rich out-group player.

Individuals in interactive social exchange often make exchange decisions as part of a strategy that supports their motives. A strategy was defined as a complex plan that guides an individual's exchange decisions such that each exchange decision conforms to the plan. For example, individuals whose motives are to strengthen in-group ties often allocate their tokens to in-group members as a strategy. An intrinsic complexity of interactive social exchange is that the same strategy can lead to several exchange decisions and several strategies can lead to the same exchange decision. For example, allocating tokens to in-group members may not be motivated by ingroup altruism but by self-enrichment, effected by building trust and eliciting reciprocation. This study develops a Theory and Data-Driven (TD2) model to simplify and enhance predictive accuracy in challenging contexts of obscure, overlapping and evolving motives. We tackle the complications by identifying definite behaviours and then seeing how these are used in combination and stack up over time to become part of complex strategies. The strategies then influence the exchange decisions individuals will make in a particular context.

The contribution of the paper is threefold. First, it provides a novel method of integrating theory into machine learning models via theory-driven data. Second, it demonstrates how theoretically grounded motives can be used to infer strategies and predict exchange decisions. Third, it contributes to the research on improving models of decision-making in Human-Agent Cooperation (HAC).

4.1.1 Decision-Making Models of Social Exchanges

Previous models of decision-making relied on either rule-based (Georgeff et al., 1998; Nilsson, 1977), theory-based (Enayat et al., 2020; Morgan et al., 2021) or data-driven (Kavak et al., 2018) algorithms. The if-then rules proposed by Nilsson (1977) consist of three basic components which are: (1) a set of rules, (2) a knowledge warehouse where information (e.g., the rule precedence) relevant to the problem domain is stored, and (3) a rule interpreter which determines the line of action that the actors will perform in a specific context. A general criticism that can be made of rule-based models is that the actors follow predefined rules and the model does not learn from the past (Phung et al., 2005); these models are not suitable for dynamic environments such as social exchange settings where information from the past exchange shapes the current decision.

Conversely, most theory-based models (e.g., Enayat et al., 2020) represent decision-making as a set of probabilities. Enayat et al. (2020) developed a theory-based computational model of social exchange as a set of probabilities based on Homans' propositions (Homans, 1961, 1974): success, value, stimuli, deprivation-satiation, aggregation-approval, and rationality. For example, the probability of performing an action is proportional to the value of the reward obtained from performing that action previously (value proposition). Although this approach showed how feedback (adjustment of behaviour based on experience) can inform decision-making, a key limitation is its inability to consider how the emergent structure or the state of the system affected the interaction among actors. For example, Enayat et al. (2020) reported that the absence of emotional factors, such as feelings and anger, causes less group distinction. Such distinctions are emergent states of the system which could impact the actor's choice during transactions but were not considered during the model formation. Also, a general problem with theory-based computational models is that mathematical representations of motives rely on assumptions that may not represent actual motives. Rather than using mathematical representations, we deduce possible motives from theoretical knowledge, represent these motives as data which feeds into the model and allow individual motives to change over time due to the stochastic nature of interactive social exchange.

Data-driven models are suited to representing the highly stochastic and dynamic (Balliet & Van Lange, 2013; Johnson & Mislin, 2011) nature of human decision-making. Kavak and colleagues(Kavak et al., 2018) have provided a template for using data from human experiments to inform and calibrate models, thereby focusing on the properties and behavioural patterns of the humans they represent. While feeding data into computational models has existed for decades, it has become popular with the increasing use of agent-based models and machine learning. Many studies (Andión et al., 2021; Augustijn et al., 2019; Augustijn et al., 2020) have integrated machine learning into agent-based models for modelling decision-making. For example, Augustijn et al. (2020) used machine learning to generate decision rules from data. Although this method ensures that the decision rules represent the actual decision-making model of the individuals being considered, it relies on the availability of a large dataset that represents a variety of behaviours in the context. There is no guarantee that the decisions will be well captured in the rule if the dataset is small. Care must be taken to ensure the quality of a small dataset to improve decision rules that may be generated from the dataset. A quality dataset (i.e., containing enough information necessary to capture decision patterns) is vital for accurate or near-accurate prediction of human decision-making.

Given that a large dataset is not available in the interactive social exchange context considered in this study, we ensure the quality of the available dataset by combining machine learning and Edmonds and Moss (2004) idea of 'Keep it descriptive, stupid' (KIDS). The idea is that models should be as descriptive as possible. This study applies KIDS by applying a theoretical framework for describing and understanding the possible motives of individuals in interactive social exchanges. Guided by the theoretical framework, we generate data that were not physically observable during the interactive social exchanges to improve the quality of the available dataset.

4.2 Method

This section presents the proposed model, the configurations, and the training and evaluation of the model. The conceptual model represented in Figure 4.1 shows the relationships between the key concepts in our proposed model. We discuss how these concepts were operationalized in our model.

Figure 4.1 shows how each stage in the model feeds into another. It also shows how the exchange decisions feedback into the model to inform subsequent decisions. 1 and 2 are collectively termed the learning phase while 3 and 4 are collectively termed the prediction phase. In the learning phase, we theoretically identified motives (internal states) for exchange decisions. The combinations of the observable differences between individuals and definite behaviours that signify the motives of the individuals are the observed behaviours in this phase. Subsequently, cluster analyses of the observed behaviours (of the learning phase) were used to impute unobserved game strategies (internal states of the prediction phase) to predict exchange decisions (observable behaviour of the prediction phase). Further explanations are provided in subsections. The material (code, data and analyses) for this model can be found via the link: OSF | Predicting human exchange decision-making with theoretically informed data and machine learning.

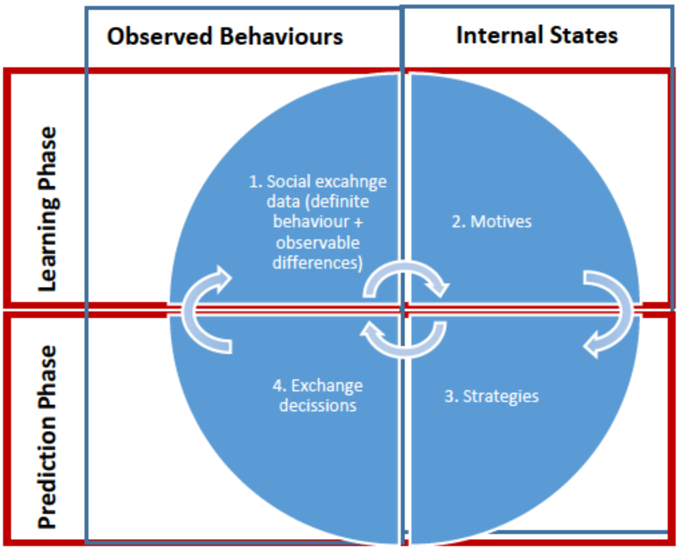


Figure 4. 3. A conceptual representation of the proposed model. 1. Social exchange data which include definite behaviours and observable differences between individuals in the game. 2. Theoretical identified motives that underlie exchange decisions in interactive exchange. 3. Individuals' game plans obtained from the combination of pre-processed social exchange data and motives via cluster analysis. 4. Combination of the definite behaviors showing individuals' choice of allocation.

4.2.1 Social Exchange Data and Exchange Decisions

The proposed model aims to predict human exchange decisions during interactive social exchanges. The model was trained and tested using interactive social exchange data and theoretical knowledge of possible motives underlying social exchange decisions. The data was obtained from Virtual Interaction Application (VIAPPL) 2013 and 2014 experiments reported by Durrheim et al. (2016). In the experiments, individuals were required to allocate a single token per round to any player in the two 7-player groups. The selection was based on the observable differences (e.g., token balance, group identity, and previous allocations) that differentiate the players.

An individual's token balance in each round shows how wealthy the individual was in that round. Thus, a token balance was regarded as a show of the individual's status. An individual's status was measured relative to the wealth of the other individuals in the game, i.e., we measure how wealthy an individual is in comparison to other individuals. Status was determined by the formula $Status = a/max(A)$, where a is the token balance of an individual in a round, and A is a vector of the token balances of all the players in that round. Thus, status is represented as a real number in the range of 0 to 1 inclusive. An individual is of high status if the individual's status is greater than the average status in the round; otherwise, the individual is of low status. Group identity simply indicates the group to which the individual belongs. The group identities are represented as 1 and 2 for group 1 and group 2 respectively. Previous allocation indicates an in-group giving, out-group giving or self-giving in the previous round, and is represented as 0, 1 or 2 respectively.

The experimental games were randomly assigned to conditions in which players and groups started with either equal or unequal token balances. The data for each allocation in each round records the game identifier, the group number, experimental condition, starting and ending token balance, and directed ties showing player to player allocations. These ties provide traces of relational interdependencies (e.g., competition and cooperation) that develop between interacting individuals and groups. The unprocessed data contains 2 (conditions) x 40 (rounds) x 14 (participants) x 5 (games) of VIAPPL 2013 data and 1 (condition) x 40 (rounds) x 14 (participants) x 4 (games) of VIAPPL 2014 data reported in (Durrheim et al., 2016). This implies 7,840 exchanges. To reduce noise in the data, the first rounds of the data were not used because players were likely to randomly allocate their tokens.

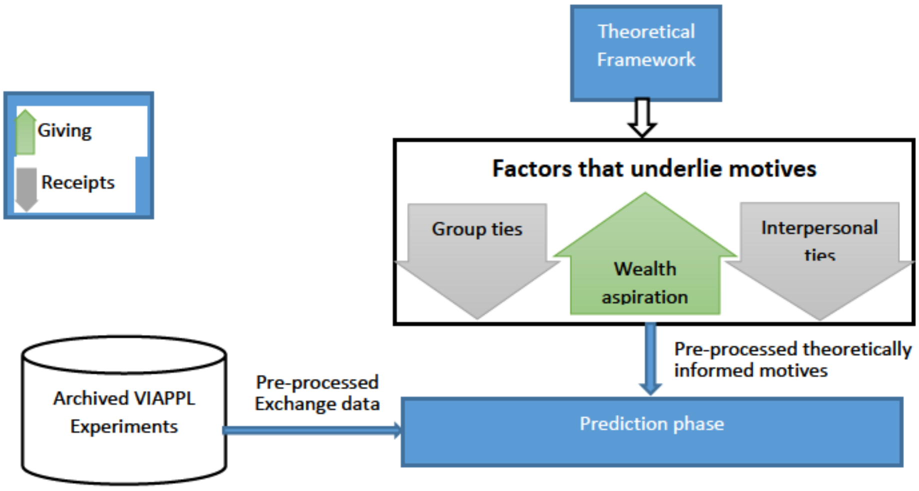
Each exchange shows the presence or absence of each definite behaviour: reciprocity, defined as giving the player who gave you in the previous round, vs non-reciprocity, (2) giving a rich vs giving a poor player and (3) giving out-group vs giving in-group. Table 4.1 presents the exchange decisions and their explanations as used in the current study. An act of reciprocation is indicated as 1 while its absence is indicated as 0. Giving a rich player (1) and giving out-group (1).

Table 4.1. The exchange decisions as determined by the definite behaviours

Reciprocation (versus Non-reciprocation)	Giving rich (versus giving poor)	Out-group (versus in-group giving)	Exchange decision	Explanation of the exchange decision
0	0	0	000	Allocate a token to a poor in-group member
0	0	1	001	Allocate a token to a poor out-group member
0	1	0	010	Allocate a token to a rich in-group member
0	1	1	011	Allocate a token to a rich out-group member
1	0	0	100	Reciprocate to a poor in-group member
1	0	1	101	Reciprocate to a poor out-group member
1	1	0	110	Reciprocate to a rich in-group member
1	1	1	111	Reciprocate to a rich out-group member

4.2.2 Motives

This section shows how definite behaviours help to signify the motives underlying exchange decisions. The theoretical framework in Figure 4.2 identifies three motives, namely, group ties, wealth aspiration, and interpersonal ties, that are likely to inform exchange behaviours. These motives, inferred from previous receipts (in the case of ties) and previous allocations (in the case of wealth aspiration), are the theoretically identified motives that account for the absence or presence of the definite behaviours in exchange decisions. The strength of each motive is measured from past observed behaviour and is used to impute strategies which in turn, is used to predict future exchange decisions. For example, an individual who has a strong motive for fairness in the game will often give a token to the poorest player (definite behaviour), irrespective of the poor player's group.


Giving
Receipts
Theoretical Framework
Factors that underlie motives
Group ties
Wealth aspiration
Interpersonal ties
Pre-processed theoretically informed motives
Archived VIAPPL Experiments
Pre-processed Exchange data
Prediction phase

Figgure.4.4. The acts of giving and the receipts that instill motives in eventual givers. Group ties (motivating giving to in-group vs out-group, Wealth aspiration (motivating giving to the rich vs giving to the poor) and Interpersonal ties (motivating reciprocation vs non-reciprocity) are the identified theoretical motives that underly exchange decisions.

Group ties measure the relationship between individuals and groups. Individuals and groups adopt various identity management strategies aimed at creating positive social identities (Ellemers, 2001; Turner & Tajfel, 1986). Social identity theory (Tajfel et al., 1971) provides a theoretical lens to understand the social psychological motives of social exchanges. The theory argues that individuals categorise as group members in intergroup contexts(Turner & Tajfel, 1986), compare the status of in-group and outgroup, and are motivated to identify with groups that are positively valued within the hierarchy (Kisfalusi et al., 2019). These processes occur as the individual seeks to achieve a positive internal perception of self or a high perception of self-worth (Tajfel, 1982). Prejudice and discrimination occur as an expression of this positive distinctiveness motive, which is expressed as in-group favouritism behaviour or parochial altruism intergroup exchange experiments (Balliet et al., 2014; Tajfel & Turner, 1979).

Bounded generalised reciprocity (BGR) theory (Yamagishi et al., 1999; Yamagishi & Mifune, 2009) argues that parochial altruism is ultimately motivated by self-interest. Individuals favour in-group members because there is an expectation to do so. To avoid acquiring a bad reputation and being excluded from exchange network, individuals will favour the in-group compared to the out-group. In other words, individuals expect profitable and advantageous interactions with in-group members because of the expectations that favours are more likely to be reciprocated by in-group members compared to the out-group members (Balliet et al., 2014; Yamagishi et al., 1999).

Both SIT and BGR expect that in-group favouring behaviour will strengthen the bond between the in-group members. We refer to this bond as *group tie*. An individual motive may be to strengthen the bond with the in-group (or even with the out-group), to maintain a good reputation. Therefore, group ties were defined in terms of in-group versus out-group exchange.

Table 4.2. Demonstrates the calculations of in-group and out-group relationships. $\boldsymbol{N_{in}} = \boldsymbol{N_{out}} = \mathbf{4}$ *representing the number of in-group and out-group members respectively.*

Participant No	**T_{in} (In-degree from in-group)**	**In-group relationship**	**T_{out} (In-degree from out-group)**	**Out-group relationship**
1	1	$1/4 = 0.25$	1	$1/4 = 0.25$
2	0	$0/4 = 0$	0	$0/4 = 0$
3	1	$1/4 = 0.25$	0	$0/4 = 0$
4	0	$0/4 = 0$	0	$0/4 = 0$
5	0	$0/4 = 0$	1	$1/4 = 0.25$
6	1	$1/4 = 0.25$	0	$0/4 = 0$
7	0	0	0	$0/4 = 0$
8	1	$1/4 = 0.25$	2	$2/4 = 0.5$

For each player, we measure the strength of each (in-group and out-group) tie by determining how often the in-group and out-group members allocate tokens to the individual (see Table 4.2). We determine in-group relationships by the ratio ($\mathrm{R_r}$) of the number of tokens received from in-group members ($\mathrm{T_{in}}$) to the number of in-group members ($\mathrm{N_{in}}$) in the round (r), given that the game rules specify one token allocation per round. Thus, $\mathrm{T_{in}} = \mathrm{N_{in}}$ will result in $\mathrm{R_r} = 1$ (a very strong bond), while $\mathrm{T_{in}} = 0$ will result in $\mathrm{R_r} = 0$ (no bond). Out-group relationship, i.e., the ratio of the number of tokens received from out-group members to the number of out-group members, is calculated in the same way as the in-group relationship, with $\mathrm{T_{in}}$ replaced by $\mathrm{T_{out}}$ and $\mathrm{N_{in}}$ replaced by $\mathrm{N_{out}}$. An individual may have a strong bond with both the in-group and out-group members. Thus, in-group and out-group relationships are two separate measures.

In most cases, the comparison between in-group and out-group is made relative to status (low/high/equal status), which is associated with power. Wealthy individuals are often seen as powerful individuals. Thus, wealth aspiration underpins two exchange behaviours: giving to the rich (or seeking power) and giving to the poor (or fairness).

Capraro et al. (2014) suggested that fairness can be the basis on which some individuals interact. According to Tajfel and Turner (1979), interaction may be moderated by the legitimacy (perceptions of fairness) of the status hierarchy. When perceptions of fairness are low, low-status group members challenge the status quo by strengthening in-group reciprocal behaviours and in-group favouritism. High status group members may either enter into intergroup competition or rectify the injustice by outgroup altruism, giving to the poor. In contrast, when the situation is viewed as legitimate, low status group members may seek to enrich themselves individually by making exchanges with rich outgroup members.

Wealth aspiration was measured relative to the wealth of the player to whom the individual allocates a token, that is, the individual's aspiration to associate with the poor (fairness) or the rich (power-seeking). Associating with the poor means allocating one’s token to a low-status individual while associating with the rich means allocating one’s token to a high status individual in the round. The former is considered as being fair, while the latter is considered as seeking power. Thus, participants’ wealth aspiration in round r is calculated based on the allocations made in round $r - 1$, using the formula $WealthAspiration = a_{r-1} / max(A_{r-1})$, where a_{r-1} is the start token (individual’s token before allocation) of the player to whom an individual allocated a token in round $r - 1$, and A_{r-1} is a vector of the start tokens of all the players in round $r - 1$.

Interpersonal ties of trust are built by reciprocity. Relationships between individuals are more trusting when exchanges occur without explicit negotiations between members (Molm et al., 2007; Molm et al., 2013). Trustworthy individuals gain positive reputations; they are likely to be rewarded by other individuals (Yoshikawa et al., 2018), and their actions are more likely to be reciprocated, especially by the in-group members. As shown by De Dreu et al. (2020), expectations of reciprocity can promote in-group interactions, and reciprocation. But powerful reciprocity norms also motivate reciprocation with individuals who are not in-group members.

Interpersonal ties were measured in terms of reciprocation motive, A's desire to allocate a token to another participant B who allocated a token to A in the previous round. This motive was measured by the history of reciprocity in terms of the presence or absence of reciprocity in the previous round indicated as 1 and 0 respectively.

Motives and experience (represented as observed behaviours) form the basis on which individuals in social interaction plan and/or adjust their plans. We refer to the combination of motives and observable differences between individuals as features used to impute strategies in social interaction. In this study, features are group ties, wealth aspiration, reciprocity, status, group identity, and previous allocation. These features form input to the cluster analysis used to determine strategies in the game. Table 4.3 summarises the features while Table 4.4 shows the representation of the features as input to the cluster analysis.

Table 4.3. Features generated from observable differences between individuals and motives in social interaction

Features	**Type of Variables**	**Range/possible values**
Group	Categorical (Nominal)	1 and 2
Status	Real	[0, 1]
Wealth Aspiration	Real	[0, 1]
In-group relationship	Real	[0, 1]
Out-group relationship	Real	[0, 1]
Reciprocity	Categorical (Nominal)	0 and 1
Previous allocation	Categorical (Nominal)	0,1 and 2

Table 4.4. Exchange data motives after pre-processing. Each row represents the participants' data in a particular round, showing (1) information used for tracking purposes (participant number, game and round),(2) features (that form input to the cluster) and (3) the exchange decision (made by each participant in each round).

				Features							
Number of exchange data	Participant No.	Game	Round	Group	Status	Wealth Aspiration	In-group relationship	Out-group relationship	Reciprocity	Previous allocation	Exchange decision
1	1	ENG1	1	1	0.75	0.75	0	1	1	2	001
2	2	ENG1	1	2	0.87	0.87	0.25	0	0	1	010
3	3	ENG1	1	1	1	0.75	0	0	0	0	110
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
7644	8	ING5	40	2	0.92	0.87	0.5	0.25	0	1	000

4.2.3 Strategies

This section identifies unobserved strategies in the minds of players. It does this by finding patterns of association between features of exchange behaviours of each player over the course of the game. Note that each row in Table 4.4 represents a single exchange decision, by one player in one round. The actual decision, recorded in the three-character depiction (see Table 4.1) is recorded in the final column. The "Features" columns record observable features that characterize the motives and game features of the decision. We use cluster analysis to identify and categorize patterns of exchanged behaviours based on the co-occurrence of features. These categories of exchange behaviours represent unseen and implicit strategies.

As shown by Zaki et al. (2020), clustering behaviours can ensure high performance in predicting exchange decisions. Partitioning around medoids (PAM) clustering algorithm in R (Team R Core, 2019) was applied with Gower distance (Gower, 1967) as the distance measure. Although other distance measures, such as Euclidean and Manhattan, can be used (see Gan et al., 2020), Gower distance is very useful and performs well in a domain with mixed data types – categorical and non-categorical data (Akay & Yüksel, 2018; Gower, 1967).

The *silhouette width* also referred to as the silhouette coefficient (Dinh et al., 2019), was used to determine the optimal number of clusters. It measures the within-cluster cohesion and the separation distance between clusters. The silhouette width of a data sample ranges from -1 to 1, where large *s* (near 1) implies well-clustered, a small s (near 0) implies that the data sample lies between clusters, and a negative *s* implies that the data sample has been placed in the wrong cluster. Thus, the higher the silhouette width, the better the cluster.

4.2.3.1 The clustering procedure and results

The optimal number of clusters was determined experimentally by obtaining the silhouette width for two to 20 clusters. Figure 4.5 plots the number of clusters on the x-axis against the silhouette width on the y-axis. The optimal number of clusters is indicated by the highest silhouette width. The plot shows that the optimal number of clusters is six, with a silhouette width of 0.685.

We interpret the clusters by identifying the two dominant exchange decisions that characterize each one (see Table 4.1). Figure 4.4 plots the stacked bar charts of these exchange decisions for each cluster, which show the dominant and recessive decisions that characterize each strategy. Table 4.5 reports the two dominant decisions for each cluster and interprets the strategy represented by this cluster. For example, Cluster 1 is represented by Ingroup-Care strategy (individuals allocate their tokens to in-group members irrespective of their status) while Cluster 6 is represented by Ingroup-Promotion (individuals allocate their tokens and reciprocate only to the poor in-group member).

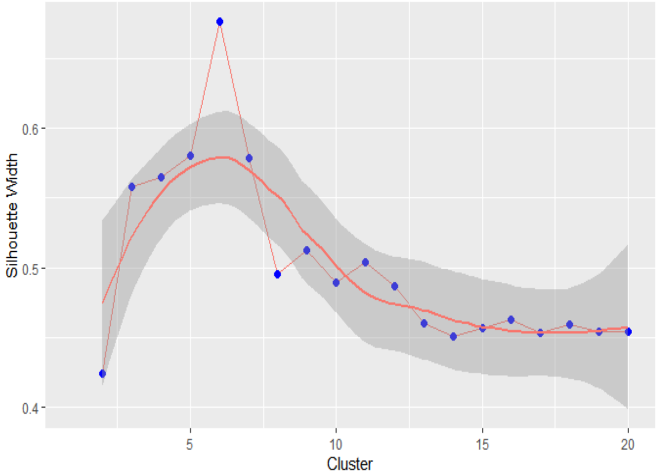


Figure 4.5. Silhouette width for determining the optimal number of clusters. Higher numbers imply a more optimal number of clusters

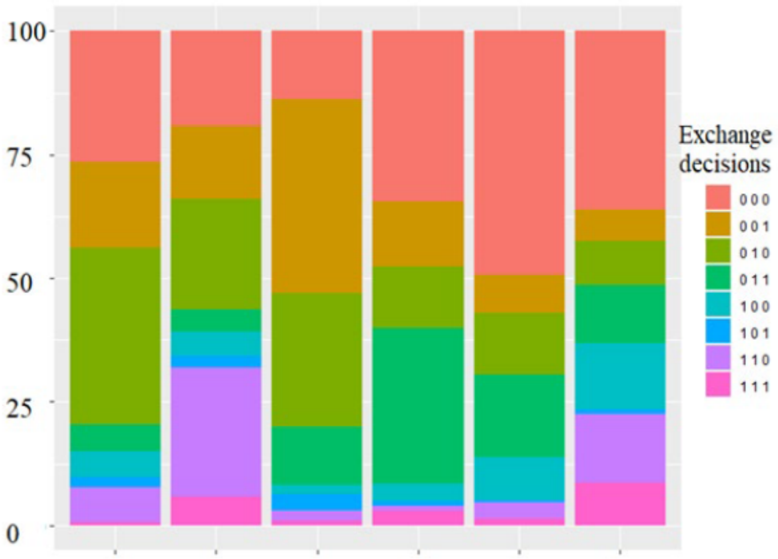


Figure 4.6. The proportion of exchange decisions in each

Table 4.5. The interpretation of clusters as strategies using the combination of the two dominant exchange decisions.

Cluster	The Two Dominant Exchange Decisions made for the cluster	Interpretation of Cluster as a Strategy	Given Name
1	000, 010	Allocate to in-group members irrespective of their status.	Ingroup-Caring
2	010, 110	Allocate and reciprocate to a rich in-group member.	Ingroup-Power-Seeking
3	001, 010	Allocate to a poor out-group member or a rich in-group member.	Outgroup-Reputation
4	000, 011	Allocate to a poor in-group member or a rich out-group member.	Outgroup-Power Seeking
5	000, 110	Allocate to a poor in-group member or reciprocate to a rich in-group member.	Equality
6	000, 100	Allocate or reciprocate to a poor in-group member.	Ingroup-Promotion

4.2.4 Predicting Strategies via Machine Learning and results

Whereas the cluster analysis identifies the strategies on the basis of all the data, our ultimate objective was to predict the strategy that motivated a single exchange behaviour. To this end, we trained an artificial neural network (ANN) (Günther & Fritsch, 2010; Hopfield, 1988; Mehlig, 2019) to predict the cluster membership of each exchange in each round by each player and represent the strategy of the player as that depicted by the predicted cluster. Input to the ANN are Features (see Table 4.4) generated, including the exchange decision, based on an individual's previous round while the output is the cluster to which the past exchange belongs. The ANN was designed to operate in real-time to classify a single exchange decision into one of the identified strategies or as a newly formed strategy. Thus, we use an artificial neural network to compute the probability that an exchange decision in the past round forms part of a complex strategy. Where the probability is below a given threshold (95%, in this study), the decision is categorised as part of a new strategy. Recognising the strategy of an individual during interactive social exchanges will improve the prediction of the individual's exchange decision. This capability has been shown to work in other domains such as traffic congestion prediction (Zaki et al., 2020). The ANN was implemented using the $Deeplearning4j$ framework (Eclipse Deeplearning4j Development Team, 2016). The study used a feed-forward artificial neural network with three layers – an input layer, a hidden layer with four neurons, and an output layer with six neurons, one for each cluster. The final artificial neural network model was trained using a batch size of 40 and a learning rate of 0.01, with 15 epochs. It makes use of the softmax activation and the NegativeLogLikelihood function in Deeplearning4j (Eclipse Deeplearning4j Development Team, 2016) for computing the error which is used to determine the direction of learning. These parameters were determined experimentally.

To train the artificial neural network, data (i.e., features and their corresponding strategies discovered by the cluster analysis) were divided into training and testing sets, each having X (the features) and Y (the strategy) components. Of the data, 70% were used for training while the remaining 30% were used for testing the artificial neural network. Both X and Y were provided to the neural network during training, but only X was provided during testing. The function of the neural network is then to classify X into one of the available clusters, irrespective of the round at which X is produced.

The artificial neural network was evaluated using the accuracy score. This simply counts the number of samples correctly classified. However, accuracy is not a true reflection when the number of samples in each class is not equal or not almost equal (imbalance dataset). To ensure a more accurate measure, the multi-class confusion matrix, detailed in (Tharwat, 2020), was used. Precision, recall and F1 scores were calculated from the confusion matrix. Precision measures the actual number of samples belonging to a class among the total number of samples the artificial neural network identified as belonging to the class. The value ranges from 0 or 0% (no identification) to 1 or 100% (perfect identification). Recall measures the artificial neural network's ability to discriminate samples that do not belong to a particular class. Again, the value ranges from 0 or 0% (no discrimination) to 1 or 100% (perfect discrimination). F1-score – measures the balance between precision and recall. It ranges from 0 to 1. A higher value indicates a better score.

We present the result of predicting the strategies applied by individuals based on the previous exchange decisions. Figure 4.5 shows the learning curve for the artificial neural network.

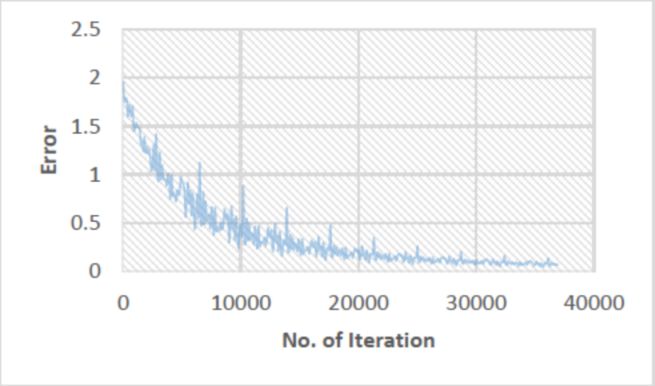


Figure 4.7. Learning curve of the artificial neural networks.

