Modelling Shared Identity and Reputation in Cooperation Systems

A thesis submitted for the partial fulfilment

for the degree of Doctor of Philosophy by

Wafi Amjed H. Bedewi

School of Computer Science and Informatics

Cardiff University

April 2021

Abstract

Cooperation is the process of working together for mutual benefit. Indirect reciprocity is an important form of cooperation because it assumes that a donation to an agent does not guarantee reciprocation. Therefore, understanding how cooperation is incentivised and sustained is of widespread interest. Reputation is known as a key mechanism to support indirect reciprocity because it is a currency through which future donations can be secured based on past behaviour. Conventional models of indirect reciprocity assume that agents have a simple identity that is uniquely defined and not shared with others. This results in a unique reputation for each agent. We generalise this assumption by allowing agents to share elements of their identity with others. This involves composing identity through traits, which can be used to represent group membership. Traits can be shared between agents and we assume that traits carry reputation in their own right, that an agent can inherit.

Our investigation of this new framework provides an insight into the effects of sharing identity on cooperation in a number of different ways. Through a breadth of simulation, we identify the extent to which agents can have an element of common identity before cooperation becomes impeded. We also discover a relationship between reputation-based cooperation and cooperation through the evolution of set-based membership, which are previously unrelated alternative perspectives on indirect reciprocity. Finally, we explore the effects of blending personal and group reputations as seen in psychological theories of identity fusion. This allows us to determine the effects of identity-driven agent motivation compared to traditional economic motivation and rational economic decision-making. These findings give new perspectives into previous studies related to identity, such as stereotyping, group identity, whitewashing, identity fusion and intrinsic motivation.

Contents

List of Figures

List of Tables

List of Algorithms

Dedication

To my parents, for their support and patience.

Acknowledgements

Alhamdulillah. Abu Huraira reported: The Prophet, peace and blessings be upon him, said, “Whoever does not thank people has not thanked God.” On this page, I would like to acknowledge and thank everyone who has helped me along my PhD journey.

I want to express my sincere gratitude and thanks to my supervisor, Professor Roger Whitaker, for going out of his way to support me. I would also like to thank my second supervisor, Professor Stuart Allen, for his encouragement and support from the beginning of my PhD. I am specially grateful to Dr Walter Colombo, who has also gone out of his way to support me from day one until the day I submitted. I do not believe I have enough words to express my appreciation for all the support that I have received from my supervisors throughout my PhD.

I would also like to thank my parents, Amjed and Ebtisam, who have provided everything for me to complete my PhD journey. There aren’t enough words that I can say that would express my thanks and gratitude to you. I would also like to thank my siblings Shaima, Sarah and Hussein and my nieces Nora, Salma and Lamees. Special thanks to my wider family, who have always checked in on me and encouraged me.

I would like to express my gratitude and thanks to my Cardiff-Bristol-France family and friends that I made here, who were always encouraging and welcoming. Thank you for forcing me to take breaks and for all the adventures that we have shared throughout these years. I would also like to thank all my other friends who have supported me, checked in on me from time to time, invited me to their weddings and shared part of their

life journey with me. Thank you for all the texts, postcards and letters of encouragement. Thank you for all the invisible soldiers I may have missed here but have been a part of my PhD journey.

Finally, I like to express my appreciation to my colleagues at King Abdulaziz University, who have supported me throughout this journey and helped me achieve this milestone. With special thanks to the Supercomputing Wales project for supporting my research.

To everyone: Thank you for being patient. Thank you for your support. Thank you for your guidance. Thank you for showing me the way. Thank you for encouraging me. Thank you for pushing me. Thank you for not letting me give up. Thank you for believing in me.

Chapter 1

Introduction

1.1 Introduction

This thesis focuses on the problem of cooperation, specifically the act of supporting a third party while incurring a cost to oneself. Cooperation is a universal phenomenon that has widespread relevance to decentralised systems, not least the human population. When cooperation flourishes, a community-wide benefit accrues, where all participants gain an “insurance” provided by the possibility of aid from others. This is a form of collective intelligence, where individuals come together to act for a social good that is shared across the population. However cooperation also leaves participants potentially exposed to risk and exploitation, where a third party can gain advantage by taking help while not donating anything to others. For this reason, understanding how cooperation is incentivised and sustained is of widespread interest. It has received attention from multiple disciplines including theoretical biology, mathematics (game theory), computer science, and morality.

In the literature, researchers have identified two types of reciprocity. The term ”direct reciprocity” refers to an exchange that takes place between two individuals. Indirect reciprocity, which involves a third person, is more complicated (see Section 2.5). We focus on the problem of downstream indirect reciprocity, where individuals may incur a small cost in order to provide a benefit that is of greater value to the recipient, and this occurs without the guarantee of future reciprocation from a recipient. This form of cooperation is frequently seen in the human population, being influenced by potentially diverse and complex mechanisms that have received considerable attention in the literature. Most notably, the importance of reputation has been established, acting as a currency and signalling system through which future donations can be secured based on past behaviour. Three-way junctions, where the right of way may not be clear, are an example scenario that helps illustrate how downstream indirect reciprocity works. If a driver allows another to turn, they are providing a benefit at a cost to themself, and an indirect reciprocal behaviour may be provided through an observer driver who may later reward this cost.

Through multi-agent simulation, this thesis concerns generalising how individuals express themselves proposing models explored through identity. In studies of downstream indirect reciprocity, it is a widespread convention that agents identify with a single reputation that is not shared with any other entity. In this work we generalise this issue and develop a flexible framework. Using inspiration from the psychological literature, we introduce a new model to represent an agent's identity (i.e., presentation to others). This involves the use of *traits*: a trait represents a feature that an agent identifies with. Each agent may align with multiple traits, and a trait may be shared by multiple agents. This provides a flexible representation through which agents can share their components of identity, in particular allowing group affiliation to be modelled. We assume that traits each hold a reputation, which represents a quality measure aligned to that trait. The reputation of an agent is then taken as a function of the reputation of traits that the individual agent possesses. As a consequence, the actions of an agent can affect the reputation of other agents that have a subset of traits in common.

Through this framework, we develop and investigate how cooperation is affected by agent identity, in particular the sharing of common traits. The approach also allows us to explore some psychologically inspired issues aligned with identity, particularly the effects of stereotyping, fused group identities and identity-driven behaviours of agents. In particular, the main contributions of this thesis are:

- a new generalised framework for agent identity that is based on combining traits that can be shared with others, such as group membership;

- through a breadth of simulation we identify the extent to which agents can have an element of common identity before cooperation becomes impeded, and relate this to the psychological concept of stereotyping;

- establishing the relationship between reputation-based cooperation and cooperation through the evolution of set membership, where cooperation evolves through the sets to which individuals belong as oppose to their reputation, which are previously disconnected explanations for the evolution of cooperation;

- exploring the effects of blending personal and group reputations as seen in identity fusion theory, and determining the effects of identity driven (as opposed to economic) motivations.

1.2 Contributions and Thesis Structure

The remaining Chapters are organised as follows:

- **Chapter 2** surveys the work related to (i) identity with a focus on social identity theory and identity fusion; (ii) cooperation with a focus on indirect reciprocity and reputation; (iii) identity within cooperation. The chapter concludes with a collection of research questions that are proposed from the gaps and limitations that have been exposed in the chapter.

- **Chapter 3** proposes a flexible approach to identity based on existing models of cooperation (indirect reciprocity). The extension focuses on shared identities through sharing of reputations. The chapter explores the current model's limitation by proposing and implementing a generalisation that allows shared identities to be examined through the use of traits.

- **Chapter 4** investigates different structures for sharing reputation through traits and their degree of impact on cooperation. Additionally, initial analysis on the generalisation, introduced in chapter 3, is carried out in this chapter.
- **Chapter 5** continues the investigations from the previous chapter by allowing agents to evolve their identity. This shifts the focus from exploring identity sharing mechanisms to how agents adopt different identities, showing the impact of that on cooperation.
- **Chapter 6** introduces a model based on the concepts related to identity fusion. It further builds on previous chapters by introducing further behavioural elements of fused agents.
- **Chapter 7** disrupts previous chapters' use of the economic convention of valuing success based on an individual's resources (payoff) to valuing their social identity. This represents an intrinsic motivation that humans often exhibit towards their social groups.
- **Chapter 8** reflects on the contributions of the thesis and provides a discussion on areas of future work.

1.3 List of Publications

The thesis includes work introduced in the following peer-reviewed publications:

1. Bedewi, W., Whitaker, R.M., Colombo, G.B., Allen, S.M. and Dunham, Y., 2019, September. Modelling stereotyping in cooperation systems. In International Conference on Computational Collective Intelligence (pp. 118-129). Springer, Cham.
2. Bedewi, W., Whitaker, R.M., Colombo, G.B., Allen, S.M. and Dunham, Y., 2020. The implications of shared identity on indirect reciprocity. Journal of Information and Telecommunication, 4(4), pp.405-424.

Chapter 2

Background

2.1 Introduction

There exists substantial research on the subject of *identity* in many different fields. Identity is defined in the Oxford English Dictionary as *"who or what something is"* [50]. The universal nature of this definition means that it is relevant to many different research areas, including philosophy [80, 144, 145], sociology [32, 82, 167], psychology [93, 166], theoretical biology [24, 111] and computer science [24, 41, 87].

The concept of identity in psychology is particularly relevant to this thesis because work in this field addresses the relationship between individual identity and identity related to collections or groups of individuals. This is relevant to complex systems and computer science because sharing of identity has high relevance to the increasingly digital world and fraudulent behaviour where identity is misleading others. For example, in complex systems, the use of shared group identity can give cover for shirkers or free-riders who exploit reputations shared with others [143], and gives an opportunity for white-washing, where agents change their identity, to take place [62].