Table 4.6. The confusion matrix of the evaluation on the test set when the number of epochs is set to 5 (all other parameters remain the same as reported). 104 data points belonging to cluster 6 were misclassified, 28 were misclassified as belonging to strategy 4, and 76 were misclassified as belonging to strategy 5.

		Actual Clusters					
Predicted Clusters		**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
	Cluster 1	**497**	0	0	0	0	0
	Cluster 2	0	**157**	0	0	0	0
	Cluster 3	0	0	**300**	0	0	0
	Cluster 4	0	0	0	**287**	0	28
	Cluster 5	0	0	0	0	**524**	76
	Cluster 6	0	0	0	0	0	**39**

Table 4.7. The performance of the artificial neural network on the test set with the number of epochs set to 5. The accuracy is 95% while the precision is 96%.

Evaluation Measures	**Actual Output**	**Percentage**
Accuracy	0.9455	**95%**
Precision	0.9641	**96%**
Recall	0.8788	**88%**
F1 Score	0.8857	**89%**

Tables 4.6 and 4.7 show the performance measures obtained in one of the experiments with the number of epochs set to 5 (all other parameters remained as reported). Tables 4.6 and 4.7 show the confusion matrix and the performance table respectively for the test set. The performance statistics

show that the neural network predicted the strategies with high accuracy of above 94%. This result is similar to that obtained for the training set. The F1 scores of 90% (see the supplementary material) and 89% on the training and test set, respectively, confirm that the performance is not biased towards any cluster. The performance over six runs were also computed to access the reliability of the neural network (see the supplementary material). The final parameter was obtained when the number of epochs was set to 15. The confusion matrices generated internally by the Deeplearning4j (Eclipse Deeplearning4j Development Team, 2016) inbuilt function for the training and test dataset are provided in the supplementary materials here.

4.2.5 Predicting future moves via a Hidden Markov Model

The ANN is trained to identify the exchange strategy that a single exchange behaviour belongs to. It can take all the exchange decisions enacted by the player at round *r* and classify them into various strategies. We now develop a hidden Markov model to predict what each player will do in the next round, *r*+1. The main aim of the hidden Markov model is to predict future moves from a player's past behaviour. Rather than taking the player's past behaviour as input in the form of Features, the hidden Markov model takes the player's past strategies as input and predicts the player's next exchange decision. This was done to improve the predictive accuracy of the model.

A hidden Markov model has hidden states on which the observables are conditioned. For example, an altruistic act can be motivated by empathic concern (FeldmanHall et al., 2015). An altruistic act is an observation while empathy is the state on which the act is conditioned. See (Mor et al., 2021) for a recent review.

A hidden Markov model process (Rabiner, 1989) is characterised by five tuples $\{Q, O, \pi, A, B\}$.

- $Q = \{q_1, q_2, q_3, \dots q_T\}$ is the set of states, each one drawn from N number of possible states, where q_t denotes the state at time $for\ 1 \leq t \leq T$, and T is the maximum number of times an observation was made.
- $O = \{O_1, O_2, O_3, \dots O_T\}$ is the set of observations, each one drawn from M number of possible observations.
- $\pi = \{\pi_1, \pi_2, \pi_3, \dots \pi_N\}$ is the initial state probability for each q in the set of all possible states.
- $A = \{a_{ij}\}$ is the transition probability. This describes the probability of moving from state $i = q_{t-1}$ to state $j = q_t$

- $B = \{b_j\,(O_t)\}$ is the emission probability. This denotes the probability of observation at time t, given state $j = q_t$
- $\lambda = \{\pi, A, B\}$ is the parameters of the hidden Markov model.

For each participant in the VIAPPL experiments, a hidden Markov model takes the previous strategies S as states and the previous exchange decisions O as observations, as shown in **Error! Reference source not found.**. It then predicts the next exchange decision of the participant. Thus, we used a time-homogenous hidden Markov model that uses round-forward chaining time-series cross-validation for training and testing. As shown in Table 4.8, round-forward chaining starts by using data from rounds 1 to r to train the hidden Markov model, which is tested by predicting the exchange decisions in round $r + 1$. Next, it includes the prediction from round $r + 1$ in the training set and predicts round $r + 2$. This process continues until the last round is predicted. The hidden Markov model is retrained after each round of the game and the transition and emission probabilities change per round. This process was implemented to accommodate changes over time in individuals' strategies.

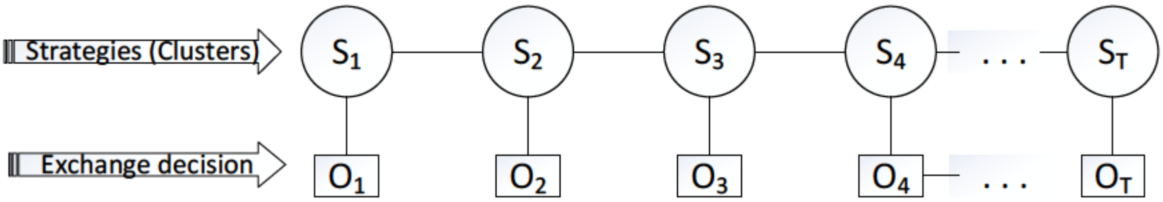


Figure 4.8. The HMM states S and the observations O, where S is the strategies and O the exchange decision.

Table 4.8. The round-forward chaining shows the proportion of the data used for training and testing at each round.

Training Set	**Test Set**
An individual's **strategy** and **exchange decisions** in rounds 1 to 3	The individual's exchange decision in round 4
An individual's **strategy** and **exchange decisions** in rounds 1 to 4	The individual's exchange decision in round 5
…	…
An individual's **strategy** and **exchange decisions** in rounds 1 to 39	The individual's exchange decision round 40

In each round of the forward chaining time-series cross-validation, the hidden Markov model is trained using the Baum-Welch expectation-maximisation algorithm described in the seminal work of Rabiner (1989). Given the exchange decisions and the strategies, training the hidden Markov model implies finding the parameters that would make the exchange decisions most likely. This is also known as parameter estimation.

Using sample observations from five participants, Table 4.9 shows the hidden Markov model evaluation. The hidden Markov model is evaluated using the average accuracy score per round in the round-forward chaining. That is, for each participant in each round, the hidden Markov model predicts the exchange decision (in the test set) of the participant. The average accuracy score per round is the number of exchange decisions correctly predicted divided by the number of participants in that round. The hidden Markov model is also evaluated on its accuracy in predicting the actions that make up the exchange decisions. The individual and combined evaluation are crucial for the application of the model, as it presents the opportunity to plan interventions based on one or more actions during interactive social exchanges.

Table 4.9. Evaluation of the HMM. The data is formulated for demonstration.

Participant ID	Observation in round r	HMM observation prediction for round r	HMM accuracy for the exchange decision	Accuracy score for each definite behaviour		
				Reciprocity score	**Fairness score**	**Favouritism score**
1	001	100	0	0	1	0
2	011	011	1	1	1	1
3	000	000	1	1	1	1
4	011	001	0	1	0	1
5	100	101	0	1	1	0
Average			**2/5 = 40%**	**4/5 = 80%**	**4/5 = 80%**	**3/5 = 60%**

The data is imbalanced, meaning that certain behaviour occurs less often than others. For example, out-group allocation occurs less often than in-group allocation. For this reason, sensitivity and specificity scores are included to measure the performance of the model more accurately. The sensitivity of the model is the percentage of the definite behaviours that are predicted as present among those that are truly present, whereas the specificity of the model is the percentage of the definite behaviour that are predicted as absent among those that are truly absent (Shreffler & Huecker, 2022).

4.3 Result

This section reports on the hidden Markov model performance in predicting a player's exchange decisions given the set of strategies previously applied by the player. Figure 4.9a. and b. plot the learning curves for the transition and emission probabilities of the hidden Markov model in the round-forward chaining cross-validation for rounds 5, 10, 20, 30 and 35. These rounds were chosen to show the differences in the convergence as the data used for training increases. The graph shows that the more rounds used for training, the better and faster the algorithm learns.

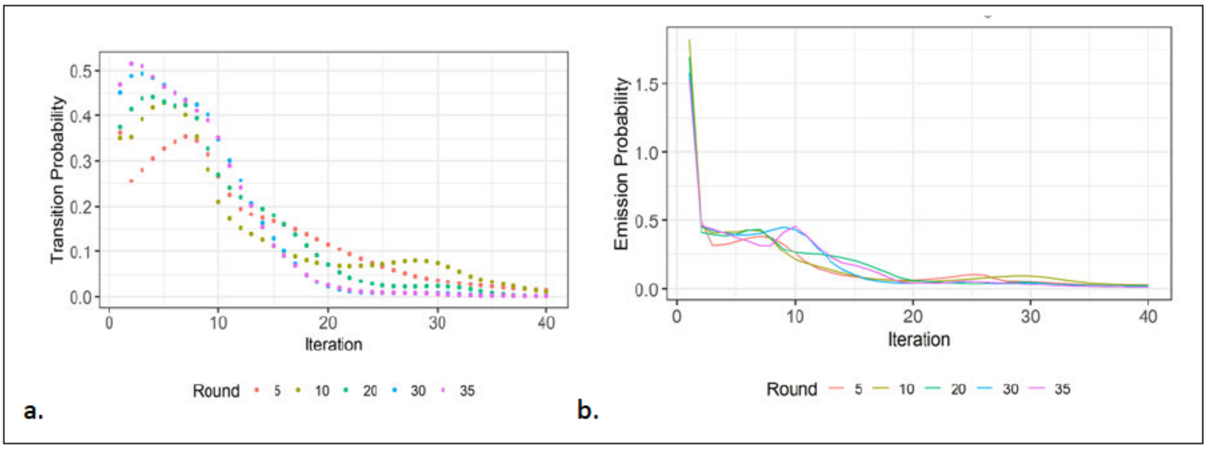


Figure .4.9a. The convergence of the transition probabilities. The convergence is slower when the first five rounds in the forward chaining are used for training and faster when more data (first 35 rounds) is used for training. Thus, the more data used, the fast the convergence. b. The convergence of the emission probabilities, shows the same trend as the transition probabilities.

As shown in Figure 4.10 a., the accuracy of the model on each definite behaviour is better than its accuracy on the exchange decision when treated as one output, i.e., when all the three definite behaviours in an exchange decision are correctly predicted. Reciprocity was predicted with higher accuracy than the others. However, the prediction accuracy decreased over time because reciprocity was very common and very easy to guess. As the model becomes more sensitive to other features, the initial accuracy declines slightly. The model accuracy on *seeking power* versus *fairness* fluctuates over time. However, the accuracy increases over time for *out-group* versus *in-group* favouritism. Overall, the model becomes more accurate over time. Compared to random guesses (Figure 4.8b.), the model accuracy on exchange decisions reached up to 40% while random guesses of exchange decisions stayed below 15% accuracy. Also, the accuracy of the model on reciprocity always stayed above 80% while random guesses of reciprocity were 50 % accurate. Thus, the model performed well.

Since model accuracy is not sufficient to draw a conclusion about the performance of the model. Figure 4.11 shows the sensitivity and specificity graphs. Overall, the specificity of the model is higher than its sensitivity, showing that the model is more confident in predicting the absence of a definite behaviour than the presence of the definite behaviour. For example, the model will predict, with more accuracy, when a player will not reciprocate than when a player will reciprocate. However, there is a gradual increase in sensitivity over rounds. This implies that the model's ability to predict the presence of a definite behaviour (i.e., identify participants that may reciprocate or allocate their tokens to out-group members) increases with an increase in the round number. Thus, more data improves the performance of the model.

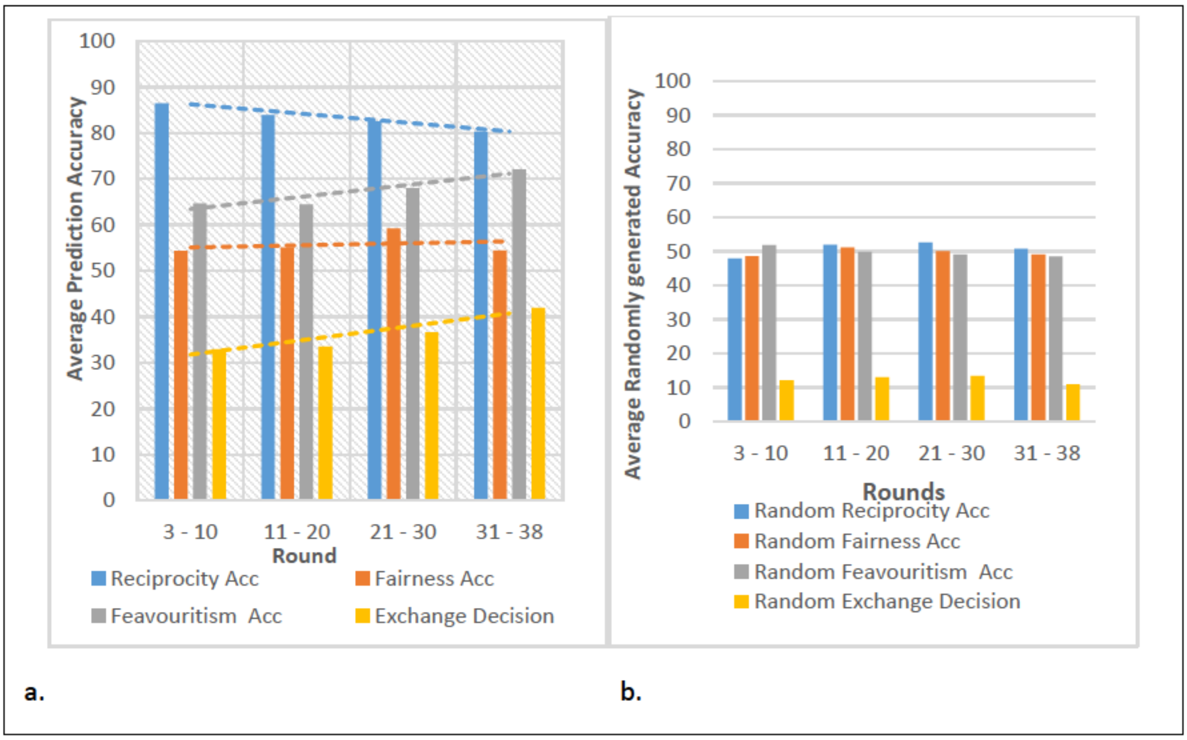


Figure 4. 10. a. The accuracy score of the model where Acc implies accuracy. b. Randomly generated accuracy score showing that exchange decisions can only be predicted randomly at an accuracy below 15% while reciprocity can be randomly guessed with a maximum accuracy of 50%. (The procedure for generating the random guesses and the script is provided in the supplementary instructions)

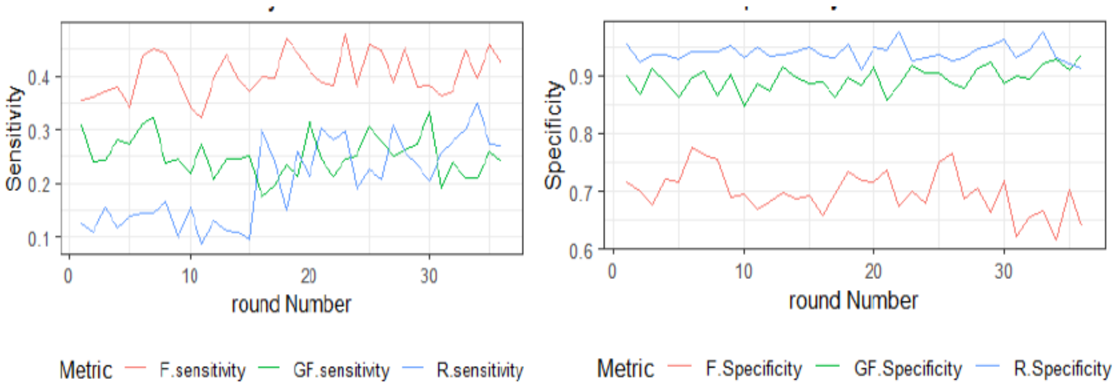


Figure 4.11. The sensitivity (Left) and specificity (Right) of the model, where F implies fairness, GF implies group favouritism and R implies reciprocity.

4.4 Discussion

This study aimed to develop a model for predicting social exchange decisions. There are many competing theories concerning the factors that contribute to decision-making in an interactive social exchange. Social identity theory (SIT) (Tajfel et al., 1971), bounded generalised reciprocity (BGR) (Yamagishi et al., 1999; Yamagishi & Mifune, 2009), and work on interdependencies (De Dreu et al., 2022; De Dreu et al., 2020) provide an understanding of how social context and social relationships impact the exchange decisions at both the individual and group levels. Based on these theories, this study developed a model that leverages the combination of technological advancements, theoretical knowledge and experimental data to predict interactive social exchange decisions. The model is discussed in two phases: the learning phase and the prediction phase.

4.4.1 The Learning Phase

In the learning phase, clustering was performed to discover strategies applied by players. The cluster performance indicated six clusters as optimal for the data provided. Indeed, this may not always be the case, as the potential exists for new strategies to emerge during interactions over a longer period. Moreover, the strategy was determined by the dominant exchange decision behaviours, which could change over time.

The neural network was trained to classify exchange data into one of the strategies. The result shows that this is an easy task for the neural network, as the final model obtained very high accuracy with one hidden layer. To show that the accuracy score is a true reflection of the neural network, the latter was also evaluated and investigated with three different metrics – precision, recall and F1-score. Each of these metrics indicates that the performance of the artificial neural networks is very good. This is also an indication that the correct number of clusters was chosen.

4.4.2 The Prediction Phase

The predictive performance of the model was based on accuracy scores (the ability to correctly predict the presence or absence of a definite behaviour), sensitivity (the percentage of the definite behaviours that are predicted as present among those that are truly present) and specificity (the percentage of the definite behaviour that is predicted as absent among those that are truly absent). Despite the accuracy of the neural network, the prediction of the exchange decision was not an easy task. The result shows that the model performed above average (i.e., above 50% on predicting fairness, above 60% on predicting favouritism, and above 80% on predicting reciprocity) on predicting the definite behaviours but lower on the exchange decisions (a combination of these behaviours as one output). This was anticipated because of the stochastic nature of interactive social exchanges. However, the gradual increase in the accuracy of the exchange decisions shows that the performance is likely to increase during an interaction over a longer period. The performance of the model is considered good, especially when compared to a random guess which produced less than 15% accuracy on exchange decision.

Although the accuracy of the model is highest in predicting reciprocity, the sensitivity and specificity graphs (Figure 4.11) show that the high accuracy after the first ten rounds was not a true reflection of the goodness of fit. The graph suggests that the output of the model is mostly 0 after the first ten rounds, while there were a few reciprocal exchanges. This leads to a very high specificity of above 0.6 (Figure 4.11 Left) but low sensitivity (Figure 4.11 Right). However, the sensitivity increased over time, which shows that the model improves over time.

The model performance on predicting token allocation to a rich or poor player did not improve over time. The study concludes that this performance was a result of status (poor or rich) changing frequently during the game from which the data was collected. However, the performance of the model on out-group vs in-group favouritism increased over time.

4.5 Limitations and Future work

Like most laboratory experiments, VIAPPL allows interaction in a very minimal context. Thus, the range of possible behaviours that can be learned is limited. Therefore, the model may not be generalisable to real-world contexts involving a higher degree of complexity. Also, due to the stochastic nature of the hidden Markov model, there is no guarantee that the performance of the model will remain the same in different settings. Finally, as with other artificial neural network-based models, a change in the model parameter, for example, the learning rate, and the number of iterations, may lead to a different outcome in the prediction of the strategies. This may lead to a change in the prediction accuracy of the model.

Future work will apply the model in a live experiment to test the performance of the co-evolution of the artificial neural networks and the hidden Markov model.

4.6 Conclusion

Understanding how human decisions are influenced by motives and actions is important for predicting human social exchange decision-making. By predicting human exchange decisions in interactive social exchanges, artificial agents can potentially help to resolve conflicts, facilitate cooperation, and promote prosocial behaviours. This study aimed to develop a model to predict exchange decisions in an interactive social exchange context. The model was trained using secondary data from game-like experiments. As part of the model, data clustering was performed to group different behaviours into a finite number of strategies. The strategies employed by players were interpreted and used and used for predicting exchange decisions. The model performance increases over time, which suggested that a better model could be realised with an increase in the number of interactions over time represented by the training data. The study provides a novel method of integrating theory-driven data into machine learning models and means by which adaptive agents can be integrated during human-agent cooperation to predict human decisions and react to them. The ability of artificial agents to predict and react to human decision is vital to the ongoing endeavours in enhancing human-agent cooperation. Thus, this study also contributes to the existing literature on cooperation in interactive social exchanges. Lastly, the study suggests the evaluation of the model in real-time experiments to test how agents applying the model can perform in a real-time interactive social exchange.

The next chapter reports on testing the model’s performance in real-time by applying it to engineer positive outcomes during interactive social exchanges. Specifically, the model was integrated into an experimental context involving humans and bots (referred to as agents), such that bots using this model are tested for their ability to reduce in-group favouritism amongst humans.

CHAPTER 5. INVESTIGATING THE ROLE OF OUT-GROUP ALTRUISM FOR REDUCING IN-GROUP FAVOURITISM IN AN INFORMATION-SCARCE ENVIRONMENT: AN AGENT-HUMAN INTERACTIONAL APPROACH

Abstract

In-group favouritism and intergroup conflict are mutually reinforcing during interactive social exchanges. This study focuses on devising means to break this cycle of influence. It asks the question of whether or not out-group altruism can reduce in-group favouritism, and, if so, to what extent.

In two studies – Study 1 and Study 2 – this research used novel methods of agent-human interaction over time via a computer-mediated experimental platform to investigate out-group altruism, a prosocial behaviour, for reducing in-group favouritism during interactions over time. Out-group altruism was introduced via: (i) non-adaptive artificial agents whose behaviour was pre-programmed to be altruistic (Study 1), and (ii) adaptive artificial agents whose altruistic behaviour was informed by a machine learning algorithm (Study 2). Turing tests were performed via a rating task to ensure that the observed behaviour was not a result of the participant's awareness of the agents.

The findings in Study 1 show that out-group altruism, via non-adaptive artificial agents, produced an ironic effect. Rather than the agents setting up out-group altruistic norms, in-group members maintained group identity by strengthening in-group favouritism and treating agents as out-group members. Similarly to the findings in Study 1, Study 2 shows that in-group members strengthened in-group favouritism during the initial stages of interactive social exchanges. Contrary to the findings in Study 1, adaptive agents in Study 2 were able to weaken in-group favouritism over time by maintaining a good reputation with both the in-group and out-group members, who perceived agents as being fairer than humans and rated agents as more human than humans.

5.1 Introduction

The early chapters (1 and 2) pointed out the role of interdependencies in reducing in-group favouritism. Subsequently, in Chapter 4, interdependencies and motives derived from theoretical knowledge were incorporated to develop and test a co-evolutionary model for predicting exchange decisions during interactive social exchanges. This model presents the opportunity to develop adaptive agents – agents that use the model to understand, predict and strategically interact with humans – to reduce in-group favouritism among humans during interactive social exchanges. This chapter aims to investigate adaptive agents in an experimental context for reducing in-group favouritism during interactive social exchanges. Firstly, it establishes a base model by developing and testing the ability of rule-based agents to reduce in-group favouritism in a similar context.

Most literature on in-group favouritism (see Chapter 1) aims to explain the factors and processes that lead to in-group favouritism, not how it may be reduced. The present study builds on the acknowledgement that the relationship between in-group favouritism and intergroup conflict is a cycle that is mutually reinforcing (Bornstein, 2003; Rapoport & Chammah, 1970). This was pointed out by Bornstein (2003), who said that "cooperation within groups inevitably contributes to the escalation of conflicts" (p. 130), and has been discussed by De Dreu et al. (2020) and others in the interdependence context. This study proposes that intervening in the interactional interdependencies (e.g. reciprocity and trust) can break the cycle. Thus, it examines non-adaptive and adaptive agents implemented via rule-based and machine learning algorithms respectively, to intervene in the interactional interdependencies.

Research (Bornstein, 2003; De Dreu et al., 2020; Greenwald & Pettigrew, 2014; Rapoport & Chammah, 1970) has shown that in-group favouritism forms the basis for intergroup bias. The work of Greenwald and Pettigrew (2014) suggests that reducing in-group favouritism may imply reducing intergroup bias. Evidence from literature (Janssens & Nuttin, 1976; Rabbie & Wilkens, 1971) has shown that the strength of intergroup bias, vis-à-vis intergroup discrimination, is an indicator of the interdependence between groups. Also, borrowing from the literature on intergroup contact (Pettigrew & Tropp, 2006), originally poised by Allport (1954), the strength of intergroup bias is dependent on the nature and outcomes of the interactions between groups (Pettigrew, 1998; Rabbie et al., 1974; Wilson & Miller, 1961) and the perceived similarity between these groups (Jetten et al., 1998).

Although Allport (1954) pointed out the conditions required for the effectiveness of positive intergroup contact, which include equal status, common goal, and intergroup cooperation, this present research does not necessarily focus on all the conditions. However, it taps from the intergroup discrimination-circumventing process originally posited by Pettigrew (1998) – and extended to form vicarious contact – that the observation of an in-group member interacting with out-group members can reduce intergroup bias and in-group favouritism. Pettigrew (1998), however, noted that a negative experience could lead to more discrimination.

Relating Pettigrew's assertion to this present study presents a promise that observing an in-group member's out-group altruism could reduce, if not extinguish, in-group favouritism in the minimal group paradigm. Nevertheless, evidence from recent experiments has shown variations in the extent of intergroup discrimination, suggesting that it does not necessarily need to be internalised by all or even the majority of the group members, but may be driven by a minority of individuals (Kranton et al., 2016, 2017).

Considering in-group favouritism as the basis for intergroup bias, as well as the volume of research and empirical evidence supporting in-group favouritism, recent research (Durrheim et al., 2016; Lee et al., 2012; Spielman, 2000) has focused on investigating how in-group favouritism emerges during intergroup interaction over time. The present research investigates ways of breaking the cycle of influence that may be capable of reducing in-group favouritism.

Some studies (for example, Tajfel & Turner, 1979; Tajfel & Turner, 1985) have suggested reasons for in-group favouritism, as well as theories behind in-group favouritism. These suggestions, however, rely on different assumptions, underlying processes and causes of in-group favouritism (Schellhaas & Dovidio, 2016), but largely ignore the power of repeated interaction. This study looks at the relevant suggestions and investigations from both theoretical and empirical perspectives; presents its investigation approach, methodology and hypothesis; presents and discusses the results; and makes some suggestions for future work.

5.2 Reducing In-group Favouritism: Theoretical Perspectives

Social identity theory (SIT) (Tajfel & Turner, 1979, 1985) postulates that in-group favouritism may occur by the mere fact that individuals are categorised as in-group or out-group members. Researchers have, therefore, derived interventions from this perspective and aimed to change individuals' level of categorisation. This aim has led to decategorisation and recategorisation theories. Hewstone et al. (2002) explained that "[d]ecategorisation seeks to eliminate categorisation via two mutual and reciprocal cognitive processes" (p. 589). The aim is therefore to make distinctions between out-group members (this is termed inter-individual *differentiation*) and see the out-group members in terms of their uniqueness and in relation to the self, not the group (i.e. thus increasing personal identity salience while decreasing the group identity salience – this is termed *personalisation*). While decategorisation focuses on *differentiation* and *personalisation, re*categorisation seeks to alter the 'initial' categorisations used, and to replace the 'us and them' categorisation with a 'we' categorisation. In other words, recategorisation seeks to move from initial subordinate categorisation to superordinate categorisation.

Studies have tested the efficacy of decategorisation and recategorisation for reducing in-group favouritism (see the following recent studies, Fritzlen et al., 2019 ; Jung et al., 2019). Although both these studies did not use the minimal group paradigm, they provided insights into categorisation threat – the fear of being categorised against one's will. Individuals are willing to accept a group when the group attributes are positive but unwilling if the group is perceived negatively. It might, therefore, be necessary to promote individuals' positive perception of the superordinate group in order to increase acceptance of the superordinate categorisation.

Although both decategorisation and recategorisation can reduce in-group favouritism, eradicating or replacing initial or original categorisations may sometimes be impossible or could threaten individuals' need for adaptation and differentiation. Thus, there is a need for alternative ways of reducing in-group favouritism.

Pettigrew and Tropp (2006) meta-analysis review of literature on contact theory, based on over 90,000 participants in 203 studies, reported that the intergroup contact hypothesis was significantly associated with a decrease in intergroup bias, which is associated with in-group favouritism (Greenwald & Pettigrew, 2014).

Although direct intergroup contact has been proven beneficial in reducing intergroup bias, it might be challenging to be justified pragmatically in some situations due to, for example, the limitation placed by borders between groups such as countries (Dovidio et al., 2011). Thus, various forms of indirect contact have been proposed that do not necessarily require in-group members to have direct contact with out-group members in order to reduce intergroup bias. These forms of indirect contact, which include extended, vicarious and imagined contact, have been proposed to reduce intergroup bias, and hence reduce in-group favouritism.

Whereas proponents of extended contact argue that the knowledge that an in-group member is interacting with out-group members is enough to reduce intergroup bias (Wright et al., 1997), proponents of imagined contact posit that imagining in-group members interacting with out-group members is enough to reduce intergroup bias. While extended and imagined contacts involve knowledge and imagination respectively, vicarious contact suggests that mere observation of in-group members interacting with out-group members is enough to reduce intergroup bias. Thus, vicarious contact emphasises the effect of the actions of an in-group member on the rest of the in-group members, when the action is directed to the out-group member. These effects can be positive or negative. Positive effects reduce in-group favouritism, while negative effects increase in-group favouritism (Dovidio et al., 2003).

The effectiveness of these various forms of contact theories has been empirically tested; for example, see (Veldhuis et al., 2014) and (Dovidio, Love, et al., 2017) for a review. This study focuses on *repeated vicarious contact* in reducing in-group favouritism in the minimal group paradigm. It is essential to know what interactional strategy is required to eliminate in-group favouritism. For instance, it is not clear whether increased interaction between some in-group and out-group members would decrease in-group favouritism in a minimal group paradigm. In other words, it is not clear whether increased inter-group allocation by some members reduces in-group favouritism by changing other members' perceptions of out-group members.

5.3 Reducing In-group Favouritism: Evidence from Empirical Studies Involving Allocation

The two popular theories that build on the minimal group paradigm – SIT and the bounded generalised reciprocity theory (Yamagishi et al., 1999; Yamagishi & Mifune, 2009) – assume that individuals are prone to perceive affiliated interests with in-group members as a result of categorisation, with bounded generalised reciprocity adding cost/benefit as yet another reason. However, most interventions that aim at reducing in-group favouritism focus on reshaping individuals' cognitive representations of group membership and how they relate their personal interests to other in-group and out-group members. Apart from reducing in-group favouritism via categorisation (for example, Crisp et al., 2006), a few studies also focus on changing incentives.

Categorisation and recategorisation have received attention both in theory and experiments. In one of the experiments, the participants in Crisp et al. (2006) were categorised arbitrarily into four different subordinate groups and one superordinate group. Crisp and colleagues' experiment shows that the role of categorisation in reducing in-group favouritism is dependent on the individuals' perceived importance of the categorisations (i.e. how important the superordinate group is to the individual).

Crisp et al. (2006) findings suggest that in-group favouritism is supported by the illusion of self-interest. People tend to think that the costly contributions they made to benefit their group are actually in their self-interest. They reason that they are part of their group and so, whatever help they render to their group, implies help to themselves. Nevertheless, they neglect the fact that, in many cases, their contributions are higher than their return.

Several studies have reported the effect of norms on in-group favouritism (Dang et al., 2019; Hertel & Kerr, 2001). The studies show that priming some norms reduces in-group favouritism, while others may increase in-group favouritism. In a real sense, intergroup bias is moderated by the participants' perceived importance of the norm.

For example, participants in (Dang et al., 2019) competed in a task comprising ten questions. These questions tested whether the in-group or out-group members had a better problem-solving ability. A self-esteem norm was primed by a negative (versus positive) evaluation of in-group performance on the test. The participants then performed another task. In the task, participants were told to allocate money (an amount between 0 and 1,000 inclusive) to in-group members and the rest of the money would belong to out-group members. The study reported less in-group favouritism for the negatively evaluated in-group compared to the positively evaluated in-group. Apart from using the minimal group paradigm, the study shows that individuals' perception – in this case, self-esteem – affects how individuals reduce bias. However, it does not consider that repeated interaction – for example, the repeated allocation – could change the influence of social self-esteem on the participants. Thus, it misses the dynamics of social interaction.

Similarly, Hertel and Kerr (2001) evaluated the impact of loyalty norms and equality norms on in-group favouritism. A loyalty norm was primed with words stressing positive aspects of loyalty, such as *team spirit*, to one's group, while the equality norm was primed with words such as *just* and *fair*. Participants in Hertel and Kerr study performed allocation tasks using the matrices developed by Tajfel and colleagues (Tajfel et al., 1971). In each matrix, there were two rows of numbers. Each number was an alternative point allocation for two persons. Hertel and Kerr (2001) reported a positive correlation between loyalty and in-group favouritism. In their report, priming of loyalty promotes in-group favouritism while priming of equality decreases in-group favouritism. The study supports the notion that "in-group favouritism might be a consequence of available social norms" (Hertel & Kerr, 2001, p. 321). Indeed, the consequences of available social norms are shaped by individual perceptions of the norms. However, this perception may change over time, thus repeated interaction (or an allocation task with the same matrix) may be necessary to evaluate the strength of the available norm.

In summary, although contact work (Allport, 1954; Pettigrew, 1998) shows the value of interaction in affecting strategies that could reduce in-group favouritism, most theoretical and experimental intervention devised to reduce in-group favouritism ignores the fact that we need an interactional strategy to eliminate bias rather than a categorisation strategy. This neglects that interdependence, often observed during interactions over time, which may develop amongst group members, could strengthen the evolution of in-group favouritism (Kelley & Thibaut, 1978)

5.4 The Present Research

This present research introduces some methodological advances with regard to experimental design and software that enable us to experimentally investigate ways of reducing in-group favouritism. The research expects that a simple but powerful interactional strategy is capable of reducing in-group favouritism. It acknowledges that "the most central, useful, powerful set of social psychological ideas is the triumvirate of imitation, conformity, and social norms" (McDonald & Crandall, 2015, p. 147). Imitation is instrumental in norm formation (McDonald & Crandall, 2015). Imitation leads to regularities of behaviour (i.e. it is recurrent), which in turn lead to the emergence of norms to which members may conform (Opp, 1982). Of course, all recurrent behaviour does not necessarily become a norm. However, 'recurrent' is a possible condition that will necessitate norm formation via imitation, and repeated interaction is a necessary condition for any recurrent behaviour.

According to Titlestad et al. (2019), the development of cooperation is best predicted by how norms are formed once social identities emerge. Titlestad et al. (2019) pointed out that, within groups, individuals follow nearly identical rules when deciding to engage in cooperative behaviour. As such, individual behaviour is seen to converge due to the quality of the social interactions within the group (Titlestad et al., 2019).

Using a public goods game (PGG) design, Titlestad et al. (2019) showed that cooperative behaviour is not a static feature, but rather emerges over time within groups as a result of social interaction. However, there are vast disparities between groups surrounding the extent to which optimal cooperation may be achieved. While some of these emergent norms may be explained by categorisation, for the most part, these norms emerge as a result of the interaction (Titlestad et al., 2019)

This research considers imitation as a social influence. Bandura (1986) pointed out that social influence may take the form of vicarious learning. Thus, social influence, in this regard, overlaps with vicarious contact. It provides a ground that "effective behaviour by one person may well be repeated by others through a process of observational learning, while choices that lead to undesirable consequences may be avoided by others" (Fulk et al., 1990, p. 122).

The primary concern of this study, and most experiments involving social stimuli, is how to ensure that the social stimuli providers remain consistent throughout the experiment. Otherwise, it is difficult to attribute the observed behaviour to the effect of social stimuli. The current study, therefore, employs artificial agents as stimuli providers. The use of artificial agents is in accord with MacDorman and Ishiguro's (2006) assertion that, by being programmable, controllable and replicable, artificial agents are better than human actors in terms of social experimental stimuli. The agents thus fulfilled the role of experimental 'stooges', allowing us to investigate the effect of their behaviours in promoting out-group altruism.

5.5 Method

5.5.1 Virtual Interaction APPLication (VIAPPL) – An Integrated Environment for Artificial Agents and Human Interaction

The research examines strategies for reducing in-group favouritism by integrating artificial agents (computer programmes, henceforth interchangeably called agents) into Visual Interaction Application, such that it allows interaction within and between agents and humans – very much like a computer game.

This research work is presented in two studies: Study 1 and Study 2, which employ different strategies to weaken in-group favouritism during interactive social exchanges. Whereas agents in Study 1 are pre-programmed to behave altruistically by allocating their token (wealth) to the out-group members randomly, agents in Study 2 learn and predict humans' interactive exchange decisions, and then devise a strategy to reduce in-group favouritism based on the predictions. Thus, Study 1 is a base study while Study 2 was conducted to overcome the shortcomings of Study 1. This was done via the application of the co-evolutionary model developed in Chapter 4.

In Study 2, agents use the model to predict exchange behaviour and devise behavioural strategies to reduce in-group favouritism.

5.6 Hypothesis

The hypotheses, which were formulated and tested, are described in the sections below.

5.6.1 Imitation Effect

The study proposed that imitation is a simple but powerful interactional strategy capable of reducing in-group bias:

Hypothesis 1: ***In-group favouritism***: The agents will produce higher reductions in in-group favouritism amongst both in-group and out-group members over time via imitation by human participants.

5.6.2 Turing Test

Due to the experimental nature of the study, minor deception was necessary as participants were led to believe, at the beginning of each game, that all the players were human. This low-risk form of deception was necessary for this study, as the study explored the effects of the strategy employed by the agents in promoting out-group altruism using minimal groups created in the experimental setting.

There is, therefore, a need to perform a Turing test. A Turing test determines whether or not a human can differentiate between the artificial agent and another human in terms of a specified objective in a specific domain. Hence, it was necessary to check whether or not participants were able to detect that one or more of the players were artificial agents. The assumption is that participants' behaviours will be more natural if they do not know that artificial agents are among the players. Thus, participants were presented with two rating tasks at the end of each game. The first was a *fairness* rating, while the second was an *agent identification* rating. The former precedes the later to avoid a possible priming effect on the former.

Since the participants were not informed of the agents' altruistic behaviours, participants were likely to view agents' altruistic behaviour (i.e. the allocation of tokens to out-group members) in the first few rounds as being inclusive, thus an act of fairness. However, the perception of fairness among agents' in-group members was likely to reduce over time.

Hypothesis 2a: ***Fairness***: The agents will be rated as fairer than humans, especially by the out-group members.

Hypothesis 2b: ***Agent identification*****:** There will be no difference between the ratings received by agents and those received by humans in identifying the agents amongst participants.

5.7 Study 1 – Rule-Based Agents

The strategy in Study 1 was to pre-programme the artificial agents to exhibit out-group altruism by associating with (i.e. allocating their tokens to) the out-group members.

5.7.1 Sample, Design and Group Assignments

5.7.1.1 Sampling

The sample (N = 400), comprised 280 students and 120 artificial agents. The 280 students, henceforth referred to as participants (149 female, 131 male; 265 Black, 10 Indian, 4 coloured, 1 white) were from the University of KwaZulu-Natal, South Africa. The participants (mean age = 20.41 years, SD = 2.83) provided written informed consent to participate in the study, which had been approved on the 11/01/2019 by the Human Sciences Research Ethics Committee of the University of KwaZulu-Natal with a protocol reference number HSS/2210/018D (see Appendix 1)

A non-probability, convenience sample method was chosen for two reasons: (i) due to the large population of students available, and (ii) the focus was on behavioural consistency, not on generalisability. The study observed consistencies in the behaviour of participants' who had been randomly assigned to a specific experimental condition, for example: Do participants imitate agents' behaviours when the number of agents equals the number of participants, but not when the number of agents is fewer than the number of participants?

Each participant was stationed in front of a computer (described in Chapter 4, Figure 4.3) in which all participants, including artificial agents, were represented with avatars, specifically, circles arranged on their screens. However, the participants were unaware of the presence of agents.

5.7.1.2 Design

There were five conditions determined by the number of agents in each group. This number is referred to as the *dosage* of altruistic agents examined for promoting out-group altruism. These conditions are shown in Table 5.1.

The study conducted by Durrheim et al. (2016) has shown that equal status (i.e. equal allocation of tokens at the start of each game) promotes in-group favouritism. That study formed a baseline condition for the present study, which replicated the design but with the addition of artificial agents, as shown in Table 5.1.

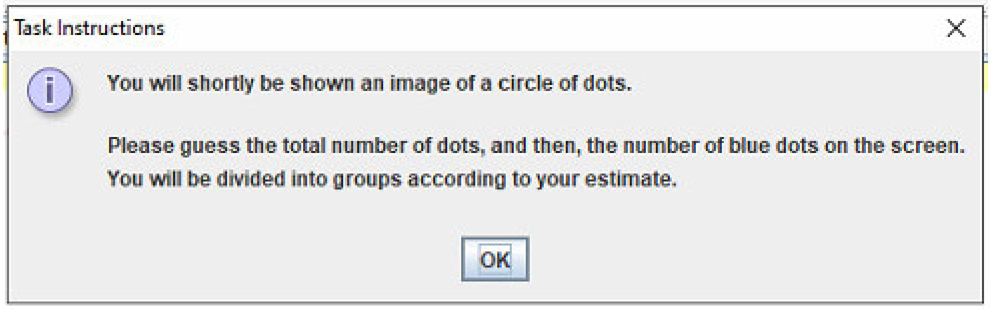
Table 5.1. Conditions under which in-group favouritism will be examined.

Condition	**Short-form**	**Description**
Condition 1:0	1:0	One agent in one group, and no agent in the other group
Condition 2:0	2:0	Two agents in one group, and no agent in the other group
Condition 1:1	1:1	One agent per group
Condition 2:1	2:1	Two agents in one group, and one in the other group.
Condition 2:2	2:2	Two agents per group

There were 50 games, ten per condition; each game had eight players and one of the conditions defined in Table 5.1. Since equal status has been shown to promote in-group favouritism (Durrheim et al., 2016), the players were allocated equal tokens – 30 tokens – representing the player's wealth at the start of the game. The games were played in the laboratory. Confederates sat at each workstation controlled by artificial agents. Thus, each game involved eight participants in the laboratory.

5.7.1.3 Group assignment

At the start of each game, the players were randomly assigned to two different groups represented by the colour of their avatar on the screen. However, participants were told that the assignment was done based on the dot estimation task they performed prior to the interactive exchange game. On login, participants see the message in Figure 5.1.



Task Instructions

You will shortly be shown an image of a circle of dots.

Please guess the total number of dots, and then, the number of blue dots on the screen.
You will be divided into groups according to your estimate.

Figure 5.1. Assignment task instruction message.

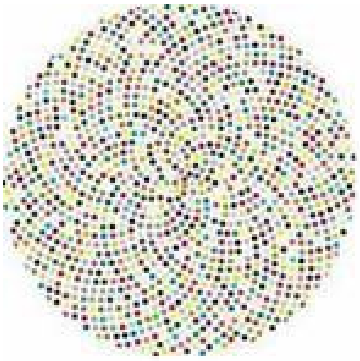


Figure 5.2. The circle used for the dot estimation task.

A circle with some blue and red dots (see Figure 5.2) was displayed on each participant's computer screen for 20 seconds. Participants were asked to estimate the total number of dots on the circle, after which they were asked to estimate the number of blue dots. Afterwards, the following message was displayed on each participant's screen:

> *Based on your estimate, you have been placed in a group with players who had similar estimates.*

Once the players had been placed in a group, they were instructed to allocate a token every round to any one of the eight players. For each game, a trial run (Trial 1) was conducted where each player played for two rounds before the actual game. This was done to ensure that players understood what was required of them. Each (actual) game continued over 30 rounds, and recorded the decision of each participant's allocation either to themselves or to another player within the network. However, participants were not informed of the last round beforehand. Thus, the end-of-game effect was avoided.

The data for each game were partitioned into five waves of six rounds each. The tokens given to self, in-group and out-group were summed for each wave and each group, resulting in a six twin--cells design representing a Visual Interaction Application game played over 30 rounds by eight players under one of the specified conditions in Table 5.1. The partitioning was necessary in order to understand how in-group favouritism changes over a specified window

5.7.2 Ethical Considerations in Sampling and Data Collection

Full ethical approval was granted by the UKZN Humanities and Social Sciences Research Ethics Committee for this study (Appendix 1). Vulnerable individuals did not participate in the study. Participants were eighteen years and above and were recruited through the use of an advert placed on the university notice boards, as well as in person, by the experimenters and research assistants.

An information letter (Appendix 2) was given to all participants, and they were encouraged to ask questions if they experienced any misunderstanding. In addition, participants were asked to sign an informed consent form (Appendix 3), stipulating that: (i) they understood that participation was voluntary and confidential, and (ii) they were able to withdraw from the study at any point.

Participants were required to scan their fingerprints to take part in the study. This was done in order to prevent them from participating in more than one game. The behavioural data were not linked to any of their personal information, such as their name or student number. Furthermore, participants had to register a Visual Interaction Application account in order to log into the game. This required their email address and name. Again, this information was in no way tied to their game data but only used to prevent unauthorised participation.

Participants were given a cash payment as an incentive and partial compensation for their time, expenses accrued and effort involved in taking part in the study. Each token in the game was valued at R1. Each participant received an incentive that coincided with their final token balance. The average cash incentive was ZAR30.00 per participant.

Participants were debriefed after they participated in order to minimise any potential stress or harm. Participants were debriefed by informing them that their assignment to a group was allocated randomly.

The Visual Interaction Application data from the games were stored on the main server in the Psychology Laboratory, and the demographic data were collected through LimeSurvey, with the data stored on the Psychology Laboratory's administration profile, which is not accessible to any third party. The Psychology Laboratory requires an alarm code and key to gain access to the room. Furthermore, the LimeSurvey profile requires the use of a username and password in order to gain access; these are not freely available to third parties and are only known to the researchers. Moreover, the demographic data are not tied to the experimental data.

5.7.3 Measures

5.7.3.1 Independent variables

The independent variables were:

- **Group**: Consisting of two levels: Group 1 and Group 2.
- **Condition**: There were five conditions (Condition 1:0, Condition 2:0, Condition 1:1, Condition 2:1, and Condition 2:2).
- **Time**: The five waves of six rounds each.

5.7.3.2 Dependent variables

The dependent variables were in-group favouritism, fairness rating and agent identification rating.

In-group favouritism

Equation 1 specifies the formula for in-group favouritism. At each wave w, IF_w computes the ratio of in-group to out-group giving for each player over the six rounds of wave w. This ratio could then be examined to see if it increased or decreased over time across experimental conditions. In-group, out-group and self-giving were count variables. Each value of these variables is in the range 0 to 6 (and jointly summed up to 6) in each wave. However, self-giving could be best described as a self-favouring strategy and not a group-favouring strategy. Hence, self-giving was not counted as in-group favouritism and was not used in the calculation. Also, the data produced by artificial agents were removed, since the interest of the study was to investigate the effect of the agents' pre-programmed behaviours in promoting out-group altruism.

Giving Equation 5.1, in-group favouritism at wave w is in the range [0, 1], t takes value from the range [1, 30] specifying the actual round, while $tmin$ and $tmax$ take values from (1, 7, 13, 19, 25) and (6, 12, 18, 24, 30) respectively, thus specifying the minimum and maximum t for each wave.

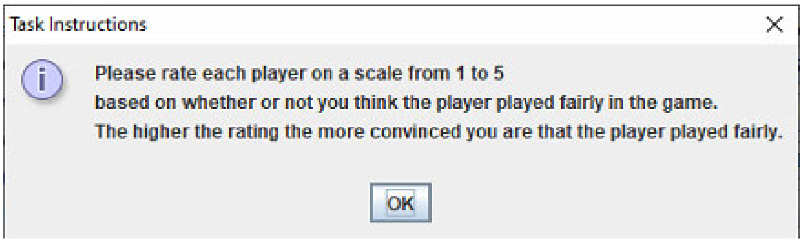
$$IF_w = \frac{\sum_{i=tmin}^{tmax} IG_i}{\sum_{i=tmin}^{tmax}(IG_i + OG_i)}$$

Equation 5.1. In-group favouritism calculation (adapted from Durrheim et al., 2016)

Where IF_w= in-group favouritism at wave w, IG_i= in-group giving at round i, and OG_i= outgroup-giving at round i, IG_i is either 1 or 0. It is 0 when the token is allocated to an out-group member or oneself, (i.e. self-giving) and 1 when the token is allocated to an in-group member. Also, OG_i is 0 when the token is allocated to an in-group member or oneself, (i.e. self-giving) and 1 when the token is allocated to an out-group member.

Fairness rating

Although participants were not given any specific definition of fairness, fairness can be thought of as non-discriminatory behaviour. This current study measures participants' perception of fairness. In order to evaluate each participant's view of other players, the message shown in Figure 5.3 was first displayed on each participant's screen.

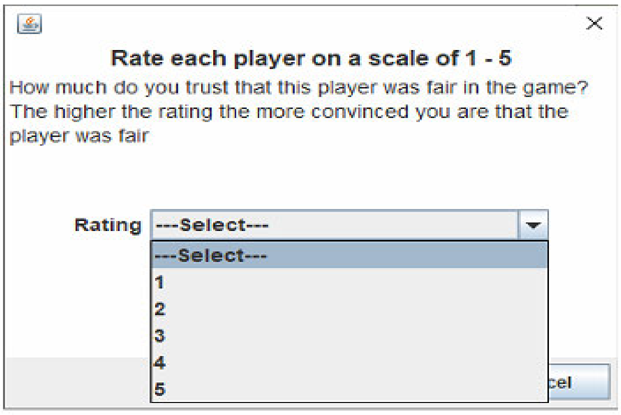


Task Instructions

Please rate each player on a scale from 1 to 5
based on whether or not you think the player played fairly in the game.
The higher the rating the more convinced you are that the player played fairly.

Figure 5.3. Fairness rating instruction.

To rate a particular player, the participant had to click on the avatar representing the player of interest. The message shown in Figure 5.4 was displayed when the participant clicked on the avatar.



Rate each player on a scale of 1 - 5

How much do you trust that this player was fair in the game? The higher the rating the more convinced you are that the player was fair

Figure 5.4. Fairness rating question.

Agent identification rating

For agent identification rating, the same procedure for fairness rating was followed. However, the following message was displayed on each participant's screen.

One or more of the players are computer programs and not humans. Please rate each player on a scale from 1 to 5 based on whether or not you think the player is a computer program. The higher the rating, the more convinced you are that the player is a computer program.

Also, the rating procedure was the same as the fairness rating except that the following message was displayed when a participant clicked on the avatar representing the player of interest.

Do you think that this player is a computer program? The higher the rating, the more convinced you are that the player is a computer program.

5.7.4 Data Analyses

5.7.4.1 Multi-level analysis

The data are nested with conditions within replicated games, and individual responses within waves, thus, the assumption of independence is violated. Consequently, analytical methods, such as analysis of variance (ANOVA), would not be appropriate. Multi-level modelling is apt for hierarchical data (Quené & Van den Bergh, 2004). It is a robust procedure that can handle moderate hetero-scedasticity and violation of sphericity. Thus, compared to ANOVA, it provides more power in estimating the effects.

Furthermore, multi-level analysis is essential to account for random effects – unexplained extraneous factors – that affect the dependent variables. Although these random factors may not have any theoretical interest, it is essential to account for them and the degree to which they influence the dependent variables (Kozlowski et al., 2013). All the analyses were performed in R, using different packages such as the *nlme* for a generalised linear mixed-effects model. These analyses can be found in the online supplementary materials via the link: OSF | A CO-EVOLUTIONARY APPROACH TO DATA-DRIVEN AGENT-BASED MODELLING: SIMULATING THE VIRTUAL INTERACTION APPLICATION EXPERIMENTS

5.7.4.2 Descriptive analysis

Descriptive statistics of the dependent variables are presented. This analysis provides information about these variables and a graphical view of how they relate to each other in different conditions. Confidence intervals were calculated using the bootstrap method. This is appropriate for large sample sizes or when the distribution is not normal.

5.7.5 Results

The results of the study are reported in two broad sections. Methods specific to each section are presented prior to the result. First, the study presents the results for the imitation effect, where the hypothesis on reducing in-group favouritism is tested. Both the multi-level model and descriptive statistics are presented. Second, the result for the Turing test is presented in two subsections, one for each of the two hypotheses under the Turing test.

5.7.6 In-group Favouritism

5.7.6.1 Method

In the multi-level model of in-group favouritism, the intervals(), fixef() and summary() method in R were used to calculate 95% confidence intervals for each condition, standardised fixed effects for the conditions, and the summary output of the model, respectively.

5.7.6.2 Multi-level model

Given that the data were hierarchical in nature, the nesting structure was considered. There were game levels and individual levels. The first model was, therefore, to have statistical evidence for nesting. Two models of in-group favouritism were built. The first, Model 1 (AIC = 529.7980, $logLik$ = -262.8990) has no random effect while the second, Model 2 (AIC = 477.9833, $logLik$ = -

235.9917) includes random effects at the game level. A comparison of Model 1 and Model 2 shows that Model 2 was a better fit to the data (p < 0.0001). This indicates that random model terms are required.

Model 3 (*AIC* = 116.6706, *logLik* = -54.3353) was built with a random effect at both the game and individual levels. A comparison of Model 2 and Model 3 shows that model 3 was a better fit to the data (p < 0.0001).

Next, an interclass correlation coefficient (ICC) for each level of nesting was calculated. The ICC at the game level is 0.079, which means 8% unexplained variance lies at the game level. While the 8% random variance may not be considered extremely high, it is high enough to justify the use of a multi-level model. Also, the ICC at the individual level was 0.471, an extremely high value, which means that 47% unexplained variance lies at the individual level.

Next, a new covariance structure is added to account for repeated measures over waves. A first-order autoregressive structure "CorAR1" was added and called Modelcor. This was significant (p < 0.0001) and improved the goodness-of-fit (*AIC* = 101.2406).

Model 4 was built with a fixed effect of the three independent variables (condition, group and time) added. Model 4 (*AIC* = 98.97578, *logLik* = -38.48789) further improved the goodness-of-fit when compared to Modelcor (*AIC* = 101.2406).

Model 5 was built with two-way interactions added to Model4. Thus, Model 5 has both fixed effects and two-way interactions of the independent variables. The model (*AIC* = 90.34137, *logLik* = -25.1707) significantly (p < 0.0016) improved the goodness-of-fit.

Lastly, Model 6 was built with fixed effect, two-way and three-way interactions between condition, group and time. Rather than improve the model fit, Model6 (*AIC* = 96.51265, *logLik* = -24.2563), decreased the goodness-of-fit. Table 5.2 compares Model 6 with Model 5. Maximum likelihood algorithms were used to obtain estimates of AIC and Log-Likelihood, as well as for model comparisons.

Table 5.2. Model 5 compared to Model 6.

Model Name	Model Description	AIC	P-value
Model 5	Fixed effect + two-way interaction	90.34137	
Model 6	Fixed effect + two-way interaction + three-way interaction	96.51265	0.7672

Since Model 6 decreased the goodness-of-fit, Model 5 was reported as the final model. The model is given in Table 5.3, which shows that *group* (*p* = 0.0018), interaction between *condition* and *group* (*p* = 0.0285), and interaction between *condition* and *time* (*p* = 0.0043) are statistically significant for determining in-group favouritism. The significant effects are further explored using descriptive analysis. The RMD files, R Markdown files and other materials for this analysis can be found in the online material.

Table 5.3. The output of the final model testing the effect of out-group altruism on in-group favouritism.

	numDF	denDF	F-value	P-value
Condition	4	45	1.2759	0.2936
Group	1	225	9.9406	0.0018
Time (Wave)	1	1114	0.0104	0.9187
Condition:Group	4	225	2.7623	0.0285
Condition:Time	4	1114	1.9435	0.101
Group:Time	1	1114	8.1999	0.0043

5.7.6.3 Group as a main effect on in-group favouritism

Results of the model for in-group favouritism can be found in Table 5.3 It shows that *group* was statistically significant as a main effect while *condition* was not. Figure 5.5 shows that in-group favouritism is very high in Group 1 compared to Group 2. It is noteworthy that Group 1 always has an equal or higher number of artificial agents compared to Group 2. With reference to the artificial agents, this implies that out-group altruism introduced via the artificial agents reduced in-group favouritism among the out-group members compared to the in-group members.

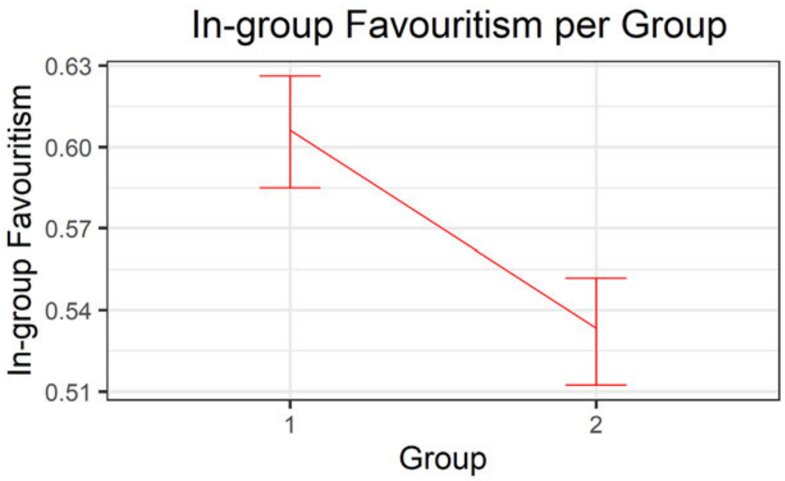


Figure 5.5. Differences in mean in-group favouritism between groups. Depending on the condition, Group 1 has equal (Conditions 1:1, and 2:2) or more (Conditions 1:0, 2:0, and 2:1) artificial agents than Group 2.

Although *condition* was not statistically significant as the main effect on in-group favouritism, the interaction between *condition* and *group* was statistically significant ($p < 0.02$). Thus, in-group favouritism increases or decreases based on the group to which each condition is applied. The study, therefore, provides further descriptive analysis in the next section to explore the interaction effect of *condition* and *group*.

5.7.6.4 Descriptive analysis of the interaction effects

Figure 5.6 shows the graph of the mean in-group favouritism per group in each condition. It shows that the reduced mean in-group favouritism found in Condition 1:0 was mainly from Group 2, a group with no agent. The graph shows more disparity in in-group favouritism when the number of agents in the groups is imbalanced. For each condition, Group 2 (the group with no agents or the lower number of agents) exhibit less in-group favouritism than the group with more agents. This further confirms that out-group altruism reduces in-group favouritism among out-group members but not among in-group members. Furthermore, it suggests that humans are reciprocating favour from agents and have lower in-group favouritism. In-group favouritism decreases in Condition 1:1, as well as Condition 2:2, which showed convergence in the mean in-group favouritism.

The disparity in in-group favouritism found in the condition with an imbalanced number of agents suggests that the perception of fairness across conditions may differ. This is explored in the next section.