In this chapter, we introduce groups and explain why they are important and how they can naturally form (Section 2.2). In human systems, groups can give a basis for an individual's identity. In Section 2.3, we consider social identity from a psychological perspective, particularly explaining different theories as to how identity is sustained and derived when individual agents are involved in groups. This has important relevance to cooperation scenarios because when agents interact they may do so while recognising the groups to which they belong. In Section 2.4, we introduce the concept of *cooperation* and the long-standing literature that supports this area. Cooperation is fundamental for distributed systems because it focuses on mechanisms and incentives that enable potentionally selfish entities to work together. Based on incentivising survival, much of this literature has origins in theoretical biology, while the formulation of cooperative dilemmas can be captured through economics and simple game theory.

We explain why cooperation is a fundamental concept and pay particular attention to *indirect reciprocity* (Section 2.5). This generalised form of cooperation is a characteristic of human cooperation and represents helping others to provide a wider community benefit. However, indirect reciprocity quickly becomes complex because it introduces the scenario of donating precious resources to others without certainty of a future pay-back. Therefore *who an agent is* plays an important role in decision making. This concept requires a reputation system (Section 2.5.1) to give a basis to differentiate between agents when making donations and many insights exist in this area. At this point the concept of identity is important, because who an entity is and how their reputation is shared, not just their status in terms of a reputation, may influence their donation behaviour. We explain this in the context of existing literature.

In conclusion, we make the observation that there is surprisingly limited previous work explicitly addressing identity in this context, despite identity and reputation being strongly intertwined.

2.2 Groups

Groups are a fundamental part of many human and biological systems as they play an important role in promoting survival prospects for individuals. From helping one another, members of a group can access benefits that are much bigger than the costs of contributing to that system, leading to mechanisms that allow individuals to work together and cooperate so that a large shared benefit accrues (see Section 2.4).

Consistent with this, there are numerous ways in which group formation can take place in human and biological systems. Groups naturally form around genetically related individuals (e.g., families) but they are often formed around traits that unrelated individuals have in common. This represents *homophily* [120], where the thing in common makes it easier for a bond to be established and a relationship to be sustained due to mutual respect. There are many examples in the literature of groups of unrelated individuals being formed on this basis [120], and these include both mutable and immutable traits. Examples of studies include friendships formed around personality traits which have been noted in humans and animals alike [42, 117], and friendships formed around a similar genotype [21, 70, 72].

The important point from this thesis's perspective is that *individuals derive some of their identity from being in a group*. This has been a long-standing area of study in terms of human behaviour (see Section 2.3) and it can be seen in terms of everyday reputations, where for example an alumnus's reputation is associated to their alma mater's. This occurs because humans are adept at making cognitive short cuts in judgement and decision making through heuristics [77, 78], which is efficient but can often lead to misjudgement in the form of stereotyping [22, 35, 100] (See Chapter 4). Conversely, the sharing of identity provides opportunities for shirkers to fraudulently exploit the reputation of a group without making contributions. In the following section, we outline the main theories behind an individual's identity, and how an individual may derive this from a group.

2.3 Individual and Group Identities

Identity is one of the cornerstones of social psychology. However, only in the last fifty years has the theory been developed in a substantive way. Initially, in psychology, identity focused on one's self, being a qualitative concept representing a person's thoughts and feelings about themselves (self-image) [114]. However, in the 1970s, Henri Tajfel and John Turner explored how individuals may align with a group's identity resulting in acceptance of extreme behaviour towards the out-group. Tajfel's interest was attributed to his passion for understanding the tragedies that occurred during the second world war alongside his interests in categorisation and social perception [178]. Specifically, Tajfel was interested in understanding why prejudice exists, as well as discrimination and stereotyping. Tajfel and Turner's work resulted in *Social Identity Theory* (SIT) [92, 182], which was the first explanation concerning inter-group conflict and the role that a person's identity plays.

The theory defined social identity as *"that part of an individual's self-concept which derives from his knowledge of his membership of a social group (or groups) together with the value and emotional significance attached to that membership"* [180]. Perceiving an individual's social identity is a process of three steps, namely these steps are *categorisation, social identification, and social comparison.*

The first step known as *social categorisation* is a cognitive process which divides individuals based on social groups [165]. Individuals identify the social groups around them and categorise themselves and others as 'us' meaning they would belong to the same group as the individual (in-group) or 'them' meaning they belong to an outside group (out-group) [179].

Categorisation enables individuals to understand behaviours that are associated with a group, such as social norms. Social norms are unwritten rules that define how individuals behave within a social environment [20]. This leads to the second step, *social identification*, where individuals adopt the social norms associated with the group. One reason for individuals to conform to the social norms of the group is their need to be identified as a group member [104].

The final step of perceiving identity is *social comparison*, this is the step where individuals begin to compare their group's situation with other groups [182]. Social comparison leads to inter-group relations, these relations can result in conflicts, such as prejudice and discrimination, or cooperation among groups and between group members [91].

Perceiving social identity often leads to an assumption about the characteristics and behaviours of an individual and attributing those characteristics to the group they are perceived to belong to, this assumption is known as stereotyping. Stereotyping is a cognitive process where third party individuals are categorised together through a perception of common characteristics [74]. This is well known to be a divisive phenomenon in the human world [52, 183, 193].

In combination, the previous steps together form the essence of social identity theory as stated above. The theory defined a person's social identity through their group memberships. It facilitates understanding of different behaviours, such as cooperation and competition, between group members and intergroup relations. Social identity theory shows that individuals favour cooperating with others in a group because of how they identify themselves as part of the group [154].

One criticism of SIT has been that the model does not explain the strong personal agency that group members often exhibit on behalf of the group [94, 177]. To comprehend how this may occur, researchers have proposed a theory known as *identity fusion* [176, 177]. Unlike social identity theory, individuals are assumed to retain a strong sense of personal identity, however this becomes strongly intersected with that of the group, or *fused*.

An individual is fused with a group when they have a feeling of oneness with the group's identity and its reason for existence. Degrees of fusion towards a group can vary from non-existing to highly fused [88]. Highly fused individuals regard themselves as an important part of the group, but similarly they regard their group membership as part of who they are as an individual [175]. This arrangement has been used to help explain the actions and behaviours behind a so-called devoted actor [8] who is willing to self-sacrifice for a group's existence. Additionally, identity fusion explains how individuals align themselves within multiple groups with various degrees of fusion [200].

While both SIT and identity fusion have strong qualitative elements such as the role of emotion on individuals, there are also elements that can inform the presentation and modelling of group identity for cooperation systems. In particular, identity fusion indicates that the composition of identity in agent-based models needs to be flexible so that it allows for identity to be shared with others on a partial basis. That is, the model needs to allow for an individual to retain a degree of personal reputation as well as having an element of group reputation. These theories have helped to inform our modelling in Chapter 6.

2.4 Cooperation

Cooperation is defined in the Oxford English Dictionary as “working together towards the same end” [49]. Cooperation is fundamental as it defines a basic interaction between individuals and groups. Traditionally cooperation has been observed between relatives, human and animals [45, 53, 102, 168], and this is known as *kin-selection* [84]. Cooperation can also be observed in other contexts such as the eradication of smallpox in the 20th century when countries, led by the World Health Organization, worked together to find and distribute a vaccine [30]. There is an increasing number of works that identify cooperation as a fundamental component of morality [44].

Cooperation has been studied using lab experiments [158, 206], surveys [164] and more recently researchers have applied computational simulations [11, 108] to logical representations of a social dilemma. In all these studies, Game Theory has been central to providing a logical formation of a cooperative decision: in other words whether an agent, human or computational, chooses and is incentivised to cooperate or defect [169, 170]. *Reciprocity*, is a central concept that has been central to developing the current understanding of cooperation.

Reciprocity is defined as “an exchange of things with others for mutual benefit” [136]. The act of reciprocity is the “return of helpful and harmful acts in kind” [127]. In early literature, researchers referred to this form of cooperation as reciprocal altruism. In reciprocal situations, an individual in an exchange is helping others while expecting a return of help in the future. This means that a time lag may exist between the first action and the action in return [13]. Researchers have identified two forms of reciprocity in the literature, referred to as *direct reciprocity*, where the exchange happens between the same individuals of the interaction, and *indirect reciprocity*, where a third party is involved (see Section 2.5).

Direct reciprocity involves repeated interactions between two agents that have the option to either cooperate with each other or defect [73, 160, 192]. Because of the repeated nature of these interactions, the parties involved can make choices that are dependent on previous track-records of the other party. The concept of direct reciprocity can be summarised with the principle, "I scratch your back and you scratch mine". Direct reciprocity has been the subject of many studies through the repeated Prisoner's Dilemma game theory framework [141]. Tit-for-tat [13], and win-stay lose-shift [123] are examples of many strategies that have been proposed by researchers to promote cooperation within the framework. Studies of direct reciprocity have been used to explain why cooperation may occur between unrelated individuals or different species.

Direct reciprocity is restricted to the notion that individuals will repeatedly exchange with each other. However, direct reciprocity does not reflect real-life interactions such as those represented in so-called "one-shot" scenarios [99], in which individuals only meet once. Indirect reciprocity is the more general scenario where individuals help those who have helped others in the past with no restriction placed on the time dimension or the number of individuals involved in the interactions.

2.5 Indirect Reciprocity

According to Nowak, [127], *indirect reciprocity* is "the return of a helpful or harmful act that was directed not at us but at others" [4]. Undertaking indirect reciprocity represents a social dilemma for the individual, who has no incentive to incur the cost of helping others unless they can use it to lever a future benefit for themselves. Despite this, high levels of indirect reciprocity are seen in human societies, indicating that mechanisms to promote this cooperation are in play. Understanding why this occurs has been long-standing [119]. Indirect reciprocity assumes that individuals have not met previously and are not expected to meet in the future.

Indirect reciprocity was first proposed in 1971 by Trivers and was originally named 'third-party altruism' and was also known as 'generalised reciprocity' [192]. Later in 1987, Alexander coined the now commonly known phase 'indirect reciprocity'. The literature states that there two forms of indirect reciprocity, upstream and downstream reciprocity. Upstream reciprocity, commonly known as "pay it forward", considers the generosity urge that the receiver of a donation has and in turn, makes a donation of their own to a third party [126]. Downstream indirect reciprocity is a mechanism for the evolution of cooperation which is allowed through reputation [28, 67, 121, 128, 131, 196]. Downstream indirect reciprocity utilises reputation systems, (see Subsection 2.5.1), that independently verify an individual's standing without the necessity of having previous interactions.Downstream indirect reciprocity facilitates exploring reputation and its links with psychology. For the rest of this thesis, we will refer to downstream indirect reciprocity simply as indirect reciprocity.

Indirect reciprocity can be modelled through the donation game theory framework [108, 127, 128, 153]. The donation game is a special case of the prisoner's dilemma, involving two actors known as the donor and the recipient. In each round of the game, an individual has to make a decision on whether or not to provide a donation. This results in a cost c to the donor, and a benefit b to the recipient, and necessarily $c < b$ [128]. Indirect reciprocity is a challenge to model due to the time lag present between donating and receiving the donation. Adopting a game theory approach allows for studies to concentrate on what makes cooperation flourish without having to focus on what rational the agents have [127]. Additionally, indirect reciprocity can be used as a basis to investigate group related issues such as prejudice, social norms and politeness [71, 198].

2.5.1 Reputation Systems

Reputation is defined as the beliefs that are "held about someone or something" [137]. Due to the nature of indirect reciprocity situations where individuals may not meet the same partner twice, the concept of reputation was established to provide a currency through which an individual can gain recognition for their good behaviour and use this to help secure future reciprocal benefits [156, 173]. Reputation systems present an essential solution that rewards good behaviour and deters bad behaviour [59, 121, 196].

Early researchers, such as Alexander in 1987 [4], noted that for cooperation to emerge in a complex social system that involves indirect reciprocity, a way of judging others based on their interactions was essential. Reputation was established as a critical mechanism that supports indirect reciprocity [127]. Individuals may not be aware of their partner's behaviour, but they can rely on their partner's reputation before making a decision. Furthermore, reputation systems prevent defectors from exploiting systems for their own gain (see Section 2.5.2).

Indirect reciprocity is frequently considered in the context of the donation game, where an agent has to decide on whether or not to provide a donation. Reputation systems act to signal an agent's overall donation behaviour to the population. Because other agents may use an agent's reputation in deciding whether or not to donate, there is an incentive for all potential recipients to maintain reputation at a sufficient level to yield future donations [59, 121, 196]. Reputation is updated, after every action taken by the donor, through assessment rules (see Section 2.5.2), this is a common approach as seen in the literature [28, 155, 199].

Models of indirect reciprocity conventionally assume that each individual is represented by a unique reputation: in other words, an individual's behaviour is entirely identified and judged by their actions. While this has been necessary to understand the dynamics, in reality, humans are prone to taking short cuts in judging the reputation of others, using cues such as group membership or common traits as a proxy. Although this is cognitively convenient, this can fuel stereotyping [74] and results in prejudicial behaviour and discrimination [130], with divisive societal consequences [103]. The sharing of reputation introduces opportunities for freeloaders to exploit the reputation without having to justify or contribute.

2.5.2 Action and Assessment Rules

In early works that studied indirect reciprocity through the donation game framework, researchers proposed several strategies concerning how to make a donation decision effectively promote cooperation [128, 171]. Later studies broke down the idea of a strategy into two components, one being action rules and the other being assessment rules [28, 132]. *Action rules* are a set of rules that determine if an agent should donate or defect based on their partner's reputations [28, 155, 199]. *Assessment rules* allow the system or other observers to judge the decision taken, while considering the reputation of both parties, and update the reputation associated with the agent acting as a potential cooperator [26, 28, 127].

Four action rules have been widely used throughout the literature; these are always cooperate, always defect, discriminate towards good reputation holders, and discriminate towards bad reputation holders [155]. The 'always' action rules indicated that agents should always act in one way regardless of their partner's reputation. In discrimination rules, action rules prescribe to agents their actions based on their partner's reputation. More recently, studies have emerged that base action rules and assessment rules on the principles of social comparison [198, 199]. In these studies, the cooperator compares their reputation and their potential recipient's reputation before making their decision to donate or not.

Following an action, assessment rules are the criteria in which donors are judged.

These are the criteria by which a donor's reputation is adjusted in light of their actions, and therefore govern the extent of reward over penalty. The update of reputation takes place after every donation decision made by a donor. Assessment rules are used as a mechanism that allows cooperation to flourish, in particular, when considering that agents only have limited knowledge of how others have behaved in the past. In this sense, they have been considered as a model for morality [4]. Three main methods for assessment of cooperative action are standing, judging and image scoring.

Sugden [171] first developed *standing*, which was originally conceived assuming that agents had binary reputations. This assessment rule effectively classifies each individual in the population as either good or bad, penalising the good if they donate to the bad. Standing allows for agents to have a good standing as long as they donate to the population and only defect to agents who have a lower standing [128]. This assessment rule has been adopted by several models of indirect reciprocity, as researchers feel that it is close to what works best in human societies. The standing rule is also simple to implement but has been considered as stable due to it preventing free-riding agents from gaining an advantage. The *judging* assessment rule adopts the same rules as the standing assessment rules; however, it additionally states that a donor is punished if they donate to an agent with a lower reputation than their own [101, 131, 140].

Image scoring [128, 196] presented the first significant alternative to the standing assessment rule, where reputation is simply incremented or decremented in response to donation or defection, respectively. A limitation of image scoring is that those who choose not to cooperate with defectors may be unfairly labelled as less cooperative [131, 132, 142]. Another criticism of image-scoring questioned whether agents would cooperate purely based on another agent's score rather than focusing on their own benefit [108].

Consequently, with their roots in the work of [171], standing [142] and judging [28] have emerged as the assessment rules that capture 'legitimate shirking', where an agent's defection is justified (e.g., on the basis of their recipient's reputation) [67, 127, 150]. Because of this property, Ohtsuki and Iwasa [132] surveyed all the possible strategies that can be represented in a binary reputation and found that only eight strategies are stable further emphasising standing and judging. However, these discrimination rules have been studied assuming that reputation has a binary representation [29, 134], although this was generalised for standing in [199] to allow for standing to apply for the integer range $[-5, 5]$ (see Section 3.5.2).

2.6 Evolutionary Processes

As agents interact through cooperative games, over time they have the opportunity to update their action rules. Evolutionary processes are one mechanism through which this can be achieved. Work in this area dates back to the theory of natural selection in 1859 [46], in which the author held the belief that selection operates at the individual level particularly in organisms. Natural selection is a process where organisms adapt to their environment by changing their genotype through selective reproduction [57]. Natural selection is also known as "survival of the fittest", which highlights the competitive nature of the changing process. Cells compete in order to survive the changes in their environment and adapt the best characteristics to produce well-adapted offspring. In the case of indirect reciprocity, the most successful agents, based on fitness, are chosen and their action rules are adopted by other agents. A renewed interest was generated after the publication of Axelrod's *The evolution of cooperation* in 1981 [13], which has attracted many researchers from different disciplines to study cooperation. Researchers were interested in understanding why individuals help each other even though natural selection favoured those who are in competition [69, 146].

Researchers have identified several evolutionary mechanisms through which cooperation can emerge and be sustained (e.g., [124]). These mechanisms structure how individuals of a population interact when they compete for reproduction and to receive benefits. In the 1960s, Hamilton [84, 85] identified two mechanisms for evolving cooperation. The first, known as *kin selection*, promotes cooperation through agents that share a gene [68].

The second was known as the *green-beard effect*, in which agents recognise other cooperator agents through tags, or labels that they have [98].

These mechanisms provide ways for cooperation to evolve without individuals having to remember any past interactions or having to know what other individuals' behaviour might be [152]. Additionally, the mechanisms explained why individuals cooperate with others who are similar to them, by belonging to the same group or sharing a gene or a label. The mechanisms, however, did not provide an explanation for observed cooperation between non-related individuals. Instead, reciprocation was seen as a primary approach to explain cooperation between unrelated individuals, especially in human societies [159] (see Section 2.4).

These mechanisms, along with implementations of indirect reciprocity, relied on selection through individuals [108, 128]. However, deviation from this assumption has occurred in the biological literature, specifically concerning the plausibility of group selection such as in [161, 202].

2.6.1 Group Selection

Group Selection is the mechanism that selection occurs not on the individual level but rather on the group level [161, 202, 205]. V. C. Wynne-Edwards was one of the first biologists to advocate for group selection [208]. The mechanism proved controversial, and early biologists argued against it by suggesting that individuals would not 'sacrifice for the sake of the group' [47, 147, 201]. Group selection was later revisited when the idea of multi-level selection was proposed.

In 1994 David Wilson and Elliot Sober introduced the idea of multi-level selection, in which they argued that selection occurs on both individual and group levels [203]. Similarly, in 2006, Traulsen and Nowak published, *Evolution of Cooperation by Multilevel Selection* [189], in which they acknowledge the debate about group selection and argued that group selection and individual selection could co-exist through multi-level selection.

However, this remains a controversial theory, as discussed by Pinker (2012) with some researchers arguing for multi-level selection [76, 129].

This thesis neither advocates nor opposes this process but is based on asexual reproduction of agents [108, 127, 199]; reproduction is discussed further in Chapter 3. However, groups are used, see Section 2.2, namely through identity which an individual uses to sustain their reputation. Reputation systems can feature in this context, allowing individuals to potentially switch between individual and group reputations [118, 172].

2.6.2 Individual and Group Identity Within Indirect Reciprocity

Most studies of reputation focus on an individual's reputation without considering groups that they may belong to [108, 128]. Individual reputations are restrictive, particularly when agents consider others whom they have not come across before. Therefore agents tend to use cognitive shortcuts, such as stereotyping, to make decisions so that our judgement of the other agents can be better informed.

Limited research has been conducted on the subject of groups, specifically where personal and group identities are combined. One of the earliest adaptions of group reputation is Baranski et al (2006) [16]. The authors introduce a reputation that is shared by members of a group, which individuals use when interacting with members of other groups. Additionally, the model allows for individuals to have a personal reputation. The model calculates a group reputation as the average of all individual reputations in a group. Therefore each interaction of an individual affects the reputation of the whole group. Additionally, agents have a cognitive memory that stores how all other agents have interacted, which is not seen in other literature. This model is limited and does not generalise to allow for members to be in multiple groups.