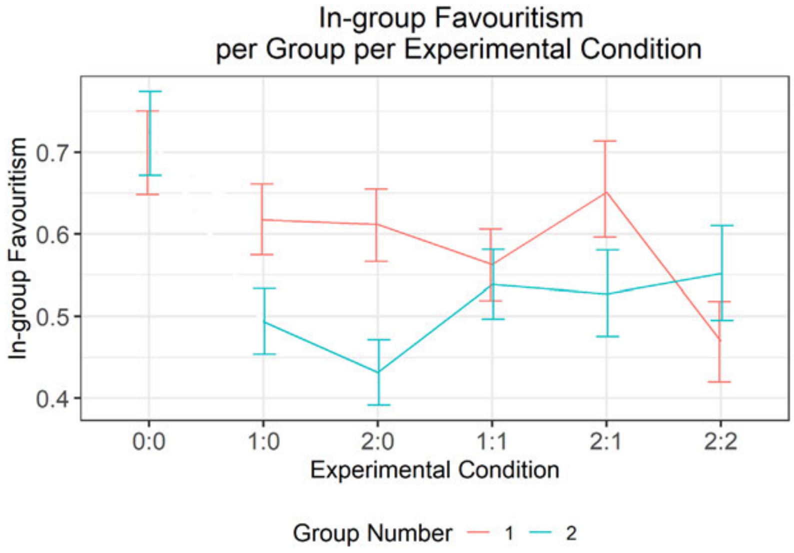


Figure 5.6. Differences in mean in-group favouritism between condition per group. The first digit (i.e. 0, 1, 2, 1, 2, 2) in each experimental condition represents the number of agents in Group 1, while the second digit (i.e. 0, 0, 0, 1, 1, 2) represents the number of agents in Group 2. Condition 0:0 was obtained from (Durrheim et al., 2016).

Figure 5.7 shows the graph of the mean in-group favouritism per group in each wave. In Group 2, in-group favouritism alternates between a slight increase and a slight decrease over time, while in Group 1, it increases steadily over time (except in the last wave). This increase was significant from the second to the last wave.

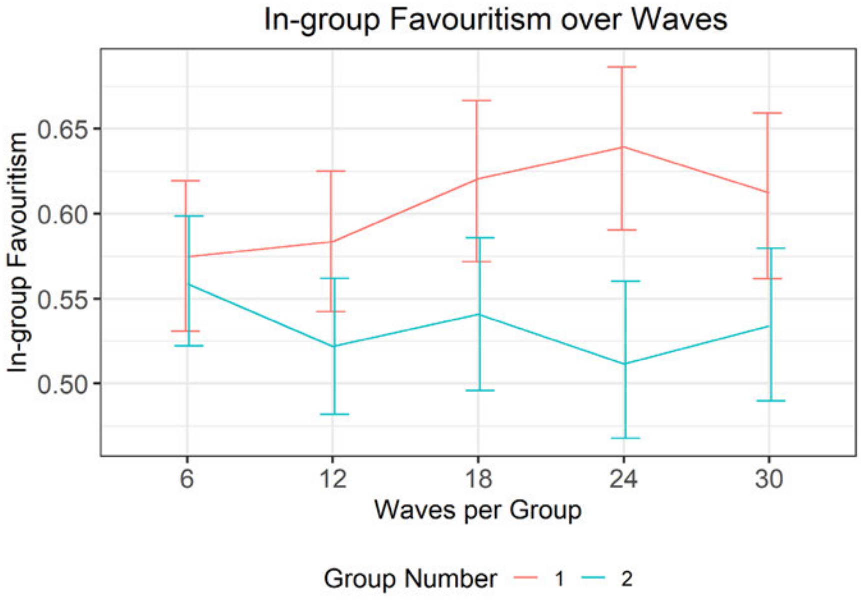


Figure 5.7. The difference in mean in-group favouritism over time. Each wave comprised five rounds of exchanges.

Overall, Group 2, with fewer agents, has lower in-group favouritism. Moreover, the in-group favouritism of Group 1 increases over time. This shows that agents may promote out-group favouritism in Group 2 but solidarity in Group 1.

5.7.7 Fairness Rating

5.7.7.1 Method

Fairness rating is a dependent variable described in the general method section. Each fairness rating received by an agent or a human is in the range of 1 to 5. To examine whether or not the agents were rated as being fairer than humans, especially by the out-group members, there is a need for a factor that could differentiate an in-group rating from an out-group rating. Thus, a level 1 factor named *TargetGroup* is created with two categories, namely, *InGroupRating* and *OutGroupRating*, as shown in Table 5.4. *InGroupRating* describes a rating from a player to another player of the same group, while *OutGroupRating* describes a rating to another player of a different group. Also, there is a need to know whether a human or an agent received the rating. Thus, an extra independent variable named *TargetParticipant* is created, with two categories, namely, *human* and *agent*.

Independent variables

- **TargetGroup**: Each human player rated an in-group or out-group member.
- **TargetParticipant**: Each rating is received by either a human or an agent.
- **Condition**: There were five conditions: Condition 1:0, Condition 2:0, Condition 1:1, Condition 2:1, and Condition 2:2.

Dependent variable

- **Rating**: The rating received by each player.

Table 5.4. Preprocessing of the rating data. Consider, as an example, four players A, B, C and D, where Players A and B are in Group 1, and Players C and D are in Group2. Player D is an agent. Agents are rated but cannot rate other players. Except for the agents, each player rated the other players. Note that both Group 1 to Group 1 ratings and Group2 to Group2 ratings are in-group while others are out-group.

Providing Rating		**Being Rated**		**A level 1 Factor**	**Independent Variable**	**Dependent Variable**
Player	**Group**	**Player**	**Group**	**TargetGroup**	**TargetParticipant**	**Rating**
A	1	B	1	InGroupRating	Human	3
A	1	C	2	OutGroupRating	Human	4
A	1	D	2	OutGroupRating	Agent	2
B	1	A	1	InGroupRating	Human	3
B	1	C	2	OutGroupRating	Human	4
B	1	D	2	OutGroupRating	Agent	4
C	2	A	1	OutGroupRating	Human	2
C	2	B	1	OutGroupRating	Human	3
C	2	D	2	InGroupRating	Agent	3

5.7.7.2 Multi-level model

As with in-group favouritism, the nesting structure was considered; thus, the first model was to have statistical evidence for nesting. Two models were built: Model 1 without random effect and Model 2 with a random effect at the game level. The comparison provided statistical evidence ($p < 0.0001$) for nesting. Model 2 (*AIC* = 6562.839, *logLik* = -3278.420) with random effect at game level has a better fit than Model 1 (*AIC* = 6606.327) with no random effect. This indicates that random model terms are required. A random effect at both the game as well as individual level was added. This model, Model 3 (*AIC* = 6474.303, *logLik* = -3233.151) has a better fit and was statistically significant ($p < 0.0001$) when compared to Model 2.

An interclass correlation coefficient (ICC) for each level of nesting was calculated. The ICC at the game level is 0.050, which means 5% unexplained variance lies at the game level. While the 5% random variance may not be considered extremely high, it is fair enough to justify the use of a multi-level model. Also, the ICC at the individual level was 0.183, a fair amount which means that 18% unexplained variance lies at the individual level.

An autoregressive structure is useful to account for the use of information from the previous round/time in predicting the value in the current round. However, there was no time factor in the fairness ratings. Hence there was no need for the first-order autoregressive structure.

The next model (i.e. Model 4) was built with a fixed effect of condition, TargetGroup and

TargetParticipant (see method in Section 5.7.7.1). This improved the goodness-of-fit (*AIC* = 6466.252, *logLik* = -3223.126) and was statistically significant (*p* = 0.0027).

Model 5 was built with two-way interactions added to Model 4. Thus, Model 5 has both fixed effects and two-way interactions of the condition, TargetGroup and TargetParticipant. The model (*AIC* = 6447.254, *logLik* = -3204.627) significantly (*p* < 0.0001) improved the goodness-of-fit.

Lastly, Model 6 was built with fixed effect, two-way and three-way interactions of condition, TargetGroup and TargtParticipant. Rather than improve the model fit, Model 6 (*AIC* = 6451.412, *logLik* = -3202.706), decreased the goodness-of-fit and was not statistically significant. Thus, Model 5, provided in Table 5.5, was reported as the final model.

Table 5.5. The output of the final model for fairness rating.

	numDF	**denDF**	**F-value**	**p-value**
(Intercept)	1	1657	4268.732	<.0001
Condition	4	45	0.997	0.4191
TargetGroup	1	1657	12.883	0.0003
TargetParticipant	1	1657	3.597	0.0581
Condition:TargetGroup	4	1657	1.49	0.2027
Condition:TargetParticipant	4	1657	4.892	0.0006

5.7.7.3 TargetGroup as a main effect on fairness ratings

Results of the model for fairness rating can be found in Table 5.5. It shows that the ratings received by humans were not significantly different from those received by artificial agents. However, out-group ratings were significantly different from in-group ratings. Figure 5.8 shows that the significant difference is a result of participants rating in-group members higher than the out-group.

Since the interest of the study is in the ratings received by *agents* versus *humans*, which is not significant, a further report of the main effects of the independent variables on fairness rating was not provided.

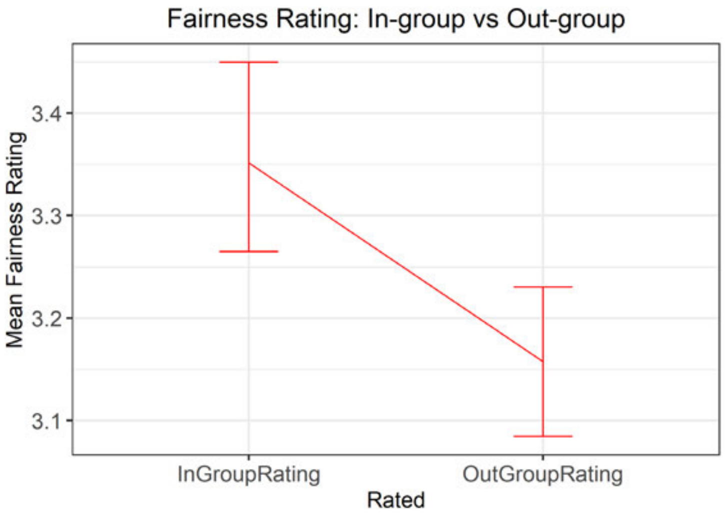


Figure 5.8. Fairness ratings of the in-group members compared to the out-group members.

5.7.7.4 Descriptive analysis of interaction effects on fairness rating

Figure 5.9 plots the interaction effect of condition and TargetParticipant on fairness rating. The figure shows that the significant effect was a result of humans being rated as fairer in one condition – with one agent in each group.

Figure 5.10 depicts the differences in fairness ratings received by humans and agents from in-group and out-group members. The figure shows that in-group members rate humans as fairer than the defecting agents. The minimum and maximum mean fairness ratings received by humans and agents from in-group and out-group members are 3.10 and 3.48 respectively, showing a difference of 0.38 between humans and agents.

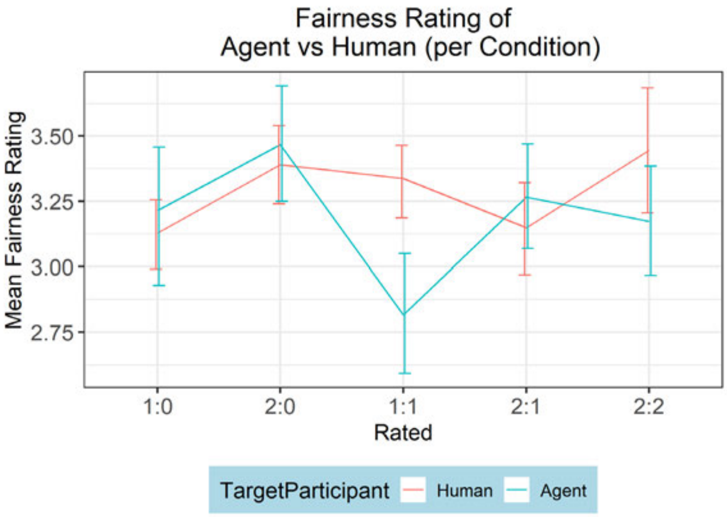


Figure 5.9. Humans were rated fairer than agents in one condition – with one agent in each group. The green line shows the ratings received by agents, while the red line shows those received by humans.

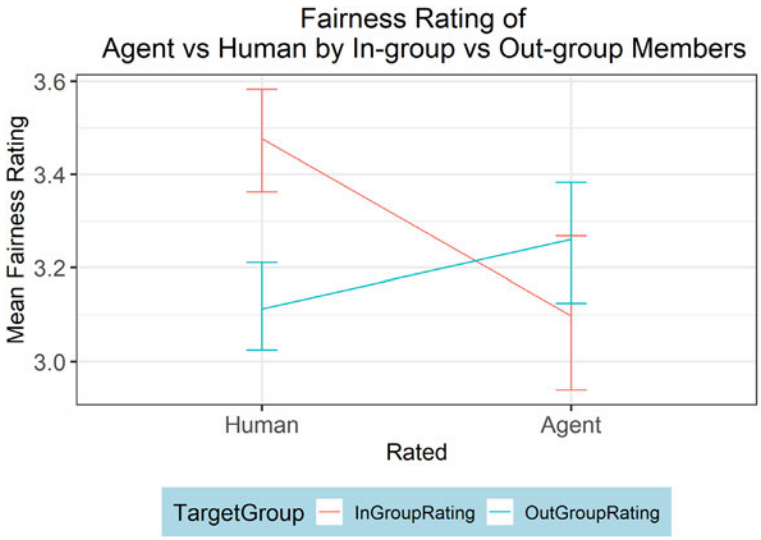


*Figure 5.10. Differences in mean fairness rating received by humans and agents from in-group and out-group members. The red line shows that humans received higher ratings from in-group members, while agents received lower ratings from in-group members. The blue line shows that humans received lower ratings from out-group members, while agents received higher ratings from out-group members***.**

The ratings received by humans, from in-group members, had the highest mean (*M = 3.48*, *SD* = 1.33). The second highest mean (*M* = 3.26, *SD* = 1.25) was for the ratings received by agents, from out-group members. The mean (*M* = 3.11, *SD* = 1.31) of the ratings received by humans from out-group members and that (*M* = 3.10, *SD* = 1.34) received by agents from in-group members have a negligible difference of 0.01, confirming the assertion that agents were treated as out-group members.

Overall, mean in-group rating (*M* = 3.351, *SD* = 1.345) is higher than the mean out-group rating (*M* = 3.157, *SD* = 1.296).

5.7.8 Agent Identification Rating

5.7.8.1 Method

The agent identification rating data was also pre-processed, as done with the fairness rating data.

5.7.8.2 Multi-level analysis

Given the nesting structure of the data, models were built to ascertain that random effects are required. Model 1 (*AIC* = 7151.047) with no random effect was built and compared to Model 2 with random effect at the game level. Model 2 (*AIC* = 7134.300, *logLik* = -3564.150) has a better fit than Model 1 (*AIC* = 7151.047, *logLik* = -3573.52) and was statistically significant ($p <$ 0.0001), thus providing evidence for nesting. Model 3 was built with a random effect at both the game as well as individual level and was compared to Model 2. Model 3 (*AIC* = 7119.979, *logLik* =-3555.989) has a better fit and was statistically significant ($p <$ 0.0001).

An interclass correlation coefficient (ICC) for each level of nesting was calculated. The ICC at the game level is 0.028, which means 3% unexplained variance lies at the game level. This is not a high value but could not rule out the use of a multi-level model. Also, the ICC at the individual level was 0.081, a fair amount which means that 8% unexplained variance lies at the individual level. This value is fair enough to justify the use of a multi-level model.

As with the fitness rating, autoregressive structure was not added because the rating data has no time or round factor.

Model 4 was built with a fixed effect of condition, TargetGroup and TargetParticipant (see method in Section 5.7.7.1). The model (*AIC* = 7128.970, *logLik* = -3554.485) did not improve the goodness-of-fit and was not statistically significant (p = 0.8078).

Model 5 was built with two-way interactions added to Model 4. Thus, Model 5 has both fixed effects and two-way interactions of the condition, TargetGroup and TargetParticipant. The model (*AIC* = 7113.697, *logLik* = -3537.848) significantly ($p <$ 0.0001) improved the goodness-of-fit.

Lastly, Model 6 was built with fixed effect, two-way and three-way interactions of condition, TargetGroup and TargetParticipant. Model 6 (*AIC* = 7111.307, *logLik* = -3532.653) improved the goodness-of-fit and was statistically significant ($p = 0.0344$. Thus, Model 6 (provided in Table 5.6) was reported as the final model.

Table 5.6. The output of the final model for agent identification rating.

	numDF	denDF	F-value	p-value
(Intercept)	1	1646	3059.862	<.0001
Condition	4	45	0.4345	0.7829
TargetGroupBias	1	1646	0.5779	0.4472
TargetParticipant	1	1646	0.7192	0.3965
Condition:TargetGroupBias	4	1646	0.4392	0.7804
Condition:TargetParticipant	4	1646	7.8791	<.0001
TargetGroupBias:TargetParticipant	1	1646	0.1811	0.6705
Condition:TargetGroupBias:TargetParticipant	4	1646	2.578	0.0359

5.7.8.3 Descriptive analysis of interaction effects on agent identification ratings

As indicated by the non-significant main effects of the independent variable on agent identification ratings, there is no difference in the ratings received by agents and humans. However, the interaction effects are explored to understand the interaction effects.

Figure 5.6 depicts the differences in mean agent identification rating received by humans and agents for each condition. Apart from Conditions 2:0 and 1:1, with significant differences in the received ratings, other conditions show that participants rated agents as humans. Interestingly to the researcher, humans were identified as agents in Condition 2:0. However, agents were rated more human than humans in Condition 1:1. Figure 5.12 indicates that in-group members were the ones that rated humans as agents in Condition 2:0; there is no significant difference in the out-group ratings.

Although there is no significant difference, the rating ($M = 2.76$, $SD = 1.49$) received by humans from out-group and those ($M = 2.85$, $SD = 1.58$) received by agents from out-group members are higher than the ratings ($M = 2.75$, $SD = 1.55$) from in-group members to agents and to humans ($M = 2.73$, $SD = 1.55$).

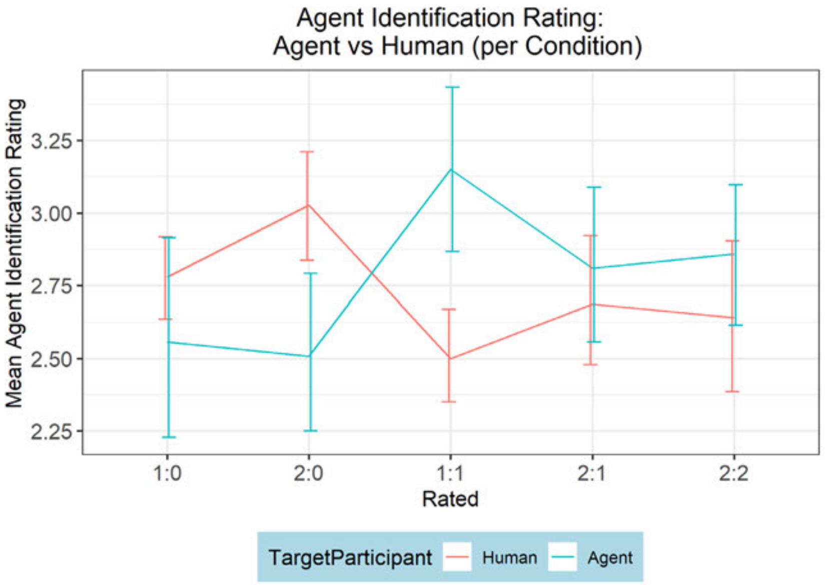


Figure 5.6. Depicts the differences in mean agent identification rating from each human to other humans and agents for each condition.

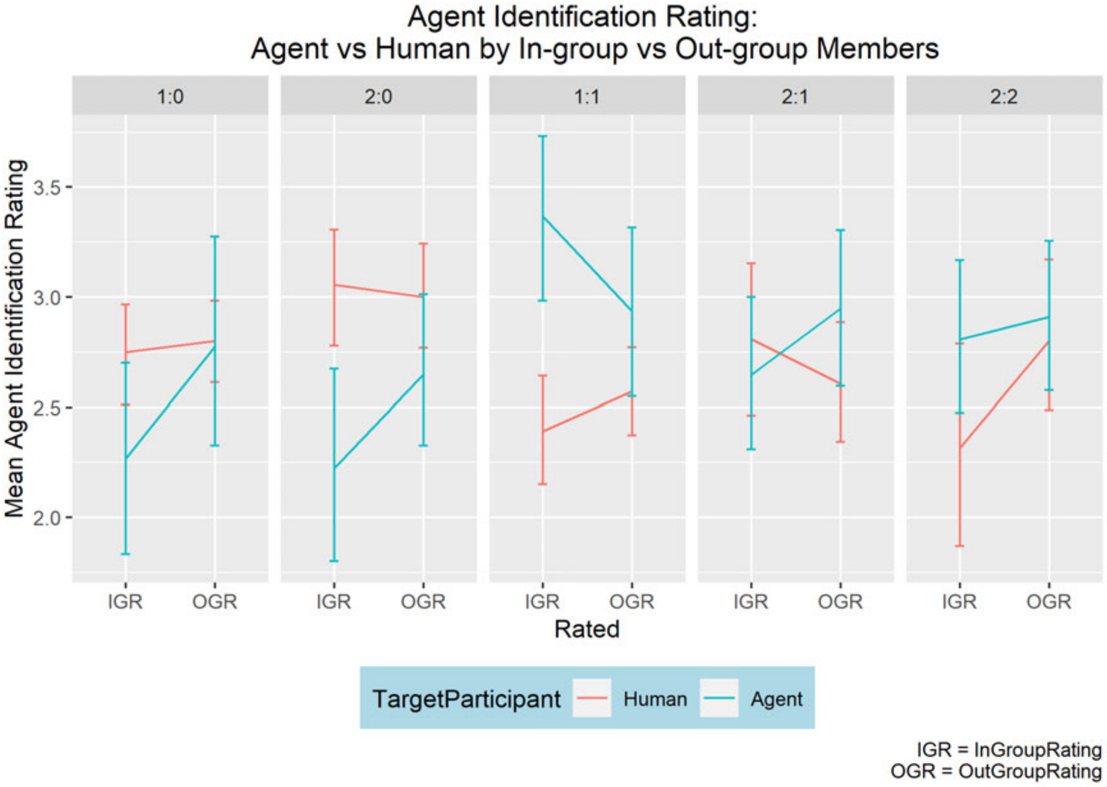


Figure 5.7. Depicts the differences in mean agent identification rating between humans and agents for in-group and out-group members.

5.7.9 *Further Test for Agent Identification: Correlation between Fairness and Agent and their Impact on In-group Favouritism Ratings*

Although the outcome of the ratings supports the hypothesis that participants did not recognise the agents (discussion in Section 5.7.8), it was anticipated that participants may indirectly identify the agents by exhibiting behaviours that indicate the presence of agents. The study assumes that, if the agents are rated as humans, but also consistently rated as not being fair, then the participants have indirectly identified the agents. Hence, it was necessary to check for a correlation between fairness and agent identification for both humans and agents.

Figure 5.13 shows the correlation between agent identification and fairness ratings for both humans and agents. Except for the positive correlation (*r* = 0.25*, p* = 0.038) found in Condition 1:0 and a negative correlation (*r* = -0.49*, p* = 0.00034) for Condition 2:1, both for humans, there were no significant correlations between the rating task. Hence, there was no proof that participants indirectly identified the agents.

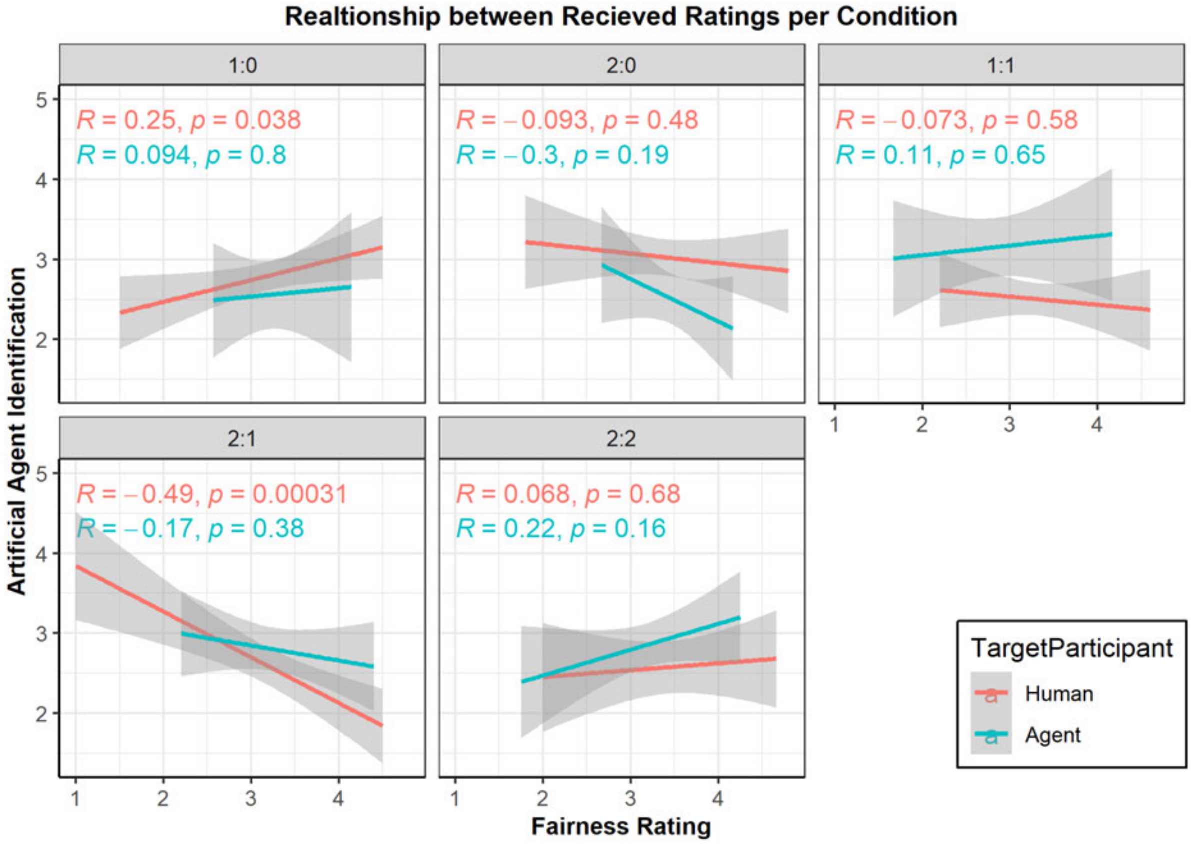


Figure 5.8. Correlation between fairness and agent identification.

5.7.10 Further Test for Punishment and Reciprocity

The analysis of in-group favouritism suggests that groups with more agents have higher in-group favouritism. This is contrary to the imitation effect hypothesised in Section 5.6.1 in which humans were expected to imitate the agents' out-group altruistic behaviour. Higher in-group favouritism amongst groups with more agents suggests that humans may have punished in-group agents for defecting (i.e. for exhibiting out-group altruism) and favoured other in-group members. On the other hand, out-group members may have reciprocated to the other groups' agents who were favouring them, thus causing a reduction in in-group favouritism in groups with fewer agents.

5.7.10.1 Method

To examine these claims, there is a need for a factor that could differentiate allocations from humans to in-group humans and from humans to in-group agents. Also, there is a need to differentiate allocations from humans to out-group humans and from humans to out-group agents. Thus, a level 1 factor named *GroupByParticipantType* was created with four categories, namely, in-groupHuman, in-groupAgent, out-groupHuman and out-groupAgent, representing the abovementioned allocations respectively. Self-allocation (allocation to oneself) is not considered.

To normalise the allocations per round, the number of tokens allocated to each category is divided by the total number of possible allocations to that category. Thus, the normalised allocation is calculated as $(Num\ of\ Allocation\ to\ X\,)/(max\ X)$, where X represents one of the *GroupByParticipantType* categories and max represents the total number of possible allocations.

For example, assume that three humans and one agent were assigned to a group. With the exception of self-allocation, two of the humans can allocate their tokens to in-group humans (max in-groupHuman = 2) while the three humans can allocate their tokens to the in-group agent (max in-groupAgent = 3). Also, the three humans can allocate their tokens to either out-group humans (max out-groupHuman = 3) or out-group agents (max out-groupAgent = 3). Assuming that two of the humans allocated their tokens to the out-group humans and one to the agent, then: in-groupHuman $= 0/2$, in-groupAgent $= 1/3$, out-groupHuman $= 2/3$, and out-groupAgent $= 0/3$.

Independent variables

- **GroupByParticipantType**: There were four categories: in-groupHuman, in-groupAgent, out-groupHuman and out-groupAgent.
- **Condition**: There were five conditions: Condition 1:0, Condition 2:0, Condition 1:1,

Condition 2:1, and Condition 2:2.

- **Group**: Each condition has two groups.

Dependent variable

- **NormAllocation: N**ormalised allocations to each category.

5.7.10.2 Multi-level analysis

The main aim of the analysis is to confirm whether or not there are differences, per condition and group, in the allocation from humans to in-group humans, in-group agents, out-group humans and out-group agents.

Model 1 (*AIC* = 2566.848, *logLik* = -1281.424) without random effects and model 2 (*AIC* = 2505.829, *logLik* = -1249.914) with random effects were built and compared to test statistical evidence for nesting. Model 2, which includes random effects at the game level, significantly ($p <$ 0.0001) improved the goodness-of-fit. Since the normalised allocation was aggregated at the round level, the individual level is not considered.

Focusing on the said aim and using *lme4* package in R, a three-way interaction of condition, group and *GroupByParticipantType* variables were modelled at the game level. The three-way interaction was significant ($p < 0.0001$), as shown in Table 5.7 (see supplementary materials here).