Another model which considers an agent's reputation within groups was presented in *Ingroup Favoritism and Intergroup Cooperation Under Indirect Reciprocity Based on Group Reputation* [118]. A group structure is proposed where individuals interact within their groups using a personal reputation. When an interaction is between groups, members adopt a group-level reputation. Therefore the paper proposes a two reputations model, one for the inner group and a group's reputation for intergroup interactions. A third-party observer updates reputations. The model is limited by not allowing members to be in multiple groups. Reputation is also limited to its binary values (good or bad).

These models do not allow for individuals to share subsets of traits, or aspects of their identity, and depend on individuals belonging to a single group. Our approach is to allow individuals to have a more complex composition of their identity, based on the assessment of multiple traits against which reputations are maintained.

2.6.3 Cooperation Within Groups

Groups have also been the focus of literature related to the free-rider problem. The free-rider problem occurs when an individual benefits from cooperation without incurring a cost, i.e. they receive benefits but do not pay any costs towards it [61]. In other words, free-riders exploit the benefits of being part of a group without incurring any costs [106]. Punishment has emerged in the literature as a mechanism that promotes cooperation within groups while deterring free-riders. Punishment is a form of cooperation, as it allows members of a group to work together to punish free-riders [25, 61, 135, 190].

The literature focuses on two types of punishment, costly punishment and pool punishment [89]. Costly punishment incurs a cost from all participants to punish the free-rider. Costly punishment is also referred to as peer-punishment because group members directly punish the free-rider. In contrast, pool punishment gathers the cost from group members and punishes free-riders through a third party. The development of pool punishment emerged from a vulnerability in peer punishment where second-order free-riders benefited from not paying any costs towards punishing the original free-riders [139].

A common theme among these works has been that individuals belong to a fixed group [105]. Fixed groups are limited in that individuals are identified through a shared group identity and reputation. As discussed in Section 2.3, studies in psychology have shown that individuals retain both a personal and social identity, where a personal identity intersects with a group identity. The work in psychology has driven our work to relax the boundaries set by the literature on groups. The relaxation allows us to expand individual identities to include both personal and group identities and reputations. Our work goes further in allowing individuals to adapt several identities, through traits, that compose an individual's identity.

2.6.4 Traits

Traits represent identities which an agent may have. Traits are defined as "a distinguishing quality or characteristic, typically one belonging to a person" [138]. Traits were often investigated in the field of social psychology, referring to persistent characteristics of individuals [2]. The surveyed literature lacks a unified approach in introducing identities, particularly identities that are shared. There exists a need to structure identities as this area has not matured yet, and one such concept that has recently emerged is the linking of agents to traits. Trait and set membership have also received attention as simple signalling mechanisms to promote the evolution of cooperation.

Traits were first introduced in [12] by Axelrod. The model used traits to represent values that a feature, that an agent's culture, may have [107]. In cooperation scenarios traits were introduced in [152] without the use of reputation, and these elements have been represented as abstract tags that are sufficient to incentivise some level of cooperation, which is known as the green-beard effect [124, 152]

The Evolutionary set theory shows that more complex set structures can promote the emergence of cooperation even in the absence of other incentives [110, 125]. Cooperation emerges through sets, groups, that individuals share with others, where cooperators belong to several sets which allows them to be more selective in choosing with whom to cooperate. Tarnita et al. [186] proposed a model based on the evolution of sets where the degree of shared membership is based on the overlapping of sets of multiple traits. In their model, the interaction is limited to traits that they have in common with others. This limitation results in agents adopting one of two action rules either to cooperate or to defect. The model allows individuals to update their action rules and set membership under evolutionary settings. Moderate levels of cooperation can be sustained with a limited mutation on traits [122, 185]. Also, Li et al. [109] adopted the same model to study the evolutionary dynamics of minimum-effort coordination games in structured populations.

Similarly, there have been studies that investigated how cooperation can evolve due to in-group favouritism [71, 154]. In Fu et al. [71], agents can move between sets as they are attracted to successful sets. Additionally, agents can adopt different strategies that allow them to differentiate between in-group and out-group interactions. Their results showed that cooperation is achieved by applying a limited mutation on sets. This finding is discussed further in Section 5.3.

In Gao et al. [75], the authors describe a model in which individuals are in groups and interactions may occur between in-group members and across groups. The model does not allow for individuals to have membership in more than one group. Individuals have two strategies that enable them to act differently toward in-group and out-group members. Although their model does not rely on reputation, it allows for mutation during the reproduction phase. Mutation, in this case, occurs on traits and strategies.

Previous observations have led to a need for a new structure of how reputations may be shared, which is one of the contributions of this thesis. Traits represent identities, and individuals may share traits which represent how groups function. Reputations can be then measured by an agent's identity, which is a better representation of how humans are judged in reality based on the combination of their traits.

2.7 Summary of the Chapter

The important point from this thesis's perspective is that individuals can derive some of their identity from being in a group. This has been a long-standing area of study in terms of human behaviour, and it can be seen in terms of everyday reputations. In this thesis, the aim is primarily to investigate cooperation in the context of identity, namely whether individuals cooperate with others without the certainty of a future pay-back, i.e. we explore cooperation using indirect reciprocity. While investigating cooperation, we additionally check if group membership affects cooperation. Groups represent one of the main ways that individuals derive their identity, and this can, in turn, affect their reputation. Furthermore, groups represent a typical way for individuals to share reputations.

The knowledge gap we aim to address in this thesis spans across two different topics: identities derived from groups and the structure of reputation. While sharing identity drives people to cooperate, the structure of how their identities are shared within a group has the potential to allow shirkers to stop cooperation from prospering. Equally, support from within a group can help to improve the payoff for individuals, and reduce risk. By looking at both the sharing of identities and the structure of identities together, we aim to gain a better and more complete understanding of how individuals make up their personal and group identities. The gaps we aim to address are as follows:

The role of identity in indirect reciprocity: One of the primary aims of the thesis is to bridge the gap between identity and structures of reputations. None of the studies carried out so far, to the best our knowledge, have examined identity in the context of indirect reciprocity. Elements of personal and group identity derive individual reputations in the psychology literature, but so far have been absent in indirect reciprocity models.

The sharing of identities: There is a focus on thesis on how agents may share identities through group membership, and we are especially interested in the structures of sharing and how these might affect cooperation. Sharing of identities may allow some agents to exploit their groups, and we investigate ways to prevent this. We further examine the role of identity by allowing agents to learn from others and adopt different identities.

Identity within groups: The relationship between individuals and their groups has been studied in psychology extensively. However, within indirect reciprocity, there has been limited research on this. Studies have not considered the possibility of agents being in multiple groups. Nor has the possibility of a partial group membership been considered. We investigate the role that individuals dedicated to their group have on cooperation, and how the group rather than economic factors may affect the willingness of individuals to cooperate.

The aim of this thesis is to fill the knowledge gap outlined above by addressing the questions:

- What benefit will the addition of traits have for current models of indirect reciprocity? (Chapter 3) How can we model identity within existing model of indirect reciprocity?
- What is the relationship between group identities and stereotyping? Do agents cooperate more with those whom they share a trait with or those outside their trait group? (Chapter 4)
- What happens when agents can move between groups? (Chapter 5)
- Are agents who are fused to their group more cooperative than those who are non-fused? (Chapter 6)
- How does pursuit of identity, rather than of resources, affect cooperation? (Chapter 7)

Chapter 3

Extending Indirect Reciprocity Models to Share Reputation

3.1 Introduction

In this chapter, we introduce a model for cooperation (indirect reciprocity). We extend the existing model so that shared identities for individual agents can be explored. This involves a new framework for modelling reputation based on *traits*. We note that the framework is general and can be applied to other cooperation scenarios that rely on reputation systems. Here our focus is on its application to indirect reciprocity because it provides a well-understood benchmark for investigation.

Our approach is to adopt an evolutionary game theory framework for indirect reciprocity which uses the recent and simple approach of *social comparison* of reputation proposed in [199]. The approach follows the natural human disposition to make relative judgements about the standing of others as compared to oneself [91]. However, the original formulation of the model is limited because it assumes that all agents have a unique identity, represented by their reputation, that is not shared with anyone. We enhance the social comparison model [128, 199] by introducing *traits* as independent components on which reputation is based. Traits may be isolated (i.e., only affiliated with one agent), resulting in the social comparison model, or traits may be shared, allowing a group's identity to be represented. Importantly we also allow an agent's identity to be composed of *mul-*

tiple traits. This approach provides a generalisation of previous approaches by allowing a non-trivial composition of identity for each agent. This also aligns with the psychological literature (Section 2.3) where it is widely acknowledged that group identities are important to an individual.

Generalising the reputation system by using traits opens up many degrees of freedom and therefore expands the range of scenarios that can be simulated and investigated. In particular this allows us to better understand how the sharing of identity impacts cooperation. Underpinning the model, the approach based on [199] involves agent-based simulation, where agents have some freedom in how they adapt their behaviour based on probabilistically copying the strategy and possibly the traits of others, based on perceived success. This represents a form of social learning [15], where agents have limited cognitive ability in their own right, but are able to learn from others in the population. This assumes that agents can change traits (this scenario is discussed in Chapter 5), although it is also possible that some of the traits affiliations could be fixed for a particular scenario (as discussed in Chapter 4).

The overall evolutionary approach allows us to explore conditions that either promote or impede cooperation, taking into account the structure of identity while assuming that agents can socially learn. This should not be confused with agent-based approaches in knowledge engineering, where protocols are sought that allow cooperation to be enforced based on individual behaviour (e.g., [207]).

We also note that the trait based framework is sufficiently flexible to allow a new examination of identity based problems that are motivated by social psychology, such as stereotyping [166, 181] (see Section 2.3). The remainder of the chapter is structured into the following sections:

- In Section 3.2, we elaborate on the significance of agent-based models and their use for research.
- In Section 3.3, we focus on existing models of indirect reciprocity and highlight the

origins of these models.

- In Section 3.4, we introduce our model and highlight how the sharing of reputations is facilitated through shared identity and traits, including:
  - explaining the importance of using traits within our model and elaborate on its usage (Subsection 3.4.1).
  - how we use agents within our models (Subsection 3.4.2).
  - agent's action rules and how they are utilised to make decisions regarding donations (Subsection 3.4.3).
- In Section 3.5, we present the setup of the experiments used in this thesis. This section is divided into three subsections where:
  - we describe the parameters of the experiments in general (Subsection 3.5.1).
  - we consider how reputations are updated using the new concept of traits (Subsection 3.5.2).
  - we present the reproduction step (Subsection 3.5.3).

3.2 Agent-Based Models

In order to study social dilemmas in a range of different scenarios, agent-based approaches have become important. Agent-based models (ABM) are a computational tool that simulate the behaviours and interactions of agents [43]. The models that are of relevance to this thesis represent a simplification of social situations [79] where agents can be imparted with a particular level of cognition and action. There are several other areas of work in multi-agent systems where the focus is to engineer protocols or rules that seek to ensure cooperation is followed. These approaches aim to disincentivise deviation from behaviours that benefit the public good [207]. ABMs have been used to study and analyse the behaviour of autonomous agents within a system, for example, the study of human behaviour [23]. The models involve several elements that enable systems to be studied in detail; these elements include game theory, complex systems, multi-agent systems, and evolutionary programming.

Each ABM is composed of a number of agents that interact within a specific environment. These agents are autonomous but have a limited cognition. Agents may represent individuals, systems, or any number of entities that are being modelled. Agents observe their environment and make simple decisions such as to cooperate or defect. Furthermore, agents can adopt successful strategies from the population. Agents are employed to consider different interactions within specific environments which allows researchers to investigate their impact.

ABMs have been used extensively to study cooperation scenarios through evolutionary game theory [9]. One of the earliest adoptions in cooperation was Axelrod's famous computer tournaments in which experts investigated strategies for the Iterated Prisoner's Dilemma game [14]. Similarly, many indirect reciprocity experiments have been adopted into an agent-based model beginning with Hamilton and Axelrod's work in [83]. These works have led to computer simulations being central to the study of cooperation with more research conducted using simulations than any other medium. This is because computer simulations and ABM, allow for greater flexibility in the setup of experiments in the study of cooperative and competitive behaviours [56].

Although ABMs have many benefits, they do not come without limitations [113]. One limitation is performance because agent-based simulation are only a model; in some circumstances, they cannot cope with large populations [17]. Some researchers question whether agents can be programmed to reflect real human behaviour, as behaviour is complex and can vary dramatically [79]. However, ABMs are an excellent tool for experiments in the field of social psychology as they can mimic real-world scenarios without compromise on the amount of detail needed as long as it can be computable [54]. Furthermore, ABMs present a simple approach for researchers to share their results, specifically when using sophisticated mathematics to solve evolutionary problems, thus giving their work more reach [9]. This is due to the nature of ABMs as they have been able to expand on the mathematical treatment of evolutionary game theory [1]. Agent-based models expands this treatment by allowing agents to be cognitive and allows agents to act independently when making decisions.

The tool we use to implement our agent-based model is Python, which is a high-level programming language that has a growing number of tools for analysis and modelling [116, 148]. As we built our model using Python, it allowed us to quickly extend the model and implement additional features.

Agent-based models present a conventional tool that allows us to simulate cooperation problems and experiment with various scenarios. Our use of ABMs allows us to implement an extended model of indirect reciprocity, where we explore different behaviours between agents that have limited cognition, and the effect of these behaviours on cooperation.

3.3 Existing Models of Indirect Reciprocity

A social dilemma is a situation in which individuals face a conflict choosing between their self-interest and the collective interest [38]. In other words, a social dilemma is a situation in which an individual faces one of two actions, one that benefits them, and the other which may harm them but benefits the collective. A variety of game theory models have been proposed by scientists to study various social dilemmas.

3.3.1 Game Theory and Cooperation

Game theory allows the study of problems of cooperation and conflict within social situations through a model [34, 163]. The study of social dilemmas through models has attracted a lot of attention since the 1960s [85, 115]. Famous models have included the basic public goods [36], prisoner's dilemma [10] and snowdrift games [51]. Similarly, indirect reciprocity has been modelled using game theory as it is a social dilemma situation where individuals choose to donate to others, paying a cost, based on their reputation without any guarantee of a benefit or to defect and not incur any cost.

Indirect reciprocity is a form of cooperation supported through reputation [4]. Indirect reciprocity is the cooperative act to help others in response to their generosity to a third party. Similarly, it can be thought of as a punishment by third party individuals in response to someone's cheating. Reputation supports indirect reciprocity as it informs that population of the individual's behaviour. Reputation systems have been studied alongside indirect reciprocity since the introduction of indirect reciprocity by Trivers [192].

Indirect reciprocity has wide-ranging models such as [108, 128, 132, 184]. Nowak and Sigmund's paper [128] is considered to be one of the first to investigate indirect reciprocity in terms of evolutionary game theory. Their approach was to implement a computer simulation of the donation game (see Section 3.4). The paper established the evolutionary significance of indirect reciprocity [6]. Additionally, the paper highlighted the importance of reputation by introducing the image score assessment rule. Image scoring updated reputation after every interaction conducted by agents as a way of mimicking gossip spread in real human behaviour [127]. Since then, researchers adopted the simulation and developed it further to implement and test various game theory approaches towards solving problems using indirect reciprocity. Table 3.1 highlights how researchers have adopted the framework and how they contributed to the further understanding of indirect reciprocity.

Alongside computer simulations, indirect reciprocity experiments with human subjects have been conducted in labs [96]. The experiments utilise a game theory approach by giving players a choice to cooperate or defect and matched players randomly. The experiments' primary objective is to demonstrate that cooperation can emerge based on indirect reciprocity. In one such experiment, the author showed that reputation is vital for cooperation to emerge in indirect reciprocity [29]. Furthermore, the author showed that defection is contagious in situations where reputation was not accessible [101]. However, since Nowak and Sigmund's paper, the focus of experiments has moved from showing that cooperation can emerge from indirect reciprocity to what strategies are best used to explain this.

A new model has been proposed by [199] which is based both on the work of Festinger's social comparison [65] and Nowak and Sigmund's model [128]. The social comparison model continued to use the donation game as the basis, however in contrast to [128] it uses social comparison behaviour which simplifies the modelling and links to work on behaviour in psychology. The key point in using social comparison is that it does not require global thresholds for action to be defined and triggered [128]. Instead, each agent takes a personal view and compares themself with others in order to make a judgement on how to act. This removes complexity from modelling - for example, in making a comparison against a function f carried by another agent, then the other agent must hold a value that is either similar, lower or greater. Additionally, this assessment is made with respect to the agent's own "world view", as defined by the value of f that it is attaining itself.

The social comparison of reputation allows cooperation to be sustained through the evolving heuristic of donating to those with similar or higher reputations. Similarly, the new model proposed a generalisation of the standing assessment rule, adopted from [171, 132], this permitted reputation to have a range between -5 and $+5$. However, the model does permit for other assessment rules to be used. The paper presented a simple model with specific parameters that enable experiments to be conducted in a swift manner, the model itself allows for an extension without restriction [198].

	Game Theory Model	**Assessment Rules (social norm)**	**The strategies Available to Agents**	**Reputation**	**Paper Focus**
The Economics of Rights, Co-operation and Welfare [171]	Mutual aid game	Good standing	Each player had one of two strategies: co-operate or defect	Binary reputation: Good or bad	The paper was credited with introducing the original standing assessment rule as a solution for the free-rider problem
Social Norms and Community Enforcement [101]	Mutual aid game	Judging	Each player can choose to cooperate or defect	Labels are used as reputation and are updated after every interaction	The paper is credited with introducing the judging assessment rule
Evolution of Indirect Reciprocity by Image Scoring [128]	The donation game	Image Scoring	A donor donates only when the recipient's score is at least the same as the donor's	Reputation based on a score that is capped between -5 to 5	The paper introduced image scoring as a way to keep track of agents' reputations
Evolution of Co-operation Through Indirect Reciprocity [108]	Island model	The paper was comparing different image-scoring with standing for evolutionary stability	Different strategies were used these compared the reputation of the donor with that of the recipient. A new strategy was introduced based solely on the donor's history and score	Reputation based on a score that is capped between -5 to 5	The paper compares different assessment rules to find the most evolutionary stable
Indirect Reciprocity Can Stabilize Co-operation Without the Second-Order Free Rider Problem [143]	Mutual aid game	Standing	Each player has one of three strategies: defector, cooperator, and ’shunner’	Binary reputation: Good or bad	The paper was attempting to solve the second-order free rider problem by suggesting a new assessment that prevents the problem
Evolution of Direct and Indirect Reciprocity [153]	The donation game based on [128] and the island model when using the standing assessment rule	Both Image-scoring and standing	Strategies were based on the score of the donor	Reputation based on an integer scale between bad (-1) and good (1) but starts with a neutral (0)	The paper compares the image-scoring assessment rule with that of standing
A Dominant Social Comparison Heuristic Unites Alternative Mechanisms for the Evolution of Indirect Reciprocity [199]	The donation game	They compare different assessment rules: image-scoring, judging, and standing.	Eight strategies are in use these allow agents to compare their reputations before taking a donation decision	Reputation based on a score that is capped between -5 to 5	Introduces social comparison
Evolution of Co-operation Under Indirect Reciprocity and Arbitrary Exploration Rates [155]	Donation Game	They use a variety to compare: Image-scoring, shunning, simple-standing, and Stern-judging.	Each player has one of four strategies: always donate, always defect, discriminate towards good, discriminate towards bad	Binary reputation: Good or bad	The paper introduced mutation to agent's strategies and explored various rates of mutation

Table 3.1: Summarises the Key Papers That Have Contributed Towards the Indirect Reciprocity Model.