Table 5.7. The output of the final model for punishment and reciprocity.

	NumDF	**DenDF**	**F value**	**Pr(>F)**	**Sign.**
Condition	4	1918.8	77.564	< 2.2e-16	***
Group	1	5558.7	257.958	< 2.2e-16	***
GroupByParticipantType	3	5567.9	192.351	< 2.2e-16	***
Condition:Group	4	5552.4	86.114	< 2.2e-16	***
Condition: GroupByParticipantType	12	5564.7	13.415	< 2.2e-16	***
Group: GroupByParticipantType	3	5567.4	21.154	1.27E-13	***
Condition:Group: GroupByParticipantType	8	5564.6	11.123	1.02E-15	***

*** = significant at 0.001

5.7.10.3 Descriptive analysis of punishment

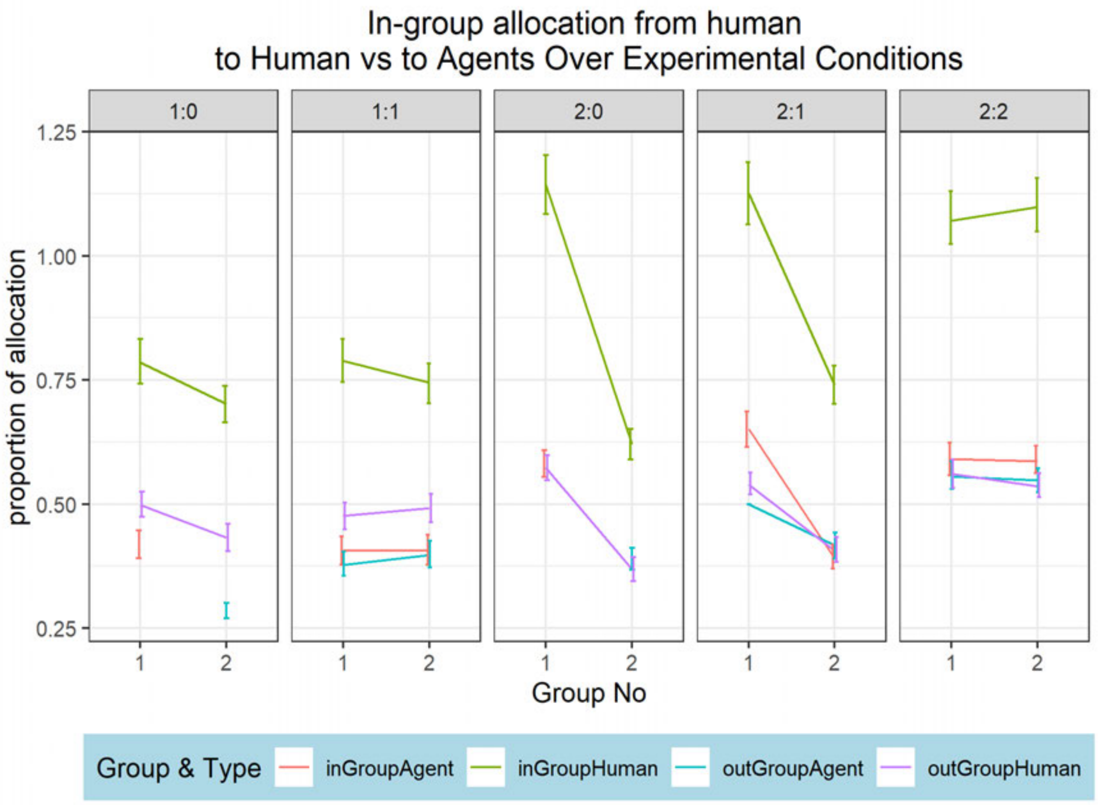


Figure 5.9. Allocation from humans to in-group and out-group humans and agents. The graph shows that in-group agents were treated as out-group members.

Figure 5.9 shows the graph of the three-way interaction of condition, group and *GroupByParticipantType*. The figure shows that in-group agents were punished for defecting. This is shown by the significant high allocations from in-group humans to other in-group humans.

Allocation from humans to in-group humans was the highest in all the conditions. However, groups with more agents show a greater tendency for in-group human allocation. In conditions 2:0, 2:1 and 2:2, the clear and significant differences between allocations to in-group humans, and others support the claim that agents were treated as out-group members.

On the other hand, there is no significant evidence supporting that out-group members may have reciprocated to the other groups' agents who were favouring them. Instead, the graph shows that out-group members reciprocate the favour equally to both humans and agents.

5.7.11 Discussion

Study 1 aimed to investigate out-group altruism as a means for reducing in-group favouritism. With many competing theories as to the underlying causes of in-group favouritism and a few suggestions about how in-group favouritism can be reduced, social influence theory suggests that a norm or behaviour can be enacted through conformity and/or imitation. In the first, an individual imitates a displayed behaviour to gain approval through conforming; in the second, an individual imitates a displayed behaviour to identify with a group. According to Bandura (1986), behaviour can be induced when an individual consistently observes the referent other exhibiting the behaviour, although Fulk et al. (1990) warned that this might lead to either rejection, where undesirable consequences are avoided, or acceptance, where the induced behaviour is accepted. Also, research (Balliet et al., 2014; Yamagishi et al., 1999) suggests that in-group defectors (in this study, in-group member exhibiting out-group altruism) will get a bad reputation in the in-group. The study expected that out-group altruism can be imitated, thereby reducing in-group favouritism.

Furthermore, the use of artificial agents, rather than humans, as experimental stooges guaranteed that the behaviour the study intended to induce was consistently displayed. However, a major disadvantage of using artificial agents is that participants may change their behaviour as a result of their awareness of the presence of the agents. Thus, the task in this study was not only to investigate out-group altruism for reducing in-group favouritism via imitation but also to ensure that the result was not influenced by participants' awareness of the presence of artificial agents. The main findings from this study will, therefore, be discussed as follows: (i) outgroup altruism reduces in-group favouritism amongst out-group members but not among in-group members; (ii) participants were not aware of the presence of agents as there was no significant difference between the fairness ratings received by agents and those received by humans; instead, in-group favouritism was displayed as the participants rated in-group members as fairer than the out-group members; (iii) participants did not identify the agents; thus, the behaviour exhibited was not a result of the participants' awareness of the presence of any artificial agent(s) among the players.

5.7.11.1 Depending on the source, out-group altruism accentuates or attenuates in-group favouritism

In this study, the multi-level model showed that time was not a statistically significant variable affecting in-group favouritism, as the main effect. However, the interaction between time and group was significant. The descriptive analysis shows that in-group favouritism increases with time amongst groups with more agents, showing evidence of solidarity (and punishments of the agents as shown in Section 5.7.10). Noting that the significant change with time was observed only at the fourth wave, this study does not make extrapolations about in-group favouritism increasing or decreasing over time. Instead, it explains how in-group favouritism changes in each condition and makes suggestions on the causes.

It is evident from the literature that mere categorisation of participants into groups has a significant effect on in-group favouritism (Durrheim et al., 2016). This study also supports this evidence in that interaction between condition and group was found to have a significant effect on in-group favouritism. This study found evidence that different conditions (i.e. different number of agents in each group, or generally put, different levels of out-group altruism), in a group context, attenuate or accentuate in-group favouritism.

Except for Condition 2:2 (i.e. two agents in both groups), in-group favouritism was higher in groups with more agents than groups with fewer agents. For example, in Condition 1:0 (i.e. one agent in Group 1 and none in Group 2), in-group favouritism was higher among Group 1 members than it is among Group 2 members. This high in-group favouritism among groups with more agents suggests that participants did not imitate the agents. Instead, they resisted the effect of imitation. This behaviour is consistent with Haslam and Reicher (2006) finding that stressing group members leads to increased shared identity amongst that group members. The increased shared identity, in turn, leads to the group members' support for each other.

According to the prison study of Haslam and Reicher (2006), people sometimes resist actions or rules that lead to undesirable consequences. Haslam and Reicher (2006) examined the behaviour of men randomly assigned to groups as guards and prisoners. The study supports that people do not conform automatically to behaviour or "their assigned role" (Haslam & Reicher, 2012, p. 2). However, they conform because "they had internalised roles and rules as aspects of a system with which they identified" (Haslam & Reicher, 2012, p. 2). The study further supports that the salience of group identity could enforce resistance rather than conformity. Thus "group identity did not mean that people simply accepted their assigned position [or a behaviour]; instead, it empowered them to resist it" (Haslam & Reicher, 2012, p. 3).

In Study 1, as a sense of shared social identity increased among members of the group with a high number of agents, they show a need to support (allocate tokens to) each other. Although the expectations of this study was that participants will imitate agents over time; it was found that in-group members see agents as traitors (see discussion on fairness rating). Thus, the study concludes that the high in-group favouritism observed amongst members of the group with a high number of agents is a result of resistance to agents' behaviour and support for each other (see the analysis in Section 5.7.10). The findings are also consistent with Fulk et al. (1990) suggestion that any attempt to induce an undesired behaviour may lead to rejection, where undesirable consequences are avoided.

5.7.11.2 Turing test

Although there has been a long-standing argument on the actual meaning of the Turing test, a common understanding is that the purpose of a Turing test is to ascertain whether or not a computer can imitate a human (Saygin et al., 2000). This study adopts the concept of the Turing test. Thus, the fairness rating was to check for 'indirect agent identification', that is, whether or not participants were able to identify agents; this involved checking whether or not they rated them differently from humans. The agent identification rating was to check for 'direct agent identification', that is, whether or not participants were able to single out the agents in the study.

Recognising 'the good one amongst them' while supporting my group: In-group favouritism in the fairness rating task

This study expected that agents would be rated as fairer than humans, especially by out-group members. This study could not find any significant evidence that agents were rated as fairer than humans. However, it was found that, irrespective of the participants' group, in-group humans rated out-group agents (and out-group humans in some cases) as fairer than in-group agents. This further shows that participants observed that 'some in-group members' (the agents) were not identifying with other in-group members during the game. Thus, rather than imitating the agents, they were treated like traitors by their in-group members but as 'good amongst others' by the out-group members.

Whereas this study aimed at checking for a difference in the ratings received by agents and humans, it found further evidence for in-group favouritism in the rating task. Overall, the in-group fairness rating was significantly different from the out-group fairness rating. The study concludes that the in-group favouritism in the fairness rating (but not in the agent identification rating – see discussion below) occurred because fairness is seen as an attribute of a positive behaviour, while being an agent amongst humans is not associated with positive behaviour.

Agents successfully hid amongst humans: A quest for inducing prosocial behaviour

Despite the failed imitation but salient resistance in Study 1, there was no evidence indicating that agents were identified among participants. Thus, it supports the idea that artificial agents with pre-programmed behaviour can successively be integrated into a computer-mediated experimental platform and used for inducing behaviour.

There was no significant difference either between humans and artificial agents or in-group and out-group agent identification ratings. Unlike the fairness rating, where participants rated in-group members as being fairer, both the in-group and out-group agent identification ratings were almost the same. Strikingly, to the researcher, the rating pattern suggests that neither in-group nor out-group members 'pointed an accusing finger' at one another. While the study failed to show any evidence that supports imitation, it suggests that artificial agents with a dynamic strategy that incorporates, for example, out-group and in-group altruism, could reduce in-group favouritism.

5.7.12 Implication

The study in this paper has several implications for research on out-group altruism and in-group favouritism. First, this study was able to demonstrate the role of out-group altruism in reducing in-group favouritism. This study suggests that out-group altruism may reduce or increase in-group favouritism depending on the source. It shows that, although altruism is a prosocial behaviour, out-group altruism that ignores the need to identify with the in-group members can create undesirable outcomes amongst in-group members. Much research (Brewer, 1999; for a review, see Greenwald & Pettigrew, 2014) has provided evidence that in-group favouritism forms the basis for intergroup bias such as discrimination, prejudice and stereotypes. Out-group altruism – which plays a role in reducing in-group favouritism among the out-group members – may play a role in reducing intergroup bias. Further research should investigate strategies that could be incorporated alongside out-group altruism to reduce in-group favouritism, both in in-group and out-group members.

Second, although the focus was on out-group altruism, the findings support notable research (Fulk et al., 1990; Haslam & Reicher, 2006; Haslam & Reicher, 2012). These research works suggest that a quest to resist undesired behaviour, common to in-group members, increases shared social identity among the in-group members. This, in turn, increases in-group members' need to support each other. Consistent with this interpretation, in-group members in this study developed the tendency to resist the agents' altruistic behaviour and support each other by allocating tokens to each other.

5.7.13 Study 1 Summary

Study 1 aimed to investigate out-group altruism for reducing in-group favouritism. Out-group altruism was implemented via artificial agents that had been pre-programmed to allocate their wealth to out-group members randomly in the game. The results partially support the initial expectation that out-group altruism is capable of weakening in-group favouritism. The findings of this study are that out-group altruism weakens in-group favouritism among out-group members (who reciprocated kindness generally among out-groups) but not among in-group members. While Study 1 acknowledges the role of out-group altruism in reducing in-group favouritism, it suggests that incorporating a strategy that promotes in-group identity while setting out-group altruistic norms may further reduce in-group favouritism.

5.8 Study 2

The overall objective of the studies is to positively exert social influence, thereby weakening in-group favouritism. In Study 1, non-adaptive (rule-based) agents were pre-programmed to randomly allocate their tokens to the out-group members. The strategy failed to weaken in-group favouritism among in-group members. Instead, it led in-group members to treat agents as out-group members. This observation is in line with the bounded generalised reciprocity theory (Yamagishi et al., 1999; Yamagishi & Mifune, 2009), which suggests that, due to reputational concern, favour is more likely to be reciprocated by in-group members compared to the out-group members (Balliet et al., 2014; Yamagishi et al., 1999), and that in-group defectors will get a bad reputation among in-group members. Reducing the effect of bounded generalised reciprocity will increase the participants' focus on between-group exchange as opposed to within-group exchange.

In this regard, Study 2 employs a strategy to break the theory of bounded generalised reciprocity. The primary intervention was to change the behaviour of out-group members by coaxing them into out-group allocation. This is done in the hope of changing the agents' in-group members' behaviour by changing their perception of interdependence which had been postulated as vital for norm emergence (Balliet et al., 2014; De Dreu et al., 2020).

The strategy is implemented via the use of a machine learning model for (adaptive) agents' decision-making. Each agent tries to identify out-group individuals: (i) most likely to reciprocate, and (ii) whose history is dominated by reciprocity. Then the agent allocates its token to one of the identified individuals. Where an agent finds no such individual in a particular round, the agent allocates its token to an in-group member. This way, Study 2 attempted to promote cross-group reciprocity while trying to change the agent's in-group members' perception of interdependence.

Figure 5.15 presents the steps in implementing the strategy. The agents in Chapter 4 were developed to predict social exchange decisions. In Line 1, the agents make use of the co-evolutionary model developed in Chapter 4 to predict participants' exchange decisions (i.e. reciprocity, allocation to in-group or out-group, and allocation to a high-status or a low-status other). Then, each agent selects out-group members predicted to reciprocate (Lines 2 and 3). For each agent, if at least one of the agent's out-group members was predicted to reciprocate, the agent checks the history – the previous allocation decisions – of the out-group member. If the individual's history was dominated by reciprocity, the agent allocates its token to this individual; otherwise, the token is allocated to the agent's in-group member (Line 4 -6).

```
1. prediction ← MlPredictExchange(Agents, History,roundNo)
2. OutgroupInPrediction  ← getOutGroup(outgroupPrediction)
3. potentialReciprocators ← getReciprocator(OutgroupInPrediction )
4. if (isNotEmpty(potentialReciprocators )){
       a. AllocateToken(oneOf(potentialReciprocators))
5. }else{
       a. if (isNotEmpty (potentialIngroupAllocator )){
              i. StrategyList ←
                 getDominatedStrategies(potentialIngroupAllocator ,History)
             ii. ingroupDominators ← grtInGroupOrineted(StrategyList)
            iii. if (isNotEmpty(ingroupDominators )){
                     1. AllocateToken(one of(ingroupDominators))
                 }
          }
       b. AllocateToRandomIngroup( )
6. }
```

Figure 5.10. Systematic interaction algorithm with two main components: Machine-learning (ML) prediction (Line 1) and rule-based algorithm that acts based on the outcome of the prediction (Lines 2 – 6).

5.8.1 Sample, Design and Group Assignments

5.8.1.1 Sampling

The sample comprised 360 student participants (261 female, 93 male; 334 Black, 21 Indian, 5 coloured, 0 white) from two South African Universities, namely, the University of KwaZulu-Natal and the Durban University of Technology. The participants (mean age = 22.24 years, SD = 3.20) provided electronic copies of informed consent to participate in the study, which had been approved on the 11/01/2019 and amended on the 02/09/2020 by the Human Sciences Research Ethics Committee of the University of KwaZulu-Natal with a protocol reference number HSS/2210/018D (see Appendix 1b). The study employed the same sampling method as Study 1.

5.8.1.2 Design

The outcome of Study 1 led to two changes to the design described in Study 1. Firstly, the descriptive analysis of the interaction effects of condition and group presented in Section 5.7.6.4 shows a more (significant) disparity in in-group favouritism when there is an imbalanced number of agents in the groups, and less (non-significant) disparity when there is a balanced number of agents in the groups. Based on this knowledge, Study 2 examines in-group favouritism under three conditions – a null condition (0:0), a condition with an imbalanced number of agents (2:0) and a condition with a balanced number of agents (2:2). These conditions and their interpretations are shown in Table 5.8.

Table 5.8. Conditions under which in-group favouritism will be examined.

Condition	**Short-form**	**Description**
Condition 0:0	0:0	Null condition with no agent
Condition 2:0	2:0	Two agents in one group, and no agent in the other group
Condition 2:0	2:2	Two agents per group

Secondly, although the interaction effect of rounds and group was significant, previous research (Titlestad et al., 2019) has shown that as few as ten rounds are enough for norm emergence. Based on this, the number of rounds per game was reduced from 30 to 15. In all, there were 60 games, 20 per condition. Each had eight players and one of the conditions defined in Table 5.8.

5.8.1.3 Group assignment

The same procedure as specified in Study 1 for group assignments was used. Firstly, participants were randomly assigned to experimental conditions. Then, as in the previous study, participants were randomly assigned to two different groups, represented by the colour of their avatar on the screen. Also, they were led to believe that the assignment was done based on the dot estimation task they performed prior to the interactive exchange game. See the previous Study 1 report for the dot estimation task.

5.8.2 Ethical Considerations in Sampling and Data Collection

Ethical considerations are the same as in Study 1 but with one change – the location of the study was changed to online. Thus, participants were able to use their personal computers and did not need to use the computer in the laboratory. This was done as a result of the COVID-19 outbreak. Also, the incentives structure was changed from R30 per participant to an average of R40 to encourage participation.

5.8.3 Measures

5.8.3.1 Independent variables

Similar to Study 1, the independent variables were:

- **Group**: Consists of two levels: Group 1 and Group 2.
- **Condition**: There were three conditions: 0:0, 2:0, and 2:2.
- **Time**: The three waves of five rounds each.

5.8.3.2 Dependent variables

As with Study 1, the dependent variables were in-group favouritism, fairness rating and agent identification rating. These variables were calculated using the same formula as presented in Study 1.

5.8.4 Data Analyses

The same procedure as Study 1 was followed to analyse the data. Also, these analyses can be found in the online supplementary materials via the link here.

5.8.5 Result

The results of the study are reported in three sections, each responding to one hypothesis. Methods specific to each section are presented prior to the result. Therefore, the result is presented in the following order: (i) test of the hypothesis and analysis of in-group favouritism; (ii) test of the hypothesis and analysis of trust (via fairness rating); and (iii) test of the hypothesis and analysis of machine learning agent Identification.

5.8.6 In-group Favouritism

5.8.6.1 Method

Condition 0:0, the null condition, was used as the reference category for the model comparison. The intervals(), fixef() and summary() method in R were used to calculate 95% confidence intervals for each condition, standardised fixed effects for the conditions, and the summary output of the model respectively.

5.8.6.2 Multi-level model

Nesting structure was considered because of the hierarchical data. Two models were built to provide statistical evidence for nesting. The first, Model 1 (*AIC* = 728.0316, *logLik* = -362.0158) has no random effect while the second, Model 2 (*AIC* = 617.5354*, logLik* = -305.7677) includes random effects at the game level. A comparison of Model 1 and Model 2 shows that Model 2 was a better fit for the data ($p < 0.0001$). This indicates that random model terms are required. Model 3 (*AIC* = 550.2124, *logLik* = -271.1062) was built with a random effect at both the game and individual levels. A comparison of Model 2 and Model 3 shows that Model 3 was a better fit to the data ($p < 0.0001$).

The interclass correlation coefficient (ICC) calculated for each level of nesting indicated that 17% (ICC = 0.166) and 38% (ICC = 0.384) of unexplained variance lies at the game level and individual level, respectively. The former is considerably high while the latter is an extremely high amount of unexplained variance.

A first-order autoregressive structure "corAR1" was added to account for repeated measures over waves. This was not significant. However, it did not deteriorate the goodness-of-fit (*AIC* = 550.5159). Thus, corAR1 was added.

As with the multi-level analysis of in-group favouritism in Study 1, Model 4 was built with a fixed effect of condition, group and time added. Model 4 (*AIC* = 530.4957, *logLik* = -256.2479) further improved the goodness-of-fit ($p < 0.0001$).

Model 5 was built with two-way interactions added to Model 4. Compared to Model4, Model5 (*AIC* = 519.8335, *logLik* = -245.9168) significantly ($p < 0.001$) improved the goodness-of-fit.

Lastly, Model 6 was built with fixed effect, two-way and three-way interactions between condition, group and time. Model 6 (*AIC* = 521.5605, *logLik* = -244.7803) decreased the goodness-of-fit and was not statistically significant. Thus, Model 5 was reported. Table 5.9 presents the final model (i.e. Model 5).

Table 5.9. The output of the final model testing the effect of out-group altruism on in-group favouritism.

	numDF	**denDF**	**F-value**	**p-value**
(Intercept)	1	716	762.3867	<.0001
Condition	2	57	5.4466	0.0068
Fromgroupno	1	297	1.4753	0.2255
Roundno	1	716	17.2718	<.0001
Condition:fromgroupno	2	297	4.5652	0.0111
Condition:roundno	2	716	5.5952	0.0039
Fromgroupno:roundno	1	716	0.5256	0.4687

5.8.6.3 Condition and time as a main effect on in-group favouritism

Table 5.9 shows that *condition* was statistically significant while *group* was not. This is not surprising; in fact, it suggests that in-group favouritism might have been weakened overall and, for this reason, there was no difference between the groups. However, the interaction between *group* and *condition* was statistically significant (p = 0.0111). Figure 5.16 shows *condition* as a main effect on in-group favouritism.

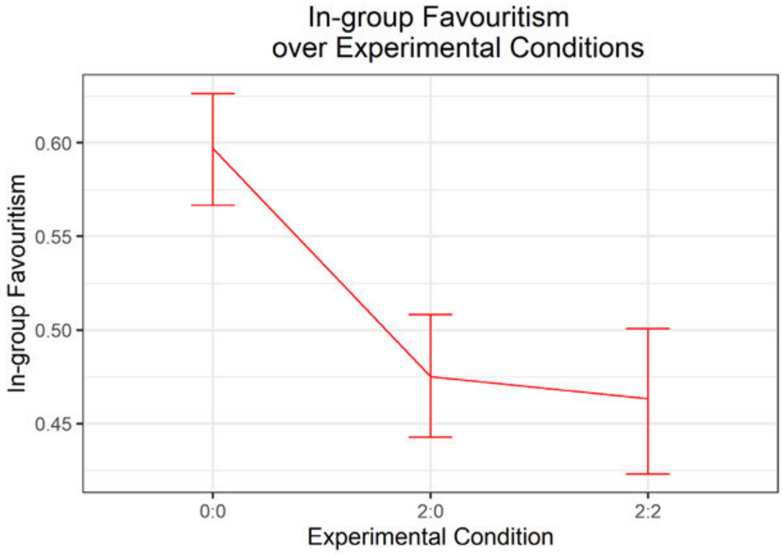


Figure 5.11. Differences in mean in-group favouritism between conditions.

The null condition (i.e. Condition 0:0) has high in-group favouritism. However, the high in-group favouritism in the null condition was significantly reduced in Conditions 2:0 and 2:2. There is no significant difference in in-ingroup favouritism in Conditions 2:0 and 2:2. This suggests that adaptive agents in one group could produce the desired effect. Thus, it is evident that agents in Study 2 were able to weaken in-group favouritism. Similarly to Study 1, in-group favouritism was high in the first wave but reduced over time in Study 2, as seen in Figure 5.17. Further descriptive analysis is given in the next section to explore the interaction effect of condition and group.

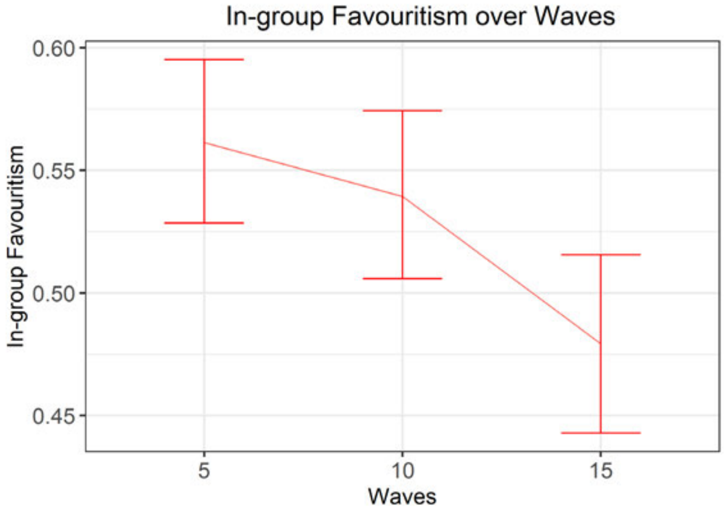


Figure 5.12. In-group favouritism over waves.

5.8.6.4 Descriptive analysis of the interaction effects

Figure 5.18 plots the graph of the interaction between *condition* and *round*. In-group favouritism is higher in the null condition. It starts high in the first wave and increases further in the second wave. Although it drops in the third wave, this drop was not significant. In-group favouritism is high at the first wave of Condition 2:0, but significantly decreases over time. The same trend was observed in Condition 2:2, which starts with less in-group favouritism compared to Condition 2:0. The trend implies that agents are able to reduce in-group favouritism, and this ability increases as interaction progress. In other words, the more the interaction progresses, the weaker in-group favouritism becomes.

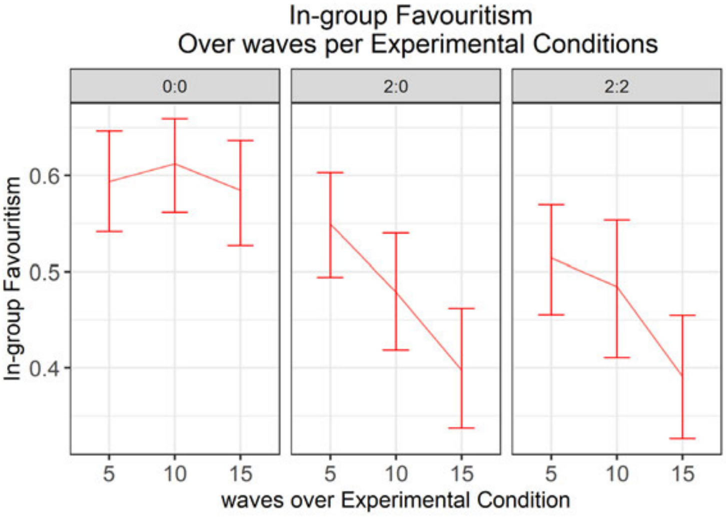


Figure 5.13. High in-group favouritism in the null condition, which persisted over time, versus in-group favouritism in Conditions 2:0 and 2:2, which reduced over time

Figure 5.19 plots in-group favouritism per group in each condition. Although Condition 2:0 reduced in-group favouritism compared to the same condition in Study 1, the reduced in-group favouritism found was mainly from Group 2, a group with no agent. In Condition 2:0, in-group favouritism is high in the group with two agents. This begs for further analysis to ascertain whether or not this was a result of humans punishing agents for defecting, as shown in Study 1.

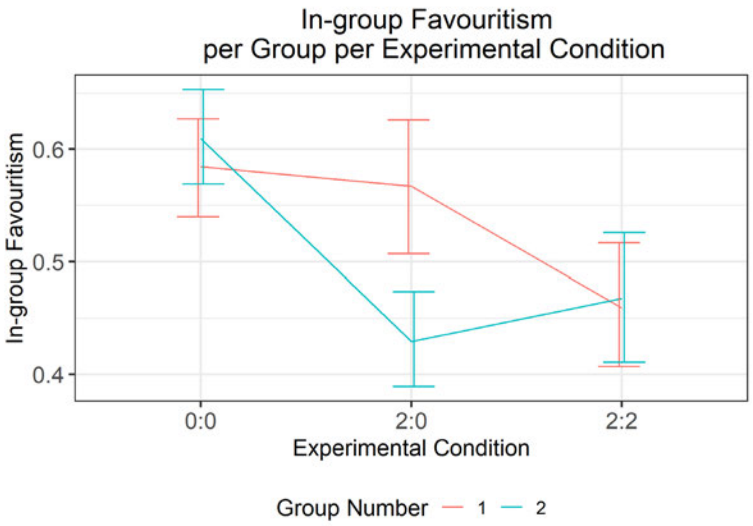


Figure 5.14. Differences in mean in-group favouritism between conditions per group. The first digit (i.e. 0, 2 and 2) in each experimental condition represents the number of agents in Group 1, while the second digit (i.e. 0, 0, and 2) represents the number of agents in Group2.