3.4 Indirect Reciprocity Model Based on Social Comparison

Since its introduction, the *donation game* has been adopted by many researchers to conduct experiments on indirect reciprocity and cooperation. Both lab and computerised versions of the donation game involve pairing up participants in a random manner and giving them various conditions to measure cooperation and reciprocity. The donation game captures interaction in its simplest form, between two individuals who take one of two decisions: either donation or defection.

The donation game is a subclass of the mutual aid game [171] where the donor incurs a cost with no guarantee of reciprocation from the beneficiary, or any other individual. While the mutual game involves one aid recipient and multiple donors, the donation game is between a donor, a recipient, and an outside observer. The donation game is modelled through prosocial donations which result in a cost c to the donor agent and a benefit b to the recipient, where $b > c > 0$. Our implementation of the donation game is performed on a set of agents, A, representing a population of individual agents and generalises the model developed in [128]. Our assumption is that agents represent individual entities that have a simple cognition. The limited cognition allows agents to take simple actions, either to donate or defect. Furthermore, we assume the cognitive ability of agents allows them to make comparisons against themselves; specifically, agents compare reputations. The cognitive limitation is intentional, as the approach is seeking to facilitate an understanding of the components that impede or enhance cooperation. A further assumption of the model is that the economics of pay-off drive agents' behaviour in the model when agents seek to update their strategy; this is a simple rational notion that is driven by agents to act selfishly. In doing so, we are assuming that agents are able to perform simple social learning (Section 3.5.3).

3.4.1 Trait Representation of Identity

Traits are features that are held by agents and represent identifiable characteristics. In this thesis, we generalise the common modelling convention that an individual agent implicitly carries just one unique trait (i.e., through a reputation that is not shared with any other agent). We calculate reputations based on traits, any number of which can be held by an individual. This better represents the fluidity that is seen in the real world, where individuals are rarely totally defined by a personal identity or group identity, but may be represented as a combination of characteristics and affiliations. Unless otherwise specified by the experiment, the traits are assumed to be immutable. The framework of using traits to model reputation is unique to this work, and the introduction of traits allows for agents to share elements of a reputation, with consequences for the agents involved. It allows for a new range of experiments to be carried out, where agents share reputations, either wholly or partially. In this manner, experiments can be conducted on behaviours of individuals who are not fully part of a group, for example, new members of a team will have a different reputation than those who are long time members.

Note that the model allows different types of reputation sharing with other agents based on the trait(s) that are held in common. We say that an agent is *dependent* if it shares at least one trait with another agent, otherwise, the agent is *independent*. Furthermore, letting T_i denote the set of traits held by an agent i, if $|T_i| > 1$ then i is a *multi-trait* agent, otherwise i is a *single-trait* agent. These key parameters are defined as follows:

Definition 1. An agent is *independent* if it does not share a trait with any other agent.

Note that independent agents may have one or more traits. Examples of independent agents are all purple agents in Figure 3.1b, each of which has a single trait. The lack of sharing means that the reputation of independent agents does not support shirking by other agents. This is because their own actions, and only their own actions, are represented by their own reputation. When this is not happening, the opportunities open up for shirkers to benefit, leading to the following definition.

Definition 2. An agent is *dependent* if it shares a trait with at least one other agent.

As an example, all green agents in Figure 3.1b are dependent agents. Agents 2 and 5 share a single trait, while agent 3, which has multiple traits, shares two traits with agent 4 and and a single trait with agent 6. Dependent agents present an opportunity for shirkers to exploit reputations that they share with agents who have the same traits as them. This allows shirkers to increase their payoff and avoid paying costs while maintaining their reputation.

It is also useful to consider the number of traits that represent an agent, as in the following definitions.

Definition 3. An agent is a *single-trait* agent if it has only one trait, i.e. an agent i is a single-trait agent when $|T_i| = 1$.

Definition 4. An agent is a *multi-trait* agent if it holds more than one trait, i.e. an agent i is a multi-trait agent when $|T_i| > 1$.

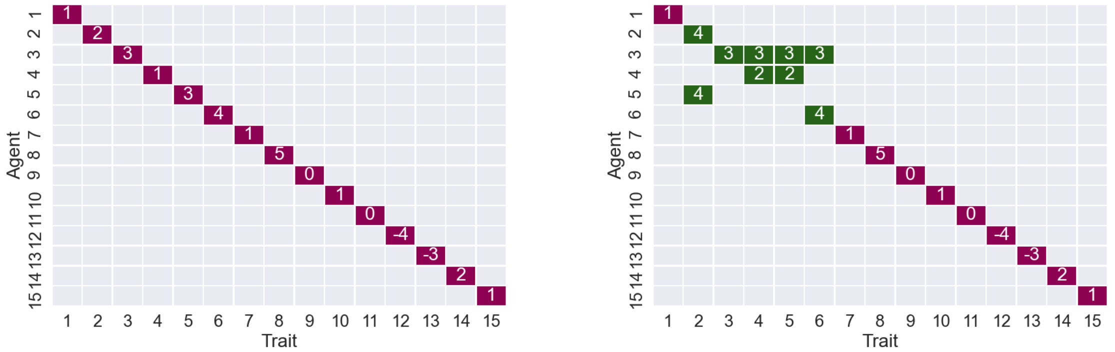
Figure 3.2 shows an example of agents sharing traits. Agents A and B are dependent agents as they share one trait between them, whereas Agent C is independent. Agent A is a multi-trait agent as they have two traits, but agent B is a single-trait agent. Also, note that agent A does not share trait 1 with another agent in this example. This figure shows that our model can be flexible to accommodate a range of experiments whether agents share traits or not. The sharing of traits among agents allows for the sharing of reputation. The model can be adapted to allow agents to share traits in different ways, and it allows individuals not to share traits if required by an experiment.

For an individual's reputation to be represented by a single overall value, there is a need to combine the reputation values that each trait represents. We assume that each trait $t \in T$, where T is the set of all traits in the model, has associated with it a reputation r_t, represented by an integer in the range $[-5, 5]$, and an agent i derives its reputation r^i from the reputations of the traits associated with i through a process of averaging (see Table

3.2 below). Specifically, let T_i (with $|T_i| > 0$) denote the associated set of traits for agent i, then its reputation r^i is defined as:

$$r^i = \frac{1}{|T_i|} \sum_{t \in T_i} r_t$$

In other words, an agent's reputation is the average of the reputation of its associated traits. While not able to capture all scenarios, this assumption allows for reputations to be modelled aligning with the literature, which assumes that all agents' reputations are known.



(a) A simple visualisation of 15 agents showing the distribution of agents and traits in conventional models of indirect reciprocity structure their experiments. Each agent has a single reputation that they do not share, i.e. all agents are independent single-trait agents. As such, sharing of traits or reputation is not permitted. Agents 1 and 3 have reputations of 1 and 3, respectively.

(b) A simple visualisation of 15 agents showing different ways that we can structure the sharing of traits among agents. Agent 1 is a single-trait agent that does not share any trait and therefore does not share their reputation. Agents 2 and 5 are single-trait dependent agents that have one trait which they share. Agent 3, which has multiple traits, shares two traits with agent 4 and and a single trait with agent 6, making agent 3 dependent.The remaining agents all have a unique trait that they do not share.

Figure 3.1: A simple visualisation of 15 agents showing alternative agent-trait relationships showing both single-trait and multi-trait agents. The figures show the relationship between agents and their reputation through traits. The purple colour indicates that the agent does not share a trait, while green indicates that an agent shares their traits with one or more agents. All agents derive their reputations from the traits that they are associated with using the formula in Subsection 3.4.1. For example, agents 1 and 15 do not share any traits and have the same reputation in both figures.

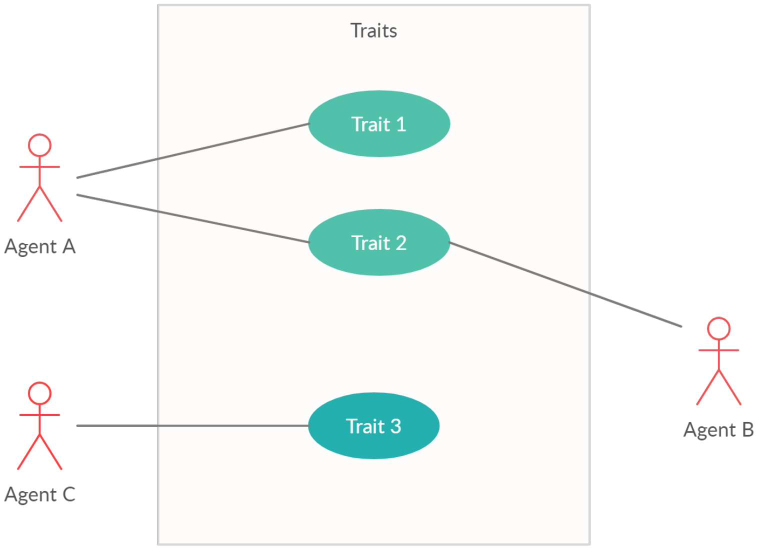


Figure 3.2: Different ways that agents may share traits. Agent B fully shares their identity by sharing trait 2 with agent A. While Agent C is an independent agent as they do not share any trait with other agents. Agent A shares some element of their identity with Agent B.

3.4.2 Agent Attributes

We assume that each agent i has four key fundamental attributes: its set of *traits* T_i, its *action rule* (s_i, u_i, d_i), its *reputation* r^i and its *fitness* f_i. Action rules and traits are defining characteristics of an agent as they dictate how an agent behaves and how they are perceived, and therefore affected, by the population. The action rules are used by an agent in deciding whether or not it should make a donation decision in respect of another agent. In our model s_i, u_i and d_i are binary variables that control the decision an agent makes based on comparison of its reputation with that of the potential receiving agent. This is described in Section 3.4.3. The actions that an agent makes affect the updating of its reputation (or more specifically the updating of the reputation held by traits associated with the agent) - see Section 3.5.2. Fitness represents the economic payoff as the accumulation of costs and benefits that are paid and received by i, based on the total donations received in a generation (see Subsection 3.5.1), less the total cost of making donations in a generation. This is used at the reproduction step (see Section 3.5.3) in order for an agent to update its action rule. Thus an agent i strives to maximise its payoff in a selfish manner, which may or may not result in cooperation emerging, based on the way in which traits are shared.

3.4.3 Agent's Action Rules

Recall that action rules represent the behaviour that agents take when deciding to donate or defect - see Section 2.5.2. Our model adopts the social comparison method from [199] which compares the reputation of donors and recipients before a donor decides to donate or to defect. Each agent i carries a binary vector of variables (s_i, u_i, d_i) which represents i's current *action rule* with respect to i's donation behaviour when it is called upon to consider making a donation to another agent j (see Algorithm 1 line 14). The action rule indicates whether or not i donates when similarity (s_i), upward (u_i), or downward self-comparison (d_i) is observed by i in respect of j's reputation (r^j), as compared to i's own reputation value (r^i). This results in a set of eight action rules: $AR = \{(0,0,0),(0,0,1),(0,1,0),(0,1,1),(1,0,0),(1,0,1),(1,1,0),(1,1,1)\}$. A similarity comparison takes place when $r^j = r^i$, upward self-comparison occurs when $r^j > r^i$, and downward self-comparison occurs when $r^j < r^i$. Section 3.5.2 discusses further how the assessment rules are applied to update reputations after each round of the donation game.

During reproduction, each agent updates its action rule through social learning, as a consequence of observing others in the population (see Algorithm 1 lines 23-33). It is known [199] that evolution promotes the action rule $(1,1,0)$, allowing agents to discriminate against those having a lower reputation than themselves, thereby representing a relative threat. The social learning step takes place during the reproduction phase discussed in Section 3.5.3.

Parameter(s)	Description	Role in model	Reference in Chapter 3
r^i	The reputation of an agent i	Used to represent the status of an agent and is used by agents to determine donation decisions	Subsection 3.4.1
$t \in T$	A trait t belongs to the set of all traits T	Defines a single feature that an agent may have and determines reputations of agents that hold it	Subsection 3.4.1
r_t	The reputation of a trait t	Represents a trait's status based on the actions of agents that hold the trait	Subsection 3.4.1
f_i	The payoff for i	Accumulates costs and benefits from making and receiving donations	Subsection 3.4.2
s_i, u_i, d_i	Action Rules for an agent i based on self-comparison of the potential recipient reputation	Governs i's donation behaviour	Subsection 3.4.3
S	The probability that the selected recipient agent chooses a donor with a trait in common	Governs the selection of a donor agent with different traits	Subsection 3.5.1
N_j	The set of agents that share at least one trait with agent j	Governs the selection of a donor agent that shares at least one trait with the agent	Subsection 3.5.1
c	The cost of donating	When an agent makes a donation the incur this cost and it is reflected in their payoff.	Subsections 3.4 & 3.4.2
b	The benefit received from a donation	When an agent receives a donation they receive a benefit that is reflected in their payoff.	Subsections 3.4 & 3.4.2

Table 3.2: Summarises the Key Parameters Used Throughout the Thesis.

3.5 Performing the Game

This section describes the steps taken to perform experimentation in our model. Recall that our model is based on the donation game scenario with the addition of traits to represent agents' identities. The model has an evolutionary game theory framework based on a combination of game theory and evolutionary dynamics. We further explain the steps of the model by breaking it into several subsections representing different stages encountered during a simulation. In Section 3.5.1, we describe the parameters that set up each experiment and describe the basic interaction that occurs between the agents in the donation game. In Section 3.5.2, we consider how reputations are updated based on their traits and how agents are assessed. Finally, in Subsection 3.5.3, we present the reproduction step that occurs at the end of each generation, in which agents probabilistically copy the action rules of successful agents.

3.5.1 Player Selection

Donation games are played in rounds. Each round is composed of a single interaction between a donor i and recipient j and we assume that the set of agents is denoted A. Each generation is composed of m rounds, and there are M generations in each experiment. We denote $U(X)$ as a function for uniform random selection of an element from the set $X = \{x_1, \ldots, x_n\}$. Algorithm 1 shows the simulation framework used in this thesis.

At the beginning of each round a potential recipient, j, is selected at random from the population (see Algorithm 1 line 4). As j is selected at random from the population, j may be selected multiple times. Similarly, a potential donor agent, i, is selected from the sub-population of agents having at least one common trait with j from the set of traits T_j, with probability S (see Algorithm 1 lines 5-12). Here S is a global parameter (not to be confused with s_i) that governs the extent to which an agent is disposed to playing in-group (i.e., with other agents that have the same traits). Otherwise, when $S = 0$, the donor agent i can be selected at random from the population.

To describe this formally, we let $N_j = \{x \in A - \{j\} : T_j \cap T_x \neq \emptyset\}$ be the set of agents that share at least one trait with agent j and $\bar{N}_j = \{x \in A - \{j\} : T_j \cap T_x = \emptyset\}$ be the set of agents that do not share any traits with agent j. With probability S, the potential recipient j attempts to select the potential donor i from the set N_j. With probability $1 - S$, the potential recipient j attempts to select the potential donor i from the set $\bar{N}_j$. If no suitable donors are found then i is randomly selected from A-$\{j\}$.

Once an agent i is selected to donate to a potential recipient j, i's donation decision depends on i's action rule, (s_i, u_i, d_i) which dictates the conditions that invoke a donation when i compares its reputation with that of j. Recall that an agent i derives its reputation r^i from the reputations of their trait set T_i. Agent i will then compare their reputation r^i with that of agent j (r^j) to determine the donation decision. Figure 3.3 highlights how agent 2 compares its reputation with agent 5 when they are sharing a trait, and when they do not share a trait. Alternatively, Algorithm 1 line 14 highlights how an agent i compares their reputation with that of agent j.

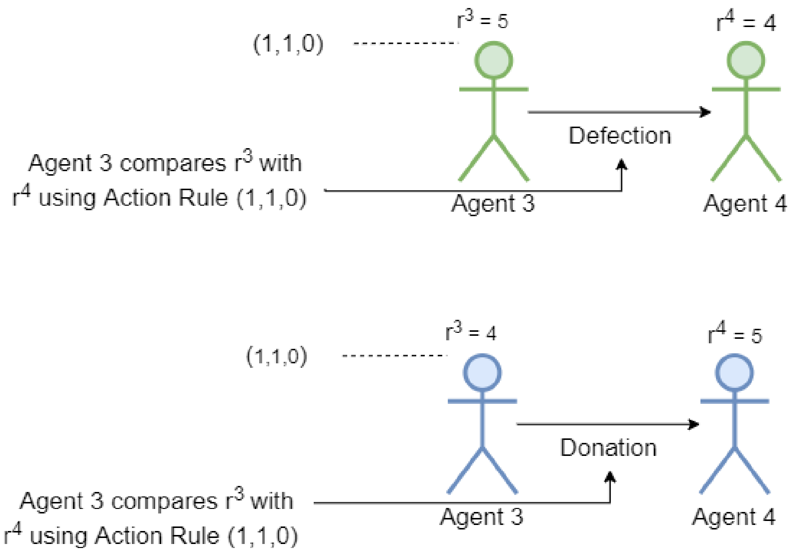

(1,1,0)
$r^3 = 5$
$r^4 = 4$
Defection
Agent 3 compares r^3 with r^4 using Action Rule (1,1,0)
Agent 3
Agent 4
(1,1,0)
$r^3 = 4$
$r^4 = 5$
Donation
Agent 3 compares r^3 with r^4 using Action Rule (1,1,0)
Agent 3
Agent 4

Figure 3.3: Two instances of the donation game based on Figure 3.1. In the first instance, agent 3 compares its reputation with the reputation of agent 4, then makes a decision to defect, as agent 3 has a higher reputation than agent 4, based on its action rule as outlined in Section 3.4.3. In the second instance, agent 4 has a higher reputation than agent 3. Agent 3 makes a decision to donate based on its action rule.

3.5.2 Updating Reputation and Assessment

After every round of the donation game, the reputations of traits that comprise i's reputation (i.e., members of T_i) are updated based on i's donation decision. An assessment rule is applied to these reputations to capture the potential benefit that the population derives from the individual's actions. These can be interpreted as rewards or punishment that an agent i receives based on their donation behaviour. Our model applies the standing rule as it is stable and is simple, being a fundamental assessment rule [142, 171]. This rule has been developed to reward those who donate, while allowing justifiable defections to take place. This could be envisaged in the case of not donating to a shirker (i.e., an agent who never donates to others). In such a case penalising one's reputation due to not donating would be harsh.

Because we are are dealing with the comparison of reputation, the standing rule functions as follows: if i donates, then r_t is incremented, for all $t \in T_i$. If $r^j \geq r^i$ and i defects then the reputation of trait t, r_t is decremented, for all $t \in T_i$. This is shown in Algorithm 1 on lines 14-21.

Note that the standing assessment rule ensures that a reduction in reputations does not occur when i fails to donate when j is of a lesser reputation, providing a defence against possible shirkers. Applying the assessment rule to the reputations of T_i means that an individual's actions equally affect the traits by which it is represented. Each trait's reputation is capped and allowed to vary in the integer range $[-5, 5]$, as developed in the original paper [199].

3.5.3 Reproduction

Once a generation has been completed (i.e., m rounds of the donation game have been played) the reproduction phase takes place. Reproduction is the step where action rules are updated based on individual agents, each probabilistically copying other population members based on their relative success, as measured by fitness (denoted f_i for each agent as introduced in Subsection 3.4.3). The Wright-Fisher model is adopted where the offspring of the previous generation replaces the current population [58, 66, 204, 205]. In performing this, our assumption is that agents are capable of socially learning from others around them by copying other agents, weighted by the relative success of agents in the population. Specifically, each agent i in the population copies the action rule of another agent j with a roulette wheel selection, upon which i adopts j's action rule for the next generation. We define the roulette wheel function as follows:

Definition 5. $R(X, f)$ denotes a random selection from the elements of the set $X = \{x_1, \ldots, x_n\}$ weighted by $f_1, \ldots, f_n$. That is:

$$p(R(X, f) = x_i) = \frac{f_i}{\sum_{x_j \in X} f_j}$$

Similarly, in specific experiments, each agent may update its trait set T_i to reflect the trait set of those that are deemed successful. Fitness for agent i represents the total donations received by i in a generation, less the total costs i incurs in making donations throughout a generation.