5.8.7 Fairness Rating

5.8.7.1 Method

As with Study 1, each fairness rating received by an agent or a human is in the range of 1 to 5. To examine participants' perception of fairness, the study creates an independent variable called *TargetGroup*, as described in Study 1. The study also creates another independent variable named *TargetParticipant*, as described in Study 1.

Additional independent variables

- **TargetGroup**: Each human player rated an in-group or out-group member.
- **TargetParticipant**: Each rating is received by either a human or an agent.
- **Condition**: There were three conditions: Condition 0:0, Condition 2:0, and Condition 2:2.

Dependent variable

- **Rating:** The rating received by each player.

5.8.7.2 Multi-level model

Again, the nesting structure was considered because players, which provide in-group or out-group ratings, were nested within games. Two models – Model 1 and Model 2 – were first built and compared to determine the need for random effects. Thus, Model 1 was built with no random effect while Model 2 was built with random effect at a game level. Model 2 (*AIC* = 4861.364, *logLik* = -2427.682) has a better fit than Model 1 (*AIC* = 5004.060). The difference was statistically significant ($p <$ 0.0001), which is evidence for nesting. Model 3 added random effect at both game and individual levels. Compared to Model 2, Model 3 improved the goodness-of-fit and was significant ($p <$ 0.0001).

An interclass correlation coefficient (ICC) for each level of nesting was calculated. The ICC at the game level is 0.139, which means 14% unexplained variance lies at the game level. This number is high enough to justify the use of a multi-level model. Also, the ICC at the individual level was 0.345, a fair amount, which means that 35% unexplained variance lies at the individual level.

As with Study 1, an autoregressive structure was not used because there was no time factor in the rating data. Thus, the next model (i.e. Model 4) was built with a fixed effect of condition, TargetGroup and TargetParticipant. Model 4 (*AIC* = 4711.299, *logLik* = -2348.650) was compared to Model 3. Although it was not statistically significant, the model slightly improved the goodness-of-fit. Moreover, it is the base model. Thus, it is used as a reference to compare subsequent models.

Model 5 was built with two-way interactions added to Model 4. Thus, Model 5 has both fixed effects and two-way interactions of the condition, TargetGroup and TargetParticipant. The model (*AIC* = 4714.448, *logLik* = -2347.224) did not improve the goodness-of-fit and was not significant

Lastly, Model 6 was built with fixed effect, two-way and three-way interactions of condition, TargetGroup and TargetParticipant. Model 6 (*AIC* = 4715.743, *logLik* = -2346.872) did not improve the goodness-of-fit and was not statistically significant. Thus, Model 4 was reported as the final model. Table 5.10 presents the model.

Table 5.10. The output of the final model for fairness rating.

	numDF	denDF	F-value	p-value
(Intercept)	1	1185	1140.3792	<.0001
Condition	1	38	0.9239	0.3425
TargetGroup	1	1185	0.3523	0.5529
TargetParticipant	1	1185	5.4441	0.0198

5.8.7.3 TargetParticipant as a main effect on fairness ratings

Although it has been proven that there was no difference between in-group and out-group ratings (shown by non-significance of *TargetGroup* variable), there is, however, a significant difference between the ratings received by agents and humans. Further descriptive analysis (Figure 5.20) shows that agents were rated as fairer than humans resulting to the observed significant difference.

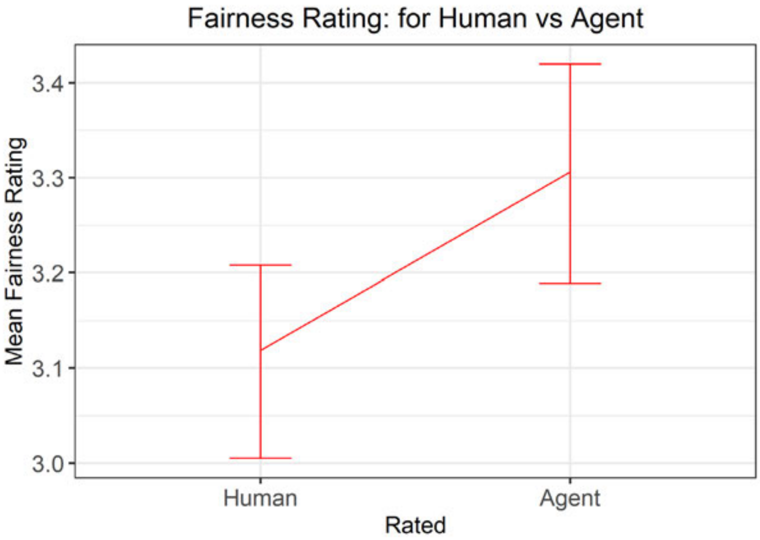


Figure 5.20. Showing that agents are rated as fairer than humans.

Compared to Study 1 where there is no significant difference between the ratings received by agents and humans, the fact that agents are rated as fairer than humans in Study 2 is an indication that the adaptive agents can exert social influence more than the rule-based agents.

Also, compared to Study 1 where agents were treated as out-group members, there was no significant effect of the interaction between in-group and out-group ratings in Study 2. This implies that adaptive agents (i.e. agents in Study 2) were perceived as fairer than rule-based agents. Thus, the adaptive agents have a better chance of influencing humans.

5.8.8 Agent Identification Rating

5.8.8.1 Method

The data for the agent identification rating was also pre-processed, as was done with the fairness rating data.

5.8.8.2 Multi-level analysis

As with the previous analysis, two models were built to ascertain that a random effect is required. Model 1 (*AIC* = 5165.340) with no random effect was built and compared to Model 2 with random effect at the game level. Model 2 (*AIC* = 5150.171, *logLik* = -2572.086) has a better fit than Model 1 and was statistically significant ($p < 0.0001$).

Model 3 was built with a random effect at both the game as well as the individual level. Surprisingly to the researcher, Model 3 (*AIC* = 5149.098, *logLik* = -2570.549) was not statistically significant compared to Model 2. However, Model 3 slightly improved the goodness-of-fit as seen by the AIC. Thus, subsequent models were built with random effects at both the game and individual levels

The interclass correlation coefficient (ICC) for each level of nesting was calculated. The ICC at the game level is 0.035, which means 4% unexplained variance lies at the game level. This is not a high value but could not rule out the use of a multi-level model. Also, the ICC at the individual level was 0.060, a fair amount which means that 6% unexplained variance lies at the individual level. This value is fair enough to justify the use of a multi-level model.

As with the fairness rating analysis, an autoregressive structure was not used because there was no time factor in the rating data. Next, Model 4 (*AIC* = 5143.925, *logLik* = -2564.963), built with a fixed effect of condition, TargetGroup and TargetParticipant, was compared to Model 3. Model 4 was statistically significant and slightly improved the goodness-of-fit (p = 0.0108)

Model 5 was built with two-way interactions added to Model 4. Model 5 (*AIC* = 5145.092, *logLik* = -2562.546) did not improve the goodness-of-fit and was not significant. Also, Model 6 (*AIC* = 5146.321, *logLik* = -2562.161), built with fixed effects, two-way and three-way interactions of condition, TargetGroup and TargetParticipant, did not improve the goodness-of-fit and was not statistically significant. Thus, Model 4 in Table 5.11 was reported as the final model.

Table 5.11. The output of the final model for agent identification rating.

	numDF	denDF	F-value	p-value
(Intercept)	1	1188	1651.1986	<.0001
Condition	1	38	0.2295	0.6347
TargetGroupBias	1	1188	0.4199	0.5171
TargetParticipant	1	1188	10.5454	0.0012

5.8.8.3 TargetParticipant as a main effect

Since there is a significant difference between the ratings received by humans and agents, the study performs descriptive analysis to further investigate this difference. As shown in Figure 5.21, the ratings received by humans are higher than those received by agents. Since 5 in the rating means that the participant thinks the player being rated is an agent, and 1 means that the participant thinks the player is a human, the graph, therefore, implies that participants were more confident that agents were human. Thus, it is evidence that agents were not identified as agents; instead, they were identified more as humans.

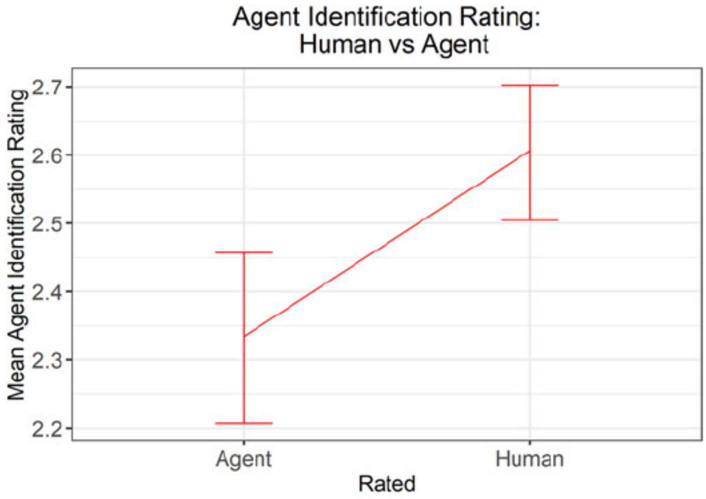


Figure 5.21. Depicts the differences in mean agent identification rating received by humans and agents.

5.8.9 Correlation between Trust and Agent Identification

As with Study 1, it was necessary to check for a correlation between fairness and agent identification for both humans and agents. Figure 5.22 shows the correlation between agent identification and fairness ratings for both humans and agents. Statistically significant negative correlations (r = -0.25, p = 0.0022) and (r = -0.38, p = 0.00053) were found in Condition 2:0

and Condition 2:2 respectively. However, both these correlations are for humans. There were no significant correlations between the rating tasks for the agents. Hence, there was no proof that participants indirectly identified the agents.

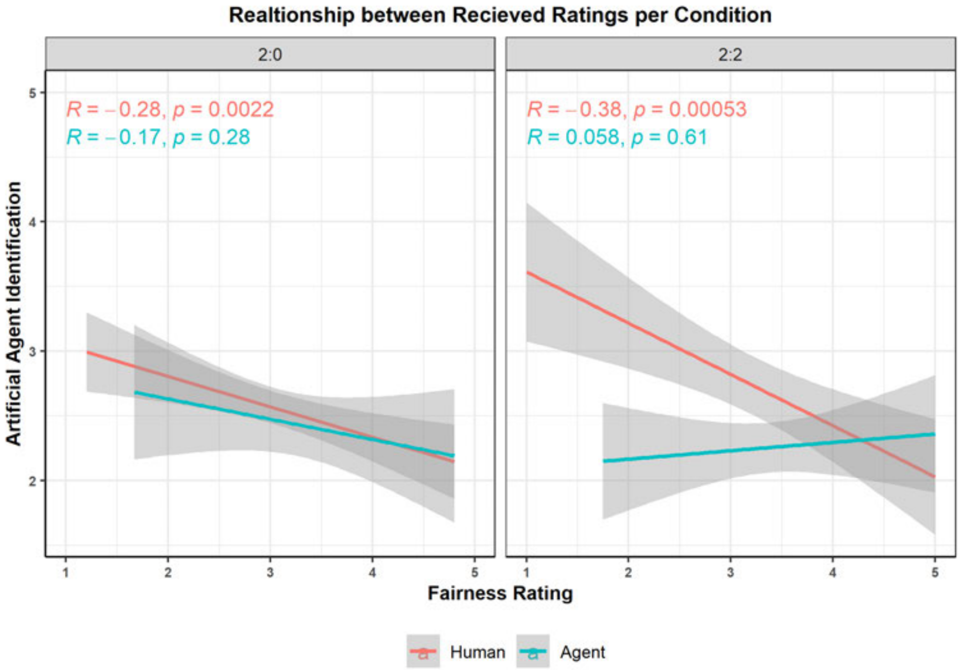


Figure 5.15. Correlation between fairness and agent identification.

It is noteworthy that the more humans are rated as fair, the more they were viewed as human, as shown by Figure 5.22. Relating this to the fairness rating in Study 2, where agents were rated as fairer than humans, Figure 5.22 confirms that adaptive agents were rated as being more human than human. Thus, when the agents behave fairly, they are viewed as human and stand a better chance of producing social interaction.

5.8.10 Further Test for Punishment and Reciprocity

As in Study 1, tests were conducted to: (i) test whether or not humans punished in-group agents for defecting (i.e. for exhibiting out-group altruism) and favoured other in-group members, and (ii) whether or not out-group members reciprocated to the other groups' agents who were favouring them.

Also, a multi-level analysis was performed. The first two models were built to have statistical evidence for nesting. The second model, model 2 (*AIC* = 5759.760, *logLik* = -2876.880) with random effects at the game level has a better fit ($p < 0.0001$) than model 1 (*AIC* = 6646.498, *logLik*

= -3321.249). Thus, it shows statistical evidence for nesting. However, the three-way interaction of the variables was not significant (p = 0.24872) as shown in Table 5.12 (see the *KCI 2021 Punishment Analysis* via: here).

Table 5.12. The output of the final model for punishment and reciprocity.

	NumDF	DenDF	F value	Pr(>F)	
Condition	1	236.83	30.6609	8.14E-08	***
Group	1	2681.46	5.2879	0.02155	*
GroupByParticipantType	3	2705.03	30.413	< 2.2e-16	***
Condition:Group	1	2687.29	2.0261	0.15473	
Condition: GroupByParticipantType	3	2717.45	8.4788	1.32E-05	***
Group: GroupByParticipantType	3	2685.27	1.543	0.20133	
Condition:Group: GroupByParticipantType	1	2691.97	1.331	0.24872	

* = significant at 0.05; *** = significant at 0.001

No further extrapolations were made of the non-significant three-way interaction. Instead, the two-way interaction between condition and *GroupByParticipantType* was significant (p < 0.0001) and therefore was presented in Figure 5.23.

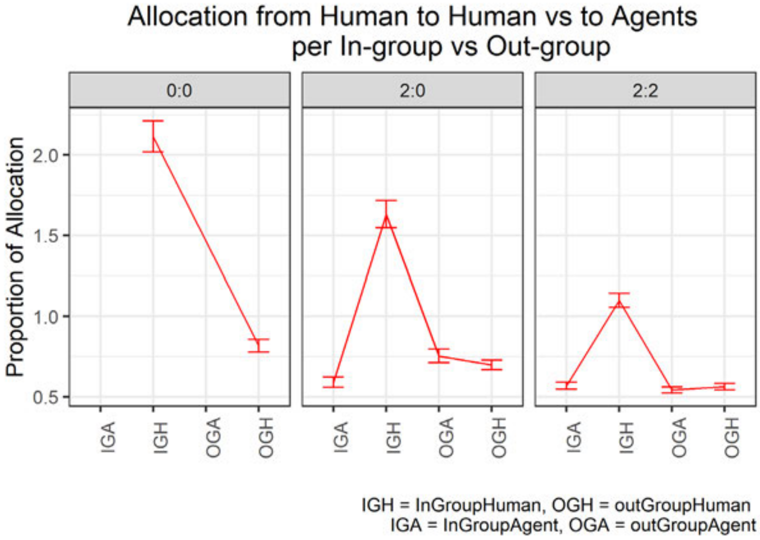


Figure 5.16. Allocation from humans to in-group and out-group humans and agents. The graph shows that in-group humans received more tokens across all conditions.

Similarly to Study 1, allocations from humans to in-group humans were the highest in all the conditions, as shown in Figure 5.23. However, this reduced significantly in Condition 2:2 of Study 2. The result of the analysis implies that adaptive agents in Study 2 were still punished for defecting, even though agents in some rounds allocated tokens to in-group. Also, the out-group reciprocated equally to the other group and not specifically to agents. This is shown by the non-significant difference between out-group agents and out-group humans.

5.8.11 Discussion

Study 2 aimed to investigate adaptive agents' ability to reduce in-group bias during interactive social exchanges. Adaptive agents were introduced to set the norm of out-group altruism via interactions (see Section 5.8) with both the in-group and out-group members. Specifically, Study 2 aimed to weaken in-group favouritism and promote out-group altruism in an intergroup context.

Study 1 had demonstrated that out-group altruism can strengthen in-group favouritism and that the defectors (perpetrators of out-group altruism) can be treated as out-group members, if the perpetrators fail to maintain a positive relationship with the in-group while associating with the out-group. Study 2 introduced out-group altruism while maintaining in-group reputation via adaptive agents that used the co-evolutionary model developed in Chapter 4 (see Section 5.8 the strategy).

Overall, Study 2 is similar to Study 1, as agents in Study 2 weaken in-group favouritism amongst the out-group but not among the in-group. Contrary to Study 1, humans in Study 2 did not punish agents irrespective of their groups, as indicated by the non-significant interaction between group and *GroupByParticipantType* (see Section 5.8.10).

The remainder of this section discusses the main findings from Study 2 as follows: (i) conformity in the face of perceived fairness, and (ii) adaptive agents' interactions with humans.

5.8.11.1 Conformity in the face of perceived fairness

The multi-level model in this study shows that time (i.e. round number) was a statistically significant variable affecting in-group favouritism, as the main effect and in interaction with other variables. Further examination showed that in-group favouritism decreased over time in Conditions 2:0 and 2:2. In Condition 2:0, which has two agents in Group 1 and no agents in Group 2, it was observed that in-group favouritism slightly increased over rounds in the first wave (rounds 1 to 5). Indeed, this increase in in-group favouritism occurred because agents do not yet have good reputation with the in-group. In other words, they are not yet viewed as being fair. Thus, wave one was the reputation-building stage. Once agents built reputations with the in-group members, the in-group members tend to adopt out-group altruistic norms as set by these agents. This resulted in a sharp decrease in in-group favouritism. Although the null condition shows a slight decrease in the in-group favouritism in the last wave, the slopes in Conditions 2:0 and 2:2 are very steep compared to those found in the null condition (Condition 0:0).

As proof of the difference in the in-group favouritism, the multi-level analysis shows that Conditions 2:0 and 2:2 are significantly different from the null condition. Descriptive analysis shows that agents were perceived as human, not only by the in-group members but also by the out-group members. Thus, the perception of fairness becomes the basis on which agents were viewed as more human than humans, as explained by the correlation between fairness rating and agent rating.

Unlike Study 1, where agents were perceived as not fair by in-group members, agents in Study 2 reduced in-group favouritism amongst out-group members and were perceived as fairer than humans. These findings, therefore, support (partly) the hypothesis that agents will produce higher reductions in in-group favouritism amongst both in-group and out-group members over time (Hypothesis 1). Also, it supports that agents will be rated as fairer than humans (Hypothesis 2). Furthermore, the study supports that when an out-group member has a good reputation, behavioural responses from the in-group are less hostile (De Dreu & Kret, 2016; Ellemers & van Nunspeet, 2020).

5.8.11.2 Adaptive agents' interactions with humans

In the computer science field, a Turing test is used to ascertain whether or not a computer can imitate a human to the extent that humans are unable to differentiate between humans and machines (Saygin et al., 2000). This study applied this concept via agent identification rating. It expects that there will be no difference between the ratings received by the adaptive agents and those received by humans. That is, it is expected that agents will not be detected, especially if these agents would successfully set the out-group altruistic norms. Indeed, their ability to weaken bias (i.e. reduce in-group favouritism) in the exchange system is the first evidence that they were not detected.

Unlike in Study 1, where the agents are controlled by a static rule, the adaptive agents' ability to weaken bias is rooted in the fact that the agents were adaptive – their exchange decisions are based on the use of the co-evolutionary model to predict the state (strategies and exchange decisions) of the exchange system. In turn, the prediction guides the agents' ability to decide when and whether or not to allocate a token to an in-group member. Although this study assumes that the agents' ability to guide their token allocation based on the prediction of human exchange decision, was instrumental in weakening in-group favoritism, the fact that that adaptive agents in Study 2 were still punished for defecting begs for further research to ascertain other factors that may have contributed to the reduced in in-group favouritism.

Although Study 2 found a significant difference between the ratings received by agents and those received by humans, further probing revealed that the significant difference occurred because agents were rated significantly as humans while humans were rated more as agents. Thus, the difference provides strong support that agents were successfully hidden among humans within the exchange system. Hypothesis 3 is therefore supported.

5.8.12 Implication

This study has several implications for research on fairness, reputation, in-group favouritism and the application of machine learning-based agents in psychology experiments. Furthermore, it has implications for interdisciplinary research.

Firstly, the study advances research on fairness and reputation. It reinforces research on fairness and its implications for intergroup relations. It also provides evidence that supports Lee et al. (2021) assertion that the perception of fairness within an exchange context reinforces trust (which leads to a good reputation).

Secondly, the study demonstrates the use of machine learning-based agents as experimental 'stooges' in weakening in-group favouritism while avoiding punishment. It shows that machine learning agents were able to interact adaptively with both the in-group and the out-group members, thereby setting the out-group altruistic norms that partly reduced in-group favouritism. Research could be developed towards improving the algorithm that controls the agents' decision-making.

Finally, this research shows the importance of the relationship between fairness and being human. With regard to interdisciplinary work, Study 2 is exceptionally important in humanoid research. The perception of fairness can enhance interaction with humanoid robots. This can enhance research in other fields such as healthcare, where humanoid robots interact with patients.

5.8.13 Study 2 Summary

Study 2 investigated adaptive agents' ability to reduce in-group bias in interaction. The agents, which make use of the co-evolutionary model (Chapter 4), were introduced to set out-group altruistic norms via interactions with both the in-group and out-group members. Specifically, the study aimed to weaken in-group favouritism by setting out-group altruism in interaction over time. The result shows that the agents were perceived as fairer than humans and, as such, were rated as more human than humans. This resulted in the agents maintaining a good reputation and they were not treated as out-group members. The findings show that humans could not identify the agents. Thus, the result supports all the three hypotheses defined, but Hypothesis 1 was only partly supported.

5.9 Limitation

Limitations to be considered when reading the findings are mainly the design and non-generalisability of the study. Firstly, the study was conducted in the Visual Interaction Application platform, which, like most laboratory experiments, allows interaction in a very minimal context, thus limiting the range of possible behaviour that can be observed in humans. The findings, therefore, may not be generalisable to real-world contexts involving a higher degree of complexity. Secondly, although there were manipulation checks for the experimental manipulations, this was not extensively done. Thus, this study cannot guarantee that participants understood the manipulations and played the game to their best understanding. Thirdly, due to the age range of the students, their behaviours may not be representative of the broader population, but rather of a particular social group. Thus, a different result may be obtained if another random sample were used rather than a convenience sample.

Study 2 used a co-evolutionary model developed by integrating an artificial neural network and a hidden Markov model. The internal processes of the algorithms may not be fully explainable. Thus, there may be variables that influence the prediction, but are not explained or accounted for due to the nature of the algorithm.

5.10 Conclusion

This chapter investigated agents' ability to set out-group altruistic norms in order to weaken in-group favouritism. Two studies, Study 1 and 2, evaluated agents that used two different strategies – rule-based altruistic behaviour and adaptive altruistic behaviour – for weakening in-group favouritism. In Study 1, out-group altruism was implemented via artificial agents that have been pre-programmed to allocate their wealth to out-group members randomly in the game. Thus, Study 1 makes use of rule-based agents. The study partially supports the initial expectation that out-group altruism is capable of weakening in-group favouritism. The findings of Study 1 are that out-group altruism reduces in-group favouritism among out-group members but not among in-group members.

Contrary to Study 1, Study 2 introduced out-group altruism via adaptive and adaptive agents that make use of a machine learning-based model to predict human exchange decisions. Based on the predictions, the agents allocate their tokens either to those predicted to have a strong affinity for reciprocity and in-group favouritism or to the agent's in-group member. The findings support that the agents were able to weaken in-group favouritism in interaction as a result of their reputation, which was evidenced by agents being rated as more human than the humans.

CHAPTER 6. GENERAL CONCLUSION

Emergence is when “the high-level phenomenon arises from the low-level domain,but truths concerning that phenomenon are unexpected given the principles governing the low-level domain”.

David Chalmers (2006, p. 244)

6.1 Summary of the Study

6.1.1 Rationale

Beyond experiments, surveys and interviews which have been successfully used to study behaviours, simulations and computer-mediated experiments are methodologies that have been applied to studying the dynamics of social interaction. However, these methodologies are yet to be fully developed to cater for the dynamics of interactive social exchanges.

6.1.2 Aims and objectives

The dual aim of this dissertation was to: (i) develop adaptive agents that can interact in an information-scarce interactive social exchange environment, and (ii) to evaluate their usefulness in creating positive outcomes – reduce group bias – during interactive social exchanges. Accomplishing this dual aim required a series of sub-aims and objectives that were realised by making sense of the complexity of interactive social exchange and its dynamics, which are centred on emergence, and the complex and obscure motives of the interacting individuals.

6.1.3 Methodology

This dissertation advanced simulations and computer-mediated experiments by taking advantage of the advancements in the field of machine learning. It combined these advancements with well-established theoretical frameworks within social psychology to create efficient and effective ways of capturing and intervening in the dynamics of interactive social exchanges. Additionally, the research work presented in this dissertation considered that technological advancements have geared interest toward big data, while neglecting areas where data are rarely available – information-scarce environments.

6.1.4 Outcome

Overall, the dissertation: (i) shows how obscure motives and objectives in an interactive social exchange can be revealed, based on theoretical knowledge, thus, advancing research on modelling and simulation in an information-scarce environment; (ii) it demonstrates that artificial agents can be seamlessly integrated into human exchange environments and used to affect behaviours; (iii) the dissertation confirms that in-group members punish defectors (Study 1), and these defectors are more likely not to be imitated; (iv) furthermore, it shows that in-group members do not punish defectors who build a reputation by associating with (through token allocation to) the in-group, but such association does not guarantee that the defector will be imitated; and (v) the dissertation shows that the perception of fairness influences beliefs about being human.

The rest of this chapter elaborates on the innovativeness, contributions and findings of this dissertation.

6.2 Key Outcome

Understanding and modelling behaviours in an information-scarce environment are non-trivial – they require anticipation and consideration of possible phenomena that may emerge and should allow a feedback loop where individuals in the system internalise the emergent phenomenon before they act.

Considering the assertion that "the designers and users of a system do not have enough knowledge about why and how they *[the emergent phenomena]* occur in the system" (Kalantari et al., 2020, p. 253), this dissertation demonstrated the application of simulation that harnesses its intelligence to improve the understanding of the emergent phenomena during interactive social exchanges in an information-scarce environment.

6.3 Methodological Innovation

Studying the dynamics of interactive social exchange, and consequently harnessing the understanding to devise an intervention, is a complex undertaking. After all, the dynamics of interactive social exchange are non-trivial to predict, and emergence is fundamentally constrained by a restricted environment. Thus, to effectively build and test a model of interactive social exchange in an information-scarce environment, the study adapted and applied existing methodologies (model-building, data collection and experiments) in a novel way. Through the application of the model, the dissertation has advanced ongoing investigations into ways to weaken the intergroup bias which has pervaded our society.

6.3.1 'Model-building' - Building a Model of Interactive Social Exchange

An interactive social exchange model was developed by combining clustering (offline), an artificial neural network, and the hidden Markov model to more accurately predict exchange decisions in an interactive social exchange context. Although these algorithms have long been in existence, and their distinct applications are numerous in many fields, their applications in an interactive social exchange context are rare. This study introduced a novel method that combined the triad, such that each component contributed to an effective simulation that discovers, incorporates and uses emergent phenomena to improve the prediction of exchange decisions. This combination led to the realisation of the objectives set out in the dissertation.

These objectives proceeded from advancing the understanding of interactive social exchanges and their applications (Objective 1 – in Chapters 2 and 3), to developing and evaluating a prototype model that was built based on the suggestions in Chapter 1 (Objectives 2 and 3 – in Chapter 4). Furthermore, the prototype was integrated and applied (via an agent-based model) to create interventions in a social exchange environment (Objective 4 – in Chapter 5) and to compare the result to a similar application but with a rule-based agent.

- ***Objective 1: Advance the understanding of interactive social exchange as a system of pure generalised exchange and the importance of studying the system.***

  In Chapter 2, this dissertation explained the importance of studying interactive social exchange via simulation, taking the view that direct and generalised exchanges are combined in a pure generalised exchange. It shed light on the complexity of interactive social exchanges and pointed out three features that form the pillars of interactive social exchange models: (i) emergence, (ii) complex and obscure motives, and (iii) the context of behaviour. Furthermore, the dissertation suggested mechanisms that will advance the study of interactive social exchange.

- ***Objective 2: Develop a model for decision-making in an interactive social exchange, based on the co-evolution of two algorithms: the artificial neural network (ANN) – a biologically inspired algorithm, and the hidden Markov model (HMM) – a statistical modelling tool.***

  In line with the recommendation and suggestions of the previous studies (Chapter 2 and Chapter 3), Chapter 4 developed a model that incorporated three modules – a clustering module, an artificial neural network module and the hidden Markov model – to capture and make sense of emergent behaviours (clustering), classify exchange strategies while taking the context of behaviour into account (artificial neural network), and predict exchange decisions given the strategy and the history (the hidden Markov model). The combination was designed such that when integrated into an experimental environment (see Chapter 5) the neural network's fitness is improved by the output of the hidden Markov model and vice versa.

- ***Objective 3: Evaluate the developed model on the data collected from Visual Interaction APPLication (VIAPPL) experiments, where VIAPPL – a computer-mediated environment – has been customised to represent an information-scarce environment.***

  With regard to model evaluation, the data used were obtained from an experiment conducted in the Virtual Interaction APPLication (VIAPPL; see www.viappl.org) arena and published in (Durrheim et al., 2016). The model was calibrated using various performance evaluation methods. The model performed fairly well on predicting exchange decisions but performed better when targeted for a simple decision such as predicting whether or not an individual would reciprocate during interactive social exchanges.

6.3.2 'Data Collection and Experiments' - Integrating and Evaluating the Model on **Visual Interaction Application**

To fully realise the objectives set in Chapter 1, a study was conducted to evaluate the impact of the model on reducing in-group favouritism during intergroup experiments. Thus, in terms of innovativeness in *data collection and experiments*, this section of the dissertation lists the objectives and how they were realised:

- ***Objective 4: Develop and integrate agents that use the model in an interactive social exchange environment.***

  The model (in Chapter 4) was developed and integrated as a sub-module in Visual Interaction Application. The avatar (representing players in the Visual Interaction Application arena) makes decisions based on the output of the model. Thus, a machine learning agent-based model of interactive social exchange was developed.

  Objective 5: Conduct experiments comprised of humans and artificial agents to understand and influence humans' exchange decisions towards reducing in-group favouritism during interactive social exchanges.

  To test the model integrated as a sub-module in Visual Interaction Application, the study conducted interactive social exchange experiments involving computer agents and humans.

Firstly, the researcher developed a rule for agents whose behaviour has been pre-programmed to be altruistic. These rule-based agents were integrated into Visual Interaction Application as a means of introducing out-group altruism during interactive social exchanges (Chapter 5). The findings of the experiments, which had a sample size of 280 human participants, show the need for adaptive agents.

Then, interactive social exchange experiments involving adaptive agents and humans were conducted (Chapter 5). The agents' strategy was to break the bounded generalised reciprocity in the exchange system. Thus, the agents predict human exchange decisions; based on the prediction, agents allocate their tokens to out-group members predicted to have an affinity for reciprocity. The findings of the experiments, which had a sample size of 360 human participants, show that the novel machine learning agent-based model successfully reduced in-group favouritism amongst out-group members more than it did amongst in-group members in the intergroup exchange system.

6.4 Summary of the Contributions and Key Findings across Chapters

Three categories of studies – reviews (Chapter 2 and Chapter 3), implementations (Chapter 4) and experiments (Chapter 5) – were conducted for this dissertation. Thus, the contributions and key findings are presented below, following these three categories.

6.4.1 Review

The dissertation contributes to the body of knowledge advancing the understanding of social exchange. The contributions are pointed out below:

- **Contribution**: In Chapter 2, pure generalised exchange was analysed as the combination of direct and generalised exchanges.
- **Contribution**: Emergence, complex and obscure motives, and context of behaviour were identified as the key features of interactive social exchanges.
- **Contribution**: The dissertation makes recommendations and suggestions on how emergence, complex and obscure motives, and context of behaviour can advance the

understanding of the social exchange system.

Further to the contribution in Chapter 2, the dissertation advanced the methodological approach available for modelling social exchange as described below.

- **Contribution**: Recognising that domains exist where data are rarely available – an information-scarce environment – Chapter 3 suggested a novel way of combining machine learning and psychology theory to effectively incorporate emergent behaviours, motives, and context of behaviour in social exchange models. Although, such combinations exist and are not new (Brearcliffe & Crooks, 2021), to the best of the author's knowledge, this dissertation is the first to suggest the combination in an information-scarce interactive social exchange.

6.4.2 Implementations

Taking the suggestions made in Chapter 2 and Chapter 3, Chapter 4 implemented a machine learning agent-based model – a novel model of social exchange that incorporates a clustering algorithm, an artificial neural network and a hidden Markov model.

- **Contribution**: To the best of the author's knowledge, this is the first study that implemented a machine learning agent-based model to simulate decision-making in an information-scarce interactive social exchange context.
- **Finding**: The model's evaluation with the previously collected data from Visual Interaction Application experiments showed that the model can simulate exchange decisions, although it can be improved further for more accurate predictions.

6.4.3 Experiments

Two studies – Study 1 and Study 2 (both reported in Chapter 5) were conducted that investigated the use of social exchange models to create interventions – specifically, to reduce in-group favouritism.

6.4.3.1 Contributions and findings based on Study 1

- **Contribution**: Study 1 in Chapter 5 investigated out-group altruism for reducing in-group favouritism. To the best of the author's knowledge, this is the first study that conducted this investigation. Out-group altruism was introduced via a rule-based model integrated into Visual Interaction Application.

- **Finding**: The rule-based agents promoted in-group favouritism among the human in-group players, who resisted rather than conformed. While this study established that rule-based agents promoted in-group favouritism, it does not refute that some static behaviours, especially a prosocial behaviour such as constant allocation of tokens to the poor to promote their status, may reduce in-group favouritism. Thus, this finding implies that more research needs to be conducted to establish whether or not social exchanges thrives when behaviours are static.
- **Finding**: In-group members perceived the agents' altruistic behaviour as undesirable This perception increased the in-group members' shared social identity, which in turn led to a need to support each other to resist the undesired behaviour. This supports the bounded generalised reciprocity theory, which suggests that defectors get a bad reputation among in-group members.
- **Finding**: The in-group members treated the rule-based agents as out-group members, thus promoting the importance of maintaining group identity while promoting out-group altruism.

6.4.3.2 Contributions and findings based on Study 2

Chapter 5 Study 1 shows the need for introducing out-group altruism via adaptive agents; thus, the second experiment in Chapter 5 was conducted. The contributions and findings are as below:

- **Contribution**: This is the first study that introduced out-group altruism via machine learning-based agents, where the agents were trained to act 'adaptively' by first predicting humans' exchange decisions. Based on this prediction, the agents allocate their tokens to: (i) out-group members who are predicted to reciprocate and who have a high affinity for in-group allocation, and/or (ii) to an in-group member, when the in-group allocation practice is not historical.
- **Finding**: Agents in Study 2 reduced in-group favouritism amongst the out-group members. Although there was no significant reduction amongst the in-group members, the agents were not treated as out-group members. This shows the effectiveness of using the knowledge of the system and emergent behaviours as a

guide for action in a social exchange context.

- **Finding**: Agents in Study 2 were rated as fairer than humans, and the correlation between fairness ratings and agent identification ratings confirmed that agents were viewed as more human than humans.

6.5 Summary of the Implications

Research presented in this dissertation has several implications, some of which have been discussed. These implications are summarised, first for the study of social exchange via simulations and, second, for intergroup research focusing on devising interventions to quell intergroup bias.

Interactive social exchanges are highly stochastic and require models that can swiftly adapt to the changing objectives and intentions of the interacting individuals. Models with such capability, however, are non-trivial to build. This dissertation has demonstrated that processing theoretical knowledge into data is useful for predicting exchange behaviour in an information-scarce environment. This dissertation further demonstrated the need for more collaboration between the computer science discipline and the social psychology discipline.

Whereas intergroup relationship has been well researched, experimentally devising interventions that can be used to inform policies is still an area that requires more work. Focusing on out-group altruism, this dissertation has demonstrated that out-group altruism is harmful to intergroup relationships, when the individual exhibiting the behaviour is not perceived as being fair. However, when the individual is perceived as being fair, out-group altruism fosters improved intergroup relationships.

6.6 Limitations

Whereas this dissertation has contributed to the body of knowledge in computer science, social psychology and computational social science, it is important to discuss its limitations. First, social interactions have multiple static and dynamic factors that interact with each other and affect the outcome of the model. There may be factors affecting social interaction which were not captured by the model.

Due to the stochastic nature of the hidden Markov model, as well as other factors such as the computer system capacity and configurations, this dissertation cannot guarantee that the same result will be obtained if another simulation is conducted using the model. Although robust means of calibrating the performance of the model were used to ensure that the result reflected the actual performance, future work needs to consider testing the model on different platforms to evaluate whether or not these affect the performance.

In Chapter 5, the study measured the effect of independent variables such as conditions and groups on in-group favouritism. However, other factors – static and dynamic – exist which interact with one another to affect the outcome of the model. Thus, the study did not explore all the possible factors, but only those that are of interest to the study.

Due to the COVID-19 outbreak, experiments reported in Study 2 of Chapter 5 were conducted online. The participants were not always physically present with the experimenter or the assistant. Thus, this dissertation cannot guarantee that the choices made were solely those of the participant. For example, a participant may have been sitting with a friend, who may have made an exchange decision for the participant during the experiment.

Second, in the study, an information-scarce environment was represented in Visual Interaction Application using the minimal group paradigm, where group identities were the only information available to the participants prior to the game. In reality, an information-scarce environment may contain less or more information than depicted in Visual Interaction Application. Future work needs to consider data from a real-world environment.

6.7 Conclusion

This study aimed to develop co-evolutionary data-driven (machine learning) agents that can act in an information-scarce interactive social exchange context and to use these agents to weaken the in-group favouritism that occurs during interactive social exchanges.

This dissertation takes the perspective of social interaction as a social exchange. Thus, it first explained social exchange and its forms – direct and generalised exchanges – and ways to develop an effective model of interactive social exchange. A co-evolutionary model was developed, which takes into account motives, emergent behaviour and the context of interaction, to predict exchange behaviours within the interactive social exchange. The model was evaluated using data from the previous Visual Interaction Application experiments. Although the performance can be improved, the model was able to predict exchange decisions.

A rule-based model was developed and integrated into the Visual Interaction Application environment as a benchmark for the co-evolutionary model. Agents, using the rule-based model, were pre-programmed to be altruistic and were used to investigate out-group altruism for weakening in-group favouritism. Thus, experiments were conducted in the Visual Interaction Application environment where participants and agents interacted. These agents created undesired effects – in-group members treated agents as out-group members. This resulted in strengthening in-group social identity rather than setting out-group altruistic norms. Agents using the co-evolutionary model were tested in similar experiments. As with rule-based agents, the findings show that the agents were able to reduce in-group favouritism among the out-group. Contrary to being treated as out-group members, the agents were perceived as being fairer than humans and rated as being more human than humans.

References

Abbink, K., & Harris, D. (2019). In-group favouritism and out-group discrimination in naturally occurring groups. *PloS ONE*, *14*(9), e0221616.

Abrams, D., & Hogg, M. A. (2010). Social identity and self-categorization. In J. F. Dovidio, M. Hewstone , P. Glick, & V. M. Esses (Eds.), *The SAGE handbook of prejudice, stereotyping and discrimination* (pp. 179-193). Sage.

Adam, C., & Gaudou, B. (2016). BDI agents in social simulations: A survey. *The Knowledge Engineering Review*, *31*(3), 207-238.

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*(7), 899-917.

Akay, Ö., & Yüksel, G. (2018). Clustering the mixed panel dataset using Gower's distance and k-prototypes algorithms. *Communications in Statistics – Simulation and Computation*, *47*(10), 3031-3041.

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. In M. W. Berry, A. Mohamed, & B. W. Y. Yap (Eds.), *Supervised and unsupervised learning for data science* (3-21). Springer.

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

An, L. (2012). Modeling human decisions in coupled human and natural systems: Review of agent-based models. *Ecological Modelling*, *229*, 25-36.

Anandika, A., Mishra, S. P., & Das, M. (2021). Review on usage of Hidden Markov Model in natural language processing. In D. Mishra, R. Buyya, P. Mohapatra, & S. Patnaik (Eds.), *Intelligent and cloud computing* (pp. 415-423). Springer.

Anderson, J. R., & Lebiere, C. J. (2014). *The atomic components of thought*. Psychology Press.

Andión, J., Dueñas, J. C., & Cuadrado, F. (2021). D*iscrete-event-simulation based on machine learning predictive agents*. International Workshop on Soft Computing Models in Industrial and Environmental Applications (pp. 609-619). Springer.

Augustijn, E.-W., Kounadi, O., Kuznecova, T., & Zurita-Milla, R. (2019). *Teaching agent-based modelling and machine learning in an integrated way*. GeoComputation 2019. https://auckland.figshare.com/articles/Teaching_Agent-Based_Modelling_and_Machine_Learning_in_an_integrated_way/9848804

Augustijn, P., Abdulkareem, S. A., Sadiq, M. H., & Albabawat, A. A. (2020). *Machine learning to derive complex behaviour in agent-based modelizing*. 2020 International Conference on Computer Science and Software Engineering (CSASE) (pp. 284-289). https://doi.org/10.1109/CSASE48920.2020.9142117

Axelrod, R. (1981). The emergence of cooperation among egoists. *American Political Science review*, *75*(2), 306-318.

Axelrod, R. (1997). Advancing the art of simulation in the social sciences. In P. Davidsson & H. Verhagen (Eds.), *Simulating social phenomena* (pp. 21-40). Springer.

Balke, T., & Gilbert, N. (2014). How do agents make decisions? A survey. *Journal of Artificial Societies and Social Simulation*, *17*(4), 13. https://www.jasss.org/17/4/13.html

Balliet, D., & Van Lange, P. A. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin*, *139*(5), 1090-1112.

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin*, *140*(6), 1556-1581.

Bandura, A. (1986). *Social foundations of thought and action*. Prentice-Hall.

Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics*, *11*(2), 122-133.

Bearman, P. (1997). Generalized exchange. *American Journal of Sociology*, *102*(5), 1383-1415.

Bedau, M. A. (1997). Weak emergence. *Philosophical Perspectives*, *11*, 375-399.

Blau, P. (2017). *Exchange and power in social life*. Routledge.

Blau, P. M. (1964). *Exchange and power in social life*. Transaction Publishers.

Bonabeau, E., & Dessalles, J.-L. (1997). Detection and emergence. *Intellectica*, *25*(2), 85-94.

Bordini, R. H., Dastani, M., Dix, J., & Seghrouchni, A. E. F. (2007). *Programming multi-agent-systems.* 4th International Workshop, ProMAS 2006, Hakodate, Japan, May 9, 2006, Revised and Invited Papers (Vol. 4411). Springer.

Bornstein, G. (2003). Intergroup conflict: Individual, group, and collective interests. *Personality and Social Psychology Review*, *7*(2), 129-145.

Borshchev, A., & Filippov, A. (2004). From system dynamics and discrete event to practical agent based modeling: reasons, techniques, tools. *Proceedings of the 22nd International Conference of the System Dynamics Society*, July 25 - 29, 2004, Oxford, England. https://www2.econ.iastate.edu/tesfatsi/systemdyndiscreteeventabmcompared.borshchevfilippov04.pdf

Bourgin, D. D., Peterson, J. C., Reichman, D., Griffiths, T. L., & Russell, S. J. (2019). Cognitive model priors for predicting human decisions. *arXiv preprint arXiv:1905.09397*.

Brearcliffe, D. K., & Crooks, A. (2021). Creating intelligent agents: Combining agent-based modeling with machine learning. *Proceedings of the 2020 Conference of The Computational Social Science Society of the Americas* (pp-31-58). https://easychair.org/publications/preprint/w3H1

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues*, *55*(3), 429-444.

Brochu, P. M., Banfield, J. C., & Dovidio, J. F. (2020). Does a common ingroup identity reduce weight bias? Only when weight discrimination is salient. *Frontiers in Psychology*, *10*, Article 3020.

Broersen, J., Dastani, M., Hulstijn, J., Huang, Z., & van der Torre, L. (2001). The BOID architecture: Conflicts between beliefs, obligations, intentions and desires. *Proceedings of the Fifth International Conference on Autonomous Agents* (pp. 9-16). https://doi.org/10.1145/375735.375766

Broersen, J., Dastani, M., Hulstijn, J., & van der Torre, L. (2002). Goal generation in the BOID architecture. *Cognitive Science Quarterly*, *2*(3-4), 428-447.

Capraro, V., Jordan, J. J., & Rand, D. G. (2014). Heuristics guide the implementation of social preferences in one-shot Prisoner's Dilemma experiments. *Scientific Reports*, *4*, Article 6790.

Carley, K. M., Prietula, M. J., & Lin, Z. (1998). Design versus cognition: The interaction of agent cognition and organizational design on organizational performance. *Journal of Artificial Societies and Social Simulation*, *1*(3), 1-4.

Castelfranchi, C., Dignum, F., Jonker, C. M., & Treur, J. (1999). Deliberative normative agents: Principles and architecture. *International Workshop on Agent Theories, Architectures, and Languages* (pp. 364-378). https://doi.org/10.1007/10719619_27

Chae, J., Kim, K., Kim, Y., Lim, G., Kim, D., & Kim, H. (2022). Ingroup favoritism overrides fairness when resources are limited. *Scientific Reports*, *12*(1), Article 4560.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7-19.

Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence*, *42*(2-3), 213-261.

Collins, M. G., Juvina, I., & Gluck, K. A. (2016). Cognitive model of trust dynamics predicts human behavior within and between two games of strategic interaction with computerized confederate agents. *Frontiers in Psychology*, *7*, 49.

Cook, K. S., Cheshire, C., Rice, E. R., & Nakagawa, S. (2013). Social exchange theory. In J. DeLamater & A. Ward (Eds.), *Handbook of social psychology* (pp. 61-88). Springer.

Corning, P. A. (2002). The re-emergence of "emergence": A venerable concept in search of a theory. *Complexity*, *7*(6), 18-30.

Crisp, R. J., Walsh, J., & Hewstone, M. (2006). Crossed categorization in common ingroup contexts. *Personality and Social Psychology Bulletin*, *32*(9), 1204-1218.

Cruwys, T., Greenaway, K. H., Ferris, L. J., Rathbone, J. A., Saeri, A. K., Williams, E., & Grace, L. (2021). When trust goes wrong: A social identity model of risk taking. *Journal of Personality and Social Psychology*, *120*(1), 1-90.

Dancy, J. (2018). *Practical shape: A theory of practical reasoning*. Oxford University Press.

Dang, J., Liu, L., Zhang, Q., & Li, C. (2019). Leaving an attacked group: Authoritative criticism decreases ingroup favoritism. *Journal of Pacific Rim Psychology*, *13*, Article e7.

De Dreu, C. K., Fariña, A., Gross, J., & Romano, A. (2022). Prosociality as a foundation for intergroup conflict. *Current Opinion in Psychology*, *44*, 112-116.

De Dreu, C. K., Gross, J., Fariña, A., & Ma, Y. (2020). Group cooperation, carrying-capacity stress, and intergroup conflict. *Trends in Cognitive Sciences*, *24*(9), 760-776.

Deguet, J., Demazeau, Y., & Magnin, L. (2006). Elements about the emergence issue: A survey of emergence definitions. *ComPlexUs*, *3*(1-3), 24-31.

Deneubourg, J.-L., Lioni, A., & Detrain, C. (2002). Dynamics of aggregation and emergence of cooperation. *The Biological Bulletin*, *202*(3), 262-267.

Diehl, M. (1990). The minimal group paradigm: Theoretical explanations and empirical findings. *European Review of Social Psychology*, *1*(1), 263-292.

Dignum, M. (2004). *A model for organizational interaction: Based on agents, founded in logic*. SIKS.

Dimuro, G. P., Costa, A. C., & Goncalves, L. V. (2010). On the problem of recognizing and learning observable social exchange strategies in open societies. *2010 Second Brazilian Workshop on Social Simulation* (pp. 66-73). https://www.researchgate.net/publication/224259617

Dimuro, G. P., Costa, A. C. d. R., Gonçalves, L. V., & Hubner, A. (2008). Interval-valued Hidden Markov Models for recognizing personality traits in social exchanges in open multiagent systems. *Trends in Computational and Applied Mathematics*, *9*(1). https://tema.sbmac.org.br/tema/article/view/183

Dinh, D.-T., Fujinami, T., & Huynh, V.-N. (2019). Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. *International Symposium on Knowledge and Systems Sciences* (pp. 1-17). Springer. https://www.springerprofessional.de/en/estimating-the-optimal-number-of-clusters-in-categorical-data-cl/17338316

Do, A.-L., Rudolf, L., & Gross, T. (2010). Patterns of cooperation: Fairness and coordination in networks of interacting agents. *New Journal of Physics*, *12*(6), 063023.

Dovidio, J. F., Eller, A., & Hewstone, M. (2011). Improving intergroup relations through direct, extended and other forms of indirect contact. *Group Processes & Intergroup Relations*, *14*(2), 147-160.

Dovidio, J. F., & Gaertner, S. L. (1981). The effects of race, status, and ability on helping behavior. *Social Psychology Quarterly*, *44*(3), 192-203.

Dovidio, J. F., Gaertner, S. L., & Abad-Merino, S. (2017). Helping behaviour and subtle discrimination. In E. van Leeuwen & H. Zagefka (Eds.), *Intergroup helping* (pp. 3-22). Springer.

Dovidio, J. F., Gaertner, S. L., & Kawakami, K. (2003). Intergroup contact: The past, present, and the future. *Group Processes & Intergroup Relations*, *6*(1), 5-21.

Dovidio, J. F., Love, A., Schellhaas, F. M., & Hewstone, M. (2017). Reducing intergroup bias through intergroup contact: Twenty years of progress and future directions. *Group Processes & Intergroup Relations*, *20*(5), 606-620.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management*, *48*, 63-71.

Duffy, J. (2006). Agent-based models and human subject experiments. *Handbook of Computational Economics*, *2*, 949-1011.

Durrheim, K., Quayle, M., Tredoux, C. G., Titlestad, K., & Tooke, L. (2016). Investigating the evolution of ingroup favoritism using a minimal group interaction paradigm: The effects of inter-and intragroup interdependence. *PloS ONE*, *11*(11), e0165974.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*, 91-118.

Eclipse Deeplearning4j Development Team. (2016). *Deeplearning4j: Open-source distributed deep learning for the JVM*. Apache Software Foundation License 2.0. https://deeplearning4j.konduit.ai/

Edmonds, B. (2017). Different modelling purposes. In B. Edmonds & R. Meyer (Eds.), *Simulating social complexity* (pp. 39-58). Springer.

Edmonds, B., & Moss, S. (2004). From KISS to KIDS – an 'anti-simplistic'modelling approach. In P. Davidsson, B. Logan, & K. Takadama (Eds.), *Multi-Agent and Multi-Agent-Based Simulation*. MABS 2004. Lecture Notes in Computer Science, 3415. Springer. https://link.springer.com/chapter/10.1007/978-3-540-32243-6_11

Ekeh, P. P. (1974). *Social exchange theory: The two traditions*. Heinemann.

Ellemers, N. (2001). Legitimacy of intergroup relations. In J. T. Jost (Ed.), *The psychology of legitimacy: Emerging perspectives on ideology, justice, and intergroup relations* (pp. 205-222). Cambridge University Press.

Emlen, J. T. (1952). Flocking behavior in birds. *The Auk*, *69*(2), 160-170.

Enayat, T., Ardebili, M. M., Kivi, R. R., Amjadi, B., & Jamali, Y. (2020). A computational approach to Homans social exchange theory. *arXiv preprint arXiv:2007.14953*.

Fast, N. J., & Schroeder, J. (2020). Power and decision making: New directions for research in the age of artificial intelligence. *Current Opinion in Psychology*, *33*, 172-176.

FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *Neuroimage*, *105*, 347-356.

Fernández Domingos, E., Terrucha, I., Suchon, R., Grujić, J., Burguillo, J. C., Santos, F. C., & Lenaerts, T. (2022). Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma. *Scientific reports*, *12*(1), 1-12.

Fibbi, R., Midtbøen, A. H., & Simon, P. (2021). Theories of discrimination. In *Migration and Discrimination*, 21-41. IMISCOE Research Series. Springer.

Flache, A., & Macy, M. (2004). Social life from the bottom up: Agent modeling and the new sociology. *Social Dynamics: Interaction, Reflexivity and Emergence. Proceedings of the Agent 2004 Conference* (pp. 275-303). https://research.rug.nl/en/publications/social-life-from-the-bottom-up-agent-modeling-and-the-new-sociolo

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). The evolution of one-shot cooperation: An experiment. *Ethology and Sociobiology*, *14*(4), 247-256.

Fritzlen, K. A., Phillips, J. E., March, D. S., Grzanka, P. R., & Olson, M. A. (2019). I know (what) you are, but what am I? The effect of recategorization threat and perceived immutability on prejudice. *Personality and Social Psychology Bulletin*, *46*(1), 94-108.

Fu, F., Tarnita, C. E., Christakis, N. A., Wang, L., Rand, D. G., & Nowak, M. A. (2012). Evolution of in-group favoritism. *Scientific Reports*, *2*, Article 460.

Fulk, J., Schmitz, J., & Steinfield, C. W. (1990). A social influence model of technology use. In J. Fulk & C. Steinfield (Eds.), *Organizations and communication technology* (pp. 117-140. Sage.

Gan, G., Ma, C., & Wu, J. (2020). Data clustering: Theory, algorithms, and applications. *International Workshop on Agent Theories, Architectures, and Languages,* https://epubs.siam.org/doi/abs/10.1137/1.9780898718348

Georgeff, M., Pell, B., Pollack, M., Tambe, M., & Wooldridge, M. (1998). The belief-desire-intention model of agency. In J. P. Müller, A. S. Rao, & M. P. Singh (Eds.), *Intelligent Agents V: Agents Theories, Architectures, and Languages*. ATAL 1998. Lecture Notes in Computer Science, 1555. Springer. https://link.springer.com/chapter/10.1007/3-540-49057-4_1

Gibbs Jr, R. W., & Van Orden, G. C. (2001). Mental causation and psychological theory. *Human Development*, *44*(6). 368-374.

Gilbert, N. (2004). Agent-based social simulation: Dealing with complexity. *The Complex Systems Network of Excellence*, *9*(25), 1-14.

Gower, J. C. (1967). A comparison of some methods of cluster analysis. *Biometrics*, *23*(4), 623-637.

Greasley, A., & Owen, C. (2018). Modelling people’s behaviour using discrete-event simulation: A review. *International Journal of Operations & Production Management*, *38*(5), 1228-1244.

Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist*, *69*(7), 669-684.

Grignard, A., Taillandier, P., Gaudou, B., Vo, D. A., Huynh, N. Q., & Drogoul, A. (2013). GAMA 1.6: Advancing the art of complex agent-based modeling and simulation. In G. Boella, E. Elkind, B. T. R., Savarimuthu, F. Dignum, & M. K.Purvis (Eds.), *PRIMA 2013: Principles and Practice of Multi-Agent Systems*. PRIMA 2013. Lecture Notes in Computer Science, 8291. Springer. https://link.springer.com/chapter/10.1007/978-3-642-44927-7_9

Gunantara, N. (2018). A review of multi-objective optimization: Methods and its applications. *Cogent Engineering*, *5*(1), Article 1502242.

Günther, F., & Fritsch, S. (2010). Neuralnet: Training of neural networks. *The R Journal*, *2*(1), 30-38.

Haslam, S. A., & Reicher, S. (2006). Stressing the group: Social identity and the unfolding dynamics of responses to stress. *Journal of Applied Psychology*, *91*(5), 1037-1052.

Haslam, S. A., & Reicher, S. D. (2012). Contesting the “nature” of conformity: What Milgram and Zimbardo's studies really show. *PLoS Biology*, *10*(11), e1001426.

Hauser, O. P., Kraft-Todd, G. T., Rand, D. G., Nowak, M. A., & Norton, M. I. (2019). Invisible inequality leads to punishing the poor and rewarding the rich. *Behavioural Public Policy*, *5*(3), 333-353.

Helbing, D., & Balietti, S. (2011). How to do agent based simulations in the future: From modeling social mechanisms to emergent phenomena and interactive systems design. In H. Dirk & S. Balietti (Eds.), *Social self-organization* (pp. 25-70). Springer.

Hertel, G., & Kerr, N. L. (2001). Priming in-group favoritism: The impact of normative scripts in the minimal group paradigm. *Journal of Experimental Social Psychology*, *37*(4), 316-324.

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, *53*(1), 575-604.

Homans, G. C. (1961). *Social behavior: Its elementary forms* . Harcourt, Brace & World, New York.

Homans, G. C. (1974). *Social behavior: Its elementary forms* . Harcourt, Brace & World, New York.

Hopfield, J. J. (1988). Artificial neural networks. *IEEE Circuits and Devices Magazine*, *4*(5), 3-10.

Hula, A., Montague, P. R., & Dayan, P. (2015). Monte Carlo planning method estimates planning horizons during interactive social exchange. *PLoS Computational Biology*, *11*(6), e1004254.

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809-816.

Janssens, L., & Nuttin, J. R. (1976). Frequency perception of individual and group successes as a function of competition, coaction, and isolation. *Journal of Personality and Social Psychology*, *34*(5), 830-836.

Jenkins, A. C., Karashchuk, P., Zhu, L., & Hsu, M. (2018). Predicting human behavior toward members of different social groups. *Proceedings of the National Academy of Sciences*, *115*(39), 9696-9701. https://doi.org/10.1073/pnas.1719452115

Jetten, J., Spears, R., & Manstead, A. S. (1998). Defining dimensions of distinctiveness: Group variability makes a difference to differentiation. *Journal of Personality and Social Psychology*, *74*(6), 1481-1492.

Johnson, C. (2006). *Basic research skills in computing science*. Glasgow Interactive Systems Group (GIST), Department of Computer Science, Glasgow University, UK. http://www.dcs.gla.ac.uk/~johnson/teaching/research_skills/basics.html

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*(5), 865-889.

Jung, J., Hogg, M. A., & Choi, H. S. (2019). Recategorization and ingroup projection: Two processes of identity uncertainty reduction. *Journal of Theoretical Social Psychology*, *3*(2), 97-114.

Kalantari, S., Nazemi, E., & Masoumi, B. (2020). Emergence phenomena in self-organizing systems: A systematic literature review of concepts, researches, and future prospects. *Journal of Organizational Computing and Electronic Commerce*, *30*(3), 224-265.

Kavak, H., Padilla, J. J., Lynch, C. J., & Diallo, S. Y. (2018). *Big data, agents, and machine learning: Towards a data-driven agent-based modeling approach*. SpringSim.

Kelley, H. H., & Thibaut, J. W. (1978). *Interpersonal relations: A theory of interdependence*. Wiley.

Kisfalusi, D., Janky, B., & Takács, K. (2019). Double standards or social identity? The role of gender and ethnicity in ability perceptions in the classroom. *Journal of Early Adolescence*, *39*(5), 745-780.

Kollock, P. (1999). The economies of online cooperation. gifts and public goods in cyberspace. In P. Kollock & M. A. Smith (Eds.), *Communities in cyberspace* (pp. 220-246). Routledge.

Konovalov, A., & Ruff, C. C. (2021). Enhancing models of social and strategic decision making with process tracing and neural data. *WIREs Cognitive Science*, *13*, e1559.

Kozlowski, S. W., Chao, G. T., Grand, J. A., Braun, M. T., & Kuljanin, G. (2013). Advancing multilevel research design: Capturing the dynamics of emergence. *Organizational Research Methods*, *16*(4), 581-615.

Krafft, P. M., Macy, M., & Pentland, A. S. (2017). Bots as virtual confederates: Design and ethics. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* https://hdl.handle.net/1721.1/137796.2

Kranton, R., Pease, M., Sanders, S., & Huettel, S. (2016). Groupy and non-groupy behavior: Deconstructing bias in social preferences. *PNAS*, *117*(35), 21185-21193.

Kranton, R., Pease, M., Sanders, S., & Huettel, S. (2017). *Deconstructing bias: Individual groupiness and income allocation*. Duke University. https://sites.duke.edu/rachelkranton/files/2016/09/Deconstructing-Bias-paper-august-14-2017-final.pdf

Lamperti, F., Roventini, A., & Sani, A. (2018). Agent-based model calibration using machine learning surrogates. *Journal of Economic Dynamics and Control*, *90*, 366-389.

Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *151*, 61-72.

Lawler, E. J. (2001). An affect theory of social exchange. *American Journal of Sociology*, *107*(2), 321-352.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Lee, K., Rucker, M., Scherer, W. T., Beling, P. A., Gerber, M. S., & Kang, H. (2017). Agent-based model construction using inverse reinforcement learning. In V. Chan (Ed.), *2017 Winter Simulation Conference, WSC 2017* (pp. 1264-1275). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/WSC.2017.8247872

Lee, S., Adair, W. L., Mannix, E. A., & Kim, J. (2012). The relational versus collective "We" and intergroup allocation: The role of nested group categorization. *Journal of Experimental Social Psychology*, *48*(5), 1132-1138.

Lee, S. H., Gokalp, O. N., & Kim, J. (2021). Firm-government relationships: A social exchange view of corporate tax compliance. *Global Strategy Journal*, *11*(2), 185-209.

Li, Z., Sim, C. H., & Low, M. Y. H. (2006). A survey of emergent behavior and its impacts in agent-based systems. *IEEE International Conference on Industrial Informatics* (pp. 1295-1300). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.217.3183&rep=rep1&type=pdf

List, C., Elsholtz, C., & Seeley, T. D. (2008). Independence and interdependence in collective decision making: An agent-based model of nest-site choice by honeybee swarms. *Philosophical Transactions of The Royal Society B: Biological Sciences*, *364*(1518), 755-762.

Liu, X., & Datta, A. (2012). Modeling context-aware dynamic trust using Hidden Markov Model. *Proceedings of theTwenty-Sixth AAAI Conference on Artificial Intelligence* (pp. 1938-1944). https://ojs.aaai.org/index.php/AAAI/article/view/8395

MacCrimmon, K. R., & Messick, D. M. (1976). A framework for social motives. *Behavioral Science*, *21*(2), 86-100.

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. Interaction Studies, 7(3), 297-337.

Macy, M. W., & Flache, A. (1995). Beyond rationality in models of choice. *Annual Review of Sociology*, *21*(1), 73-91.

Mann, A. (2016). Core concept: Computational social science. *Proceedings of the National Academy of Sciences*, *113*(3), 468-470.

March, C. (2021). Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology*, *87*, Article 102426.

Marmarosh, C. L., & Sproul, A. (2021). Group cohesion: Empirical evidence from group psychotherapy for those studying other areas of group work. In C. D. Parks & G. A. Tasca (Eds.), *The psychology of groups: The intersection of social psychology and psychotherapy research* (pp. 169-189). American Psychological Association.

McDonald, R. I., & Crandall, C. S. (2015). Social norms and social influence. *Current Opinion in Behavioral Sciences*, *3*, 147-151.

Mehlig, B. (2019). Artificial neural networks. *arXiv e-prints*, arXiv: 1901.05639.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT Press.

Molm, L. D., Collett, J. L., & Schaefer, D. R. (2007). Building solidarity through generalized exchange: A theory of reciprocity. *American Journal of Sociology*, *113*(1), 205-242.

Molm, L. D., Melamed, D., & Whitham, M. M. (2013). Behavioral consequences of embeddedness: Effects of the underlying forms of exchange. *Social Psychology Quarterly*, *76*(1), 73-97.

Montoya, R. M., & Pittinsky, T. L. (2013). Individual variability in adherence to the norm of group interest predicts outgroup bias. *Group Processes & Intergroup Relations*, *16*(2), 173-191.

Mor, B., Garhwal, S., & Kumar, A. (2021). A systematic review of Hidden Markov Models and their applications. *Archives of Computational Methods in Engineering*, *28*(3), 1429-1448.

Morgan, J. H., Lebiere, C., Moody, J., & Orr, M. G. (2021). Trusty ally or faithless snake: Modeling the role of human memory and expectations in social exchange. *Proceedings of the 14th International Conference, SBP-BRiMS 2021* (pp. 268-278). https://doi.org/10.1007/978-3-030-80387-2_26

Nguyen, N., & Nguyen, D. (2021). Global stock selection with hidden Markov model. *Risks*, *9*(1), 9. https://doi.org/10.3390/risks9010009

Nilsson, N. J. (1977). *A production system for automatic deduction*. Department of Computer Science, Stanford University.

Nowak, M. A., & Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, *194*(4), 561-574.

Ntwiga, D. B., & Ogutu, C. (2018). Interaction dynamics in a social network using Hidden Markov Model. *Social Networking*, *7*(3), 147-155.

Olekalns, M., & Smith, P. L. (2021). Decision frames and the social utility of negotiation outcomes. *Current Psychology*, Article 143. https://doi.org/10.1007/s12144-021-02248-8

Opp, K. D. (1982). The evolutionary emergence of norms. *British Journal of Social Psychology*, *21*(2), 139-149.

Osoba, O., & Davis, P. K. (2019). An artificial intelligence/machine learning perspective on social simulation: new data and new challenges. In P. K. Davis, A. O'Mahony, & J. Pfautz (Eds.), *Social-behavioral modeling for complex systems* (pp. 443-476). Wiley.

Ostrom, T. M. (1988). Computer simulation: The third symbol system. *Journal of Experimental Social Psychology*, *24*(5), 381-392.

Parsell, C., & Clarke, A. (2020). Charity and shame: Towards reciprocity. *Social Problems*, *69*(2), 436-452.

Pavard, B., & Dugdale, J. (2002). An introduction to complexity in social science. *Tutorial on Complexity in Social Science, COSI project.* http://www.irit. fr/COSI/training/complexitytutorial/complexity-tutorial. htm

Perelman, C. (1979). The new rhetoric: A theory of practical reasoning. In C. Perelman (Ed.), *The new rhetoric and the humanities* (pp. 1-42). Springer.

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology*, *49*(1), 65-85.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751-783.

Phelps, S. (2013). Emergence of social networks via direct and indirect reciprocity. *Autonomous Agents and Multi-agent Systems*, *27*(3), 355-374.

Phung, T., Winikoff, M., & Padgham, L. (2005). Learning within the bdi framework: An empirical analysis. In R. Khosla, R. J. Howlett, & L. C. Jain (Eds.), *Knowledge-based intelligent information and engineering systems*. Springer. https://doi.org/10.1007/11553939_41

Piaget, J. (1967/1995). Sociological studies, L. Smith (Ed.). Lawrence Erlbaum Associates.

Quené, H., & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, *43*(1-2), 103-121.

Rabbie, J. M., Benoist, F., Oosterbaan, H., & Visser, L. (1974). Differential power and effects of expected competitive and cooperative intergroup interaction on intragroup and outgroup attitudes. *Journal of Personality and Social Psychology*, *30*(1), 46-56.

Rabbie, J. M., & Wilkens, G. (1971). Intergroup competition and its effect on intragroup and intergroup relations. *European Journal of Social Psychology*, *1*(2), 215-234.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.

Railsback, S. F., & Grimm, V. (2012). *Agent-based and individual-based modeling: A practical introduction*. Princeton University Press.

Rapoport, A., & Chammah, A. M. (1970). *Prisoner's dilemma: A study in conflict and cooperation*. University of Michigan Press.

Ray, D., King-Casas, B., Montague, P. R., & Dayan, P. (2009). Bayesian model of behaviour in economic games. *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 8-11, 2008. https://www.semanticscholar.org/paper/Bayesian-Model-of-Behaviour-in-Economic-Games-Ray-King-Casas/e2302a758da280bd72999c096c1c5df104dac9b5

Rosenfeld, A., Zuckerman, I., Azaria, A., & Kraus, S. (2012). Combining psychological models with machine learning to better predict people's decisions. *Synthese*, *189*(1), 81-93.

Santucci, V. G., Cartoni, E., da Silva, B. C., & Baldassarre, G. (2019). Autonomous open-ended learning of interdependent tasks. *arXiv preprint arXiv:1905.02690*.

Sawyer, R. K. (2002). Emergence in psychology: Lessons from the history of non-reductionist science. *Human Development*, *45*(1), 2-28.

Saygin, A. P., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and Machines*, *10*(4), 463-518.

Schellhaas, F. M., & Dovidio, J. F. (2016). Improving intergroup relations. *Current Opinion in Psychology*, *11*, 10-14.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, *1*(2), 143-186.

Shreffler, J., & Huecker, M. R. (2022). Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios. In *StatPearls [Internet]*. StatPearls Publishing.

Simpson, B., Harrell, A., Melamed, D., Heiserman, N., & Negraia, D. V. (2018). The roots of reciprocity: Gratitude and reputation in generalized exchange systems. *American Sociological Review*, *83*(1), 88-110.

Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, *11*(1), 87-104.

Sohn, D., & Leckenby, J. D. (2007). A structural solution to communication dilemmas in a virtual community. *Journal of Communication*, *57*(3), 435-449.

Spicer, J., & Sanborn, A. N. (2019). What does the mind learn? A comparison of human and machine learning representations. *Current Opinion in Neurobiology*, *55*, 97-102.

Spielman, D. A. (2000). Young children, minimal groups, and dichotomous categorization. *Personality and Social Psychology Bulletin*, *26*(11), 1433-1441.

Stangor, C., Jhangiani, R., & Tarry, H. (2022). Ingroup favoritism and prejudice. In R. Jhangiani & H. Tarry (Eds.), *Principles of social psychology -1st international H5P edition*. https://opentextbc.ca/socialpsychology/

Stürmer, S., & Siem, B. (2017). A group-level theory of helping and altruism within and across group boundaries. In E. van Leeuwen & H. Zagefka (Eds.), *Intergroup helping* (pp. 103-127). Springer.

Suzuki, S., & O'Doherty, J. P. (2020). Breaking human social decision making into multiple components and then putting them together again. *Cortex*, *127*, 221-230.

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology*, *33*, 1-39.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149-178.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *Psychology of intergroup relations* (pp. 56-65). Nelson-Hall.

Tajfel, H., & Turner, J. C. (1985). The social identity theory of intergroup behavior. In *Psychology of intergroup relations* (2nd ed., pp. 7-24). Nelson-Hall Publishers.

Takahashi, N. (2000). The emergence of generalized exchange. *American Journal of Sociology*, *105*(4), 1105-1134.

Taylor, S. J., Khan, A., Morse, K. L., Tolk, A., Yilmaz, L., Zander, J., & Mosterman, P. J. (2015). Grand challenges for modeling and simulation: Simulation everywhere – from cyberinfrastructure to clouds to citizens. *Simulation*, *91*(7), 648-665.

Team R Core. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. https://www.R-project.org

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168-192.

Titlestad, K., Snijders, T., Durrheim, K., Quayle, M., & Postmes, T. (2019). The dynamic emergence of cooperative norms in a social dilemma. *Journal of Experimental Social Psychology*, *84*, Article 103799.

Turner, J. C., & Reynolds, K. J. (2001). The social identity perspective in intergroup relations: Theories, themes, and controversies. In R. Brown & S. L. Gaertner (Eds.), *Blackwell handbook of social psychology* (pp. 133-152). Blackwell: Intergroup Processes.

Turner, J. C., & Tajfel, H. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relation* (pp. 7-24). Hall Publishers.

Vallacher, R. R., Van Geert, P., & Nowak, A. (2015). The intrinsic dynamics of psychological process. *Current Directions in Psychological Science*, *24*(1), 58-64.

van der Hoog, S. (2017). Deep learning in (and of) agent-based models: A prospectus. *arXiv preprint arXiv:1706.06302*.

Veldhuis, T. M., Gordijn, E. H., Veenstra, R., & Lindenberg, S. (2014). Vicarious group-based rejection: creating a potentially dangerous mix of humiliation, powerlessness, and anger. *PloS ONE*, *9*(4), e95421.

Walker, M. (2019, April). Implicit bias: Root cause of discrimination against women in construction. In P. Paolini (Ed.), *International Conference on Gender Research* (pp. 646-XXIV). Academic Conferences International Limited.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Paper 601. https://doi.org/10.1145/3290605.3300831

Warner, B., & Misra, M. (1996). Understanding neural networks as statistical tools. *The American Statistician*, *50*(4), 284-293.

Willer, R., Flynn, F. J., & Zak, S. (2012). Structure, identity, and solidarity: A comparative field study of generalized and direct exchange. *Administrative Science Quarterly*, *57*(1), 119-155.

Wilson, W., & Miller, N. (1961). Shifts in evaluations of participants following intergroup competition. *Journal of Abnormal and Social Psychology*, *63*(2), 428-431.

Wispe, L. G., & Freshley, H. B. (1971). Race, sex, and sympathetic helping behavior: The broken bag caper. *Journal of Personality and Social Psychology*, *17*(1), 59-65.

Wright, S. C., Aron, A., McLaughlin-Volpe, T., & Ropp, S. A. (1997). The extended contact effect: Knowledge of cross-group friendships and prejudice. *Journal of Personality and Social Psychology*, *73*(1), 73-90.

Xiang, T., Ray, D., Lohrenz, T., Dayan, P., & Montague, P. R. (2012). Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Computational Biology*, *8*(12), e1002841.

Yamagishi, T., Jin, N., & Kiyonari, T. (1999). Bounded generalized reciprocity: Ingroup boasting and ingroup favoritism. *Advances in Group Processes*, *16*(1), 161-197.

Yamagishi, T., & Mifune, N. (2009). Social exchange and solidarity: In-group love or out-group hate? *Evolution and Human Behavior*, *30*(4), 229-237.

Yiu, N. C. (2021). Decentralizing supply chain anti-counterfeiting systems using blockchain technology. *arXiv preprint arXiv:2102.01456*.

Yoshikawa, K., Wu, C.-H., & Lee, H.-J. (2018). Generalized social exchange and its relevance to new era workplace relationships. *Industrial and Organizational Psychology*, *11*(3), 486-492.

Zaki, J. F., Ali-Eldin, A., Hussein, S. E., Saraya, S. F., & Areed, F. F. (2020). Traffic congestion prediction based on Hidden Markov Models and contrast measure. *Ain Shams Engineering Journal*, *11*(3), 535-551.

Zhang, X., Li, Y., Wang, S., Fang, B., & Philip, S. Y. (2019). Enhancing stock market prediction with extended coupled hidden Markov model over multi-sourced data. *Knowledge and Information Systems*, *61*(2), 1071-1090.

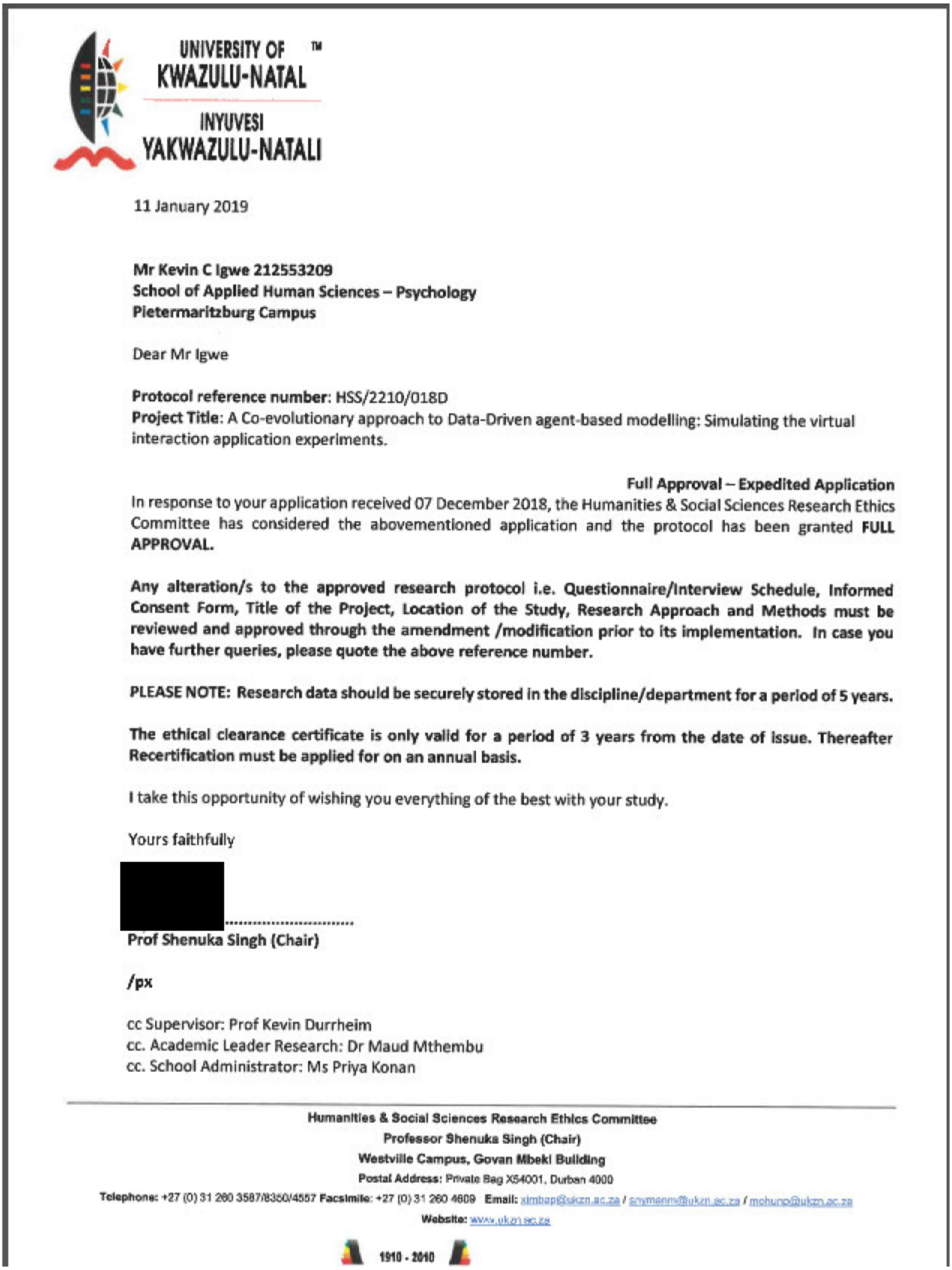
Zhang, Y., Grignard, A., Lyons, K., Aubuchon, A., & Larson, K. (2018). Real-time machine learning prediction of an agent-based model for urban decision-making. Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, 2018, July 10-15, 2018, Stockholm, Sweden. https://www.media.mit.edu/publications/real-time-machine-learning-prediction-of-an-agent-based-model-for-urban-decision-making/

Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D. C., & Socher, R. (2020). The AI economist: Improving equality and productivity with ai-driven tax policies. *arXiv preprint arXiv:2004.13332*.

Zuo, Y., Chen, B., & Zhao, Y. (2018). The destructive effect of ingroup competition on ingroup favoritism. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02207

.

Appendix 1: Letter of Ethical Approval



UNIVERSITY OF KWAZULU-NATAL™
INYUVESI YAKWAZULU-NATALI

11 January 2019

Mr Kevin C Igwe 212553209
School of Applied Human Sciences – Psychology
Pietermaritzburg Campus

Dear Mr Igwe

Protocol reference number: HSS/2210/018D
Project Title: A Co-evolutionary approach to Data-Driven agent-based modelling: Simulating the virtual interaction application experiments.

Full Approval – Expedited Application

In response to your application received 07 December 2018, the Humanities & Social Sciences Research Ethics Committee has considered the abovementioned application and the protocol has been granted **FULL APPROVAL.**

Any alteration/s to the approved research protocol i.e. Questionnaire/Interview Schedule, Informed Consent Form, Title of the Project, Location of the Study, Research Approach and Methods must be reviewed and approved through the amendment /modification prior to its implementation. In case you have further queries, please quote the above reference number.

PLEASE NOTE: Research data should be securely stored in the discipline/department for a period of 5 years.

The ethical clearance certificate is only valid for a period of 3 years from the date of issue. Thereafter Recertification must be applied for on an annual basis.

I take this opportunity of wishing you everything of the best with your study.

Yours faithfully

[redacted]............................
Prof Shenuka Singh (Chair)

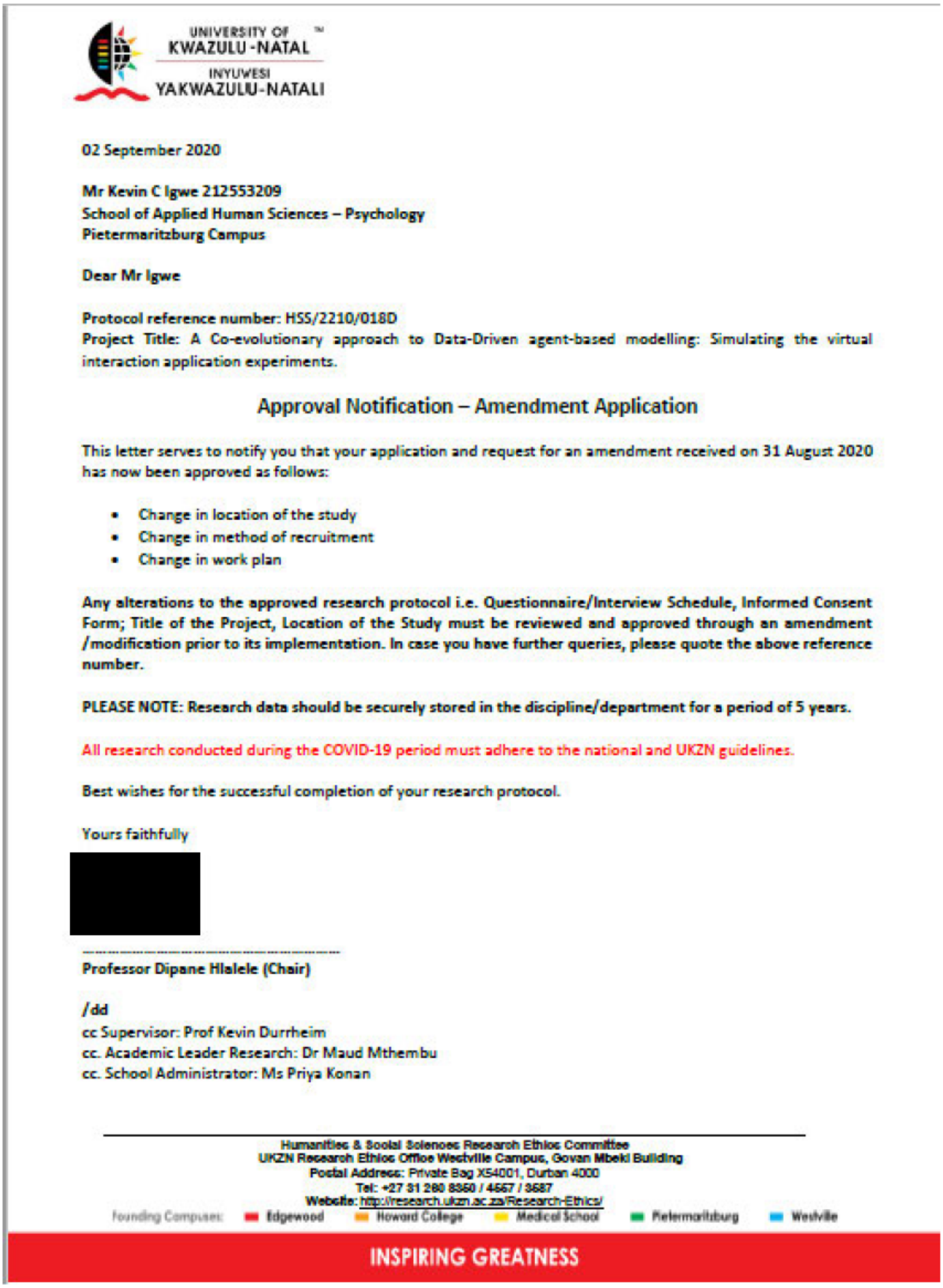
/px

cc Supervisor: Prof Kevin Durrheim
cc. Academic Leader Research: Dr Maud Mthembu
cc. School Administrator: Ms Priya Konan

Humanities & Social Sciences Research Ethics Committee
Professor Shenuka Singh (Chair)
Westville Campus, Govan Mbeki Building
Postal Address: Private Bag X54001, Durban 4000
Telephone: +27 (0) 31 260 3587/8350/4557 **Facsimile:** +27 (0) 31 260 4609 **Email:** ximbap@ukzn.ac.za / snymanm@ukzn.ac.za / mohunp@ukzn.ac.za
Website: www.ukzn.ac.za

1910 - 2010

Appendix 1b: Letter of Ethical Approval – Amended



UNIVERSITY OF KWAZULU-NATAL™
INYUVESI YAKWAZULU-NATALI

02 September 2020

Mr Kevin C Igwe 212553209
School of Applied Human Sciences – Psychology
Pietermaritzburg Campus

Dear Mr Igwe

Protocol reference number: HSS/2210/018D
Project Title: A Co-evolutionary approach to Data-Driven agent-based modelling: Simulating the virtual interaction application experiments.

Approval Notification – Amendment Application

This letter serves to notify you that your application and request for an amendment received on 31 August 2020 has now been approved as follows:

- Change in location of the study
- Change in method of recruitment
- Change in work plan

Any alterations to the approved research protocol i.e. Questionnaire/Interview Schedule, Informed Consent Form; Title of the Project, Location of the Study must be reviewed and approved through an amendment /modification prior to its implementation. In case you have further queries, please quote the above reference number.

PLEASE NOTE: Research data should be securely stored in the discipline/department for a period of 5 years.

All research conducted during the COVID-19 period must adhere to the national and UKZN guidelines.

Best wishes for the successful completion of your research protocol.

Yours faithfully

[redacted]

……………………………………………
Professor Dipane Hlalele (Chair)

/dd
cc Supervisor: Prof Kevin Durrheim
cc. Academic Leader Research: Dr Maud Mthembu
cc. School Administrator: Ms Priya Konan

Humanities & Social Sciences Research Ethics Committee
UKZN Research Ethics Office Westville Campus, Govan Mbeki Building
Postal Address: Private Bag X54001, Durban 4000
Tel: +27 31 260 8350 / 4557 / 3587
Website: http://research.ukzn.ac.za/Research-Ethics/

Founding Campuses: 🟥 Edgewood 🟧 Howard College 🟨 Medical School 🟩 Pietermaritzburg 🟦 Westville

INSPIRING GREATNESS

Appendix 2: Information Sheet

Visual Interaction Application - 2019

Dear Participant,

This is a research project on intergroup behaviour. It has been approved by the UKZN Human Social Science Research Ethics Committee and the protocol reference number is HSS/2210/018D.

Brief outline of the study: This research study aims to explore behaviour in a social setting. The study is electronically based game, played by 8 players, by giving of tokens.

What you will be required to do: The study will take place in the Psyc Lab. You will be required to play a game and answer some questions on questionnaires. This will take about 30 minutes to an hour of your time.

Voluntary participation: Your participation is voluntary and you are not being forced to take part in this study. The choice of whether or not to participate is yours alone, and there will be no consequences if you choose to not take part. You may withdraw from the research at any time by telling me that you do not want to continue. There will be no penalties for doing so, however you will not receive your incentive money unless you complete the study.

Anonymity: Although we will ask you to register as a research participant, your responses will not be linked with your name or any other information by which you can be identified. Furthermore, will we ask you to take a webcam photo at the start of the game depending on the manipulation; these photos are in no way linked to your responses and will not be used for any purpose other than game manipulation. In other words, you will remain entirely anonymous and your participation will remain confidential. There are no limits to confidentiality.

Research incentive: Participants will be given an average of R30 cash after completing the study; you will receive an incentive that corresponds with how well you do in the game.

Who to contact if you have been harmed or have any concerns: Although this research involves very little risk, if you have any questions or complaints about aspects of the research or feel that you have been harmed in any way by participating in this study, please contact:

- Human Social Science Research Ethics Committee:

- Ms. PhumeXimba (ximbap@ukzn.ac.za/ 031 260 3587)
- Project Leaders: School of Applied Human Sciences, University of KwaZulu-Natal: Professor Kevin Durrheim (Durrheim@ukzn.ac.za)

  Mr Kevin Igwe (igwekevin@gmail.com)

  Mrs. Tsitsi Chirove (chirovets@gmail.com)

Appendix 3: Consent Form

Consent form

I hereby agree to participate in research on social interaction. I am aware of what is required of me, and I understand that:

- I am participating freely and without coercion.
- This is a research project whose purpose is not necessarily to benefit me personally.
- I will remain anonymous and my participation in the study will remain confidential.
- I have a right to withdraw from the study at any time, without penalty.
- I agree to the results of my participation being used for research and teaching purposes and for presentation in reports and at conferences. My name will not appear in any of these documents.
- I agree to my photo being taken via webcam for game manipulation purposes.
- I agree/disagree to the discussion at the end of the game being recorded for research purposes.

Signature of participant:________________________________Date:_________________