Definition 6. The fitness of agent i is denoted as $f_i = b - c$ where b represents all the benefits received by agent i in a generation and c represents all the costs i incurred in the generation.

Fitness or pay-off is an essential component to keep track of the most successful action rules. The most successful action rules become widely used after several generations while action rules generating a lower pay-off are no longer selected. Successful action rules become spread through the reproduction phase when this is repeatedly applied.

Once agents have updated their action rules, mutation takes place. Mutation allows for a change in the action rules with a small probability. Recall that (s_i, u_i, d_i) is a binary vector, and therefore each agent can have one of eight action rules. Specifically, after an agent has been assigned a new set of action rules, based on the most successful action rule, mutation is applied with a probability of $\mu_A = \frac{1}{100}$. In this case, mutation randomly assigns the agent one of the eight action rules. Similarly, in specific experiments, agents may have their traits mutated so that they are assigned a random trait.

Prior to commencing a new generation, fitness f_i is set to zero ($f_i = 0, \forall i$) and for all traits t, reputation is reset $r_t = 0$ is set. Throughout a $\frac{c}{b}$ ratio of 0.7 is applied (see Table 3.2). The chosen ratio is based on analysis of [198, 199] but has been reconfirmed by this work. The ratio is considered to be reasonable but conservative because donations are relatively costly compared to ratios applied in other work (e.g., [27, 37]). Lower ratios would reduce the impact from defectors as donors would incur lower costs when donating to defectors, yielding higher payoffs in return but eliminates strategies that defect. Higher ratios would reduce the payoff received by agents and would deter agents from donating therefore, reducing cooperation. Note that 1 is an upper bound on $\frac{c}{b}$ ratio since when costs are greater than benefits, there is no collective benefits accrue across the population.

3.6 Experimental Assumptions

Once the number of generations M is reached, the evolutionary simulation ends. To evaluate the experiment, we compare the total number of instances of cooperation (i.e., anytime a donation has been made where i donates to j in a donation game) across all generations. Furthermore, we run each experiment over five randomly seeded runs and average the results. Throughout our experiments we use the default parameters of $A = 100$, $M = 100000$ and $m = 5000$, this results in each agent participating in an average of 50 games per generation. These numbers are adopted from [198, 199]. Algorithm 1 shows a simple run of the experiment with lines 23-33 showing how agents evolve through the reproduction system. The algorithm was validated against the works of [199] by running an experiment where all agents i are single-trait independent agents. This scenario used the standing assessment rule as discussed in 3.5.2, along with a cost-benefit ratio of $\frac{c}{b} = 0.7$. The results were consistent with that of [199], producing an average cooperation of more than 90%.

Aligned to the wider literature, we follow a number of conventions concerning the implementation of the model. Firstly we implement a population size of $A = 100$. This scale of population is widely used in the literature (e.g., [3, 5, 23, 71, 81, 128, 133, 153, 188, 198, 199]) because it allows a good balance between providing experimental insights (e.g., non-trivial subgroups can be defined and observed) and completion of a significant number of games with available computational resources.

A second important convention that we follow is obtaining observations from a sample of runs, each with a different random seed. This approach is also used across the literature (e.g., [39, 48, 55, 60, 71, 81, 86, 95, 152, 153, 184]). Due to mutation and the underlying evolutionary forces, alternative starting points make limited difference to the long term evolution of the simulation (as seen in Sections 5.3, 5.4, and 6.5). This is also reflected in the lack, to the best of our knowledge, of variance measures in the literature (e.g., standard deviations) on snapshots of results. We note that most papers in the field do not typically report deviation measures between runs (e.g., [18, 71, 86, 108, 128, 151, 152, 153, 184, 198, 199]), with greater emphasis placed on ensuring that simulations have a sufficient number of generations to achieve convergence. We follow this convention, with $M = 100,000$ generations applied and $m = 5000$ rounds in each generation. This represents each agent participating in an average of 50 games per generation. Where we observe slow convergence, we increase M. Table 3.3 presents the different parameter values used in Chapters 4 to 7.

In terms of reporting results, our primary metric is an assessment of cooperation. We typically report the average cooperation over five runs. Cooperation is the number of donations made as a percentage of the total number of games played.

Parameter	**Description**	**Role in Model**	**The range of values used**	**Notes**
A	The number of agents	It represents the size of the population	$\|A\| = 100$	The value is constant in all experiments
m	The number of iterations within a generation	Each iteration consists of two agents taking part in the donation game	$m = 5000$	The value is constant in all experiments
M	The number of generations	A specific number of generations must be chosen to control the experimental run	$M = 100,000$	The value is constant in all experiments except for Section 7.5 where the number of generations is increased to a million
T_i	The set of traits held by an agent i	These traits make up the reputation of the agent	Agents are classified as single-trait agents, i.e. $\|T_I\| = 1$ or as multi-trait agents, i.e. $\|A\| \leq \|T_I\| \geq 1$	The default in this thesis is to assume that all agents i are single-trait agents, such as in Sections 4.4, 4.4.1, 5.3 and 5.4. However, specific experiments agents allow agents to be multi-trait agents such as in Section 4.5 and Chapters 6 and 7
b	The benefit of receiving a donation	The benefit is received by the recipient	1	
c	The cost of donation	The cost is incurred by the donor	0.7	
r_n	The reputation of a trait or an agent	Reputation determines the value of a trait or an agent and is used for comparison when donation decisions are made	Reputation is capped and allowed to vary in the integer range $[-5, 5]$	
μ_A	Mutation rate of action rules	Mutation randomly assigns the agent one of the eight action rules	1%	
μ_T	Mutation rate of traits	Mutation randomly changes each agent i's trait into any other trait	10% in all sections of Chapter 5 except for Section 5.3.3 where various rates are used	
μ_B	Mutation rate of blending levels	Mutation randomly assigns agents one of seven blending levels identified in Subsection 6.3.2	1%	

Table 3.3: An Experiment Scenario Table That Summarises Parameter Values Used in Chapters 4 to 7.

3.7 Conclusions

In this chapter, we discussed existing models of indirect reciprocity and described how the models are limited and need to be extended in order to explore shared identities.

The limitation transpired in agents not being able to share reputations. We generalised the reputation system by introducing traits to the existing model, thus allowing us to investigate reputations based on traits rather than being based on the individual. Traits represent features or characteristics that agents have; therefore, traits can be unique to an agent or shared with others. The addition of traits vastly expands the range of scenarios that can be simulated using the model.

Additionally, the chapter described the experimental setup for the following chapters. We explained how action rules are used based on trait reputations. Furthermore, we summarised how reputations are updated based on the donation decisions by agents. Finally, we presented the reproduction step and discussed how agents adopt successful action rules for selfish gain. This model provides a new basis for identity and assesses its impact on cooperation, using the model of indirect reciprocity. This also provides the basis to further develop understanding of concepts related to identity, such as stereotyping.

Algorithm 1 Algorithm for the Basic Model of Indirect Reciprocity Based on the Reputation of Traits

Require: Number of iterations m; Number of generations M; set of agents A; set of traits T; set of binary action rules $AR = (s_i, u_i, d_i)$; cost c; benefit b; in-group interaction probability S; mutation rate of action rules μ_A; the set of agents that share at least one trait with agent j = N_j

```
for $M$ generations do                                    ▷ Perform evolutionary simulation
    Set $f_i = 0 \ \forall i \in A$ and $r_t = 0 \ \forall t \in T$       ▷ Reset fitness and reputation
    for $m$ iterations do
        $j \leftarrow U(A)$                                ▷ Select recipient (see Section 3.5.1)
        $p \leftarrow U([0,1])$
        if $p < S$ and $|N_j| > 0$ then
            $i \leftarrow U(N_j)$                          ▷ Select random in-group donor
        else if $p \geq S$ and $|\bar{N}_j| > 0$ then
            $i \leftarrow U(\bar{N}_j)$                    ▷ Select random out-group donor
        else
            $i \leftarrow U(A - \{j\})$                    ▷ Select random donor
        end if
        ▷ Apply action Rules (see Sections 3.4.3 and 3.5.2)
        if ($r^i$=$r^j$ and $s_i = 1$)                      ▷ Compare equal
              or ($r^i < r^j$ and $u_i = 1$)                ▷ Compare upwards
              or ($r^i > r^j$ and $d_i = 1$) then           ▷ Compare downwards
            $r_t \leftarrow min(5, r_t + 1)$                ▷ $i$ donates, increase reputation
            $f_i \leftarrow f_i - c$; $f_j \leftarrow f_j + b$        ▷ Update fitness
        else                                               ▷ $i$ defects
            if $r^j \geq r^i$ then                         ▷ Detect unjustified defection.
                $r_t \leftarrow max(-5, r_t - 1)$          ▷ Decrease reputation
            end if
        end if
    end for
    ▷ Reproduction stage (see Section 3.5.3)
    for $i \in A$ do
        $j \leftarrow R(A, f)$                 ▷ Roulette wheel based on fitness (see Section 3.5.3)
        $(s'_i, u'_i, d'_i) \leftarrow (s_j, u_j, d_j)$       ▷ $i$ copies $j$'s action rules
        if $U([0,1]) < \mu_A$ then
            $(s'_i, u'_i, d'_i) \leftarrow U(AR)$          ▷ Mutate action rules
        end if
    end for
    for $i \in A$ do
        $(s_i, u_i, d_i) \leftarrow (s'_i, u'_i, d'_i)$       ▷ Update action rules for all agents
    end for
end for
```

Chapter 4

Sharing Identity Through Traits and Reputation

4.1 Introduction

Three interconnected concepts characterise an agent. These are *identity, traits* and *reputation*. Traits are used as the basis for an individual's identity, and traits carry a reputation in their own right. This chapter investigates different structures for sharing reputation through traits and their degree of impact on cooperation. Sharing can be disruptive to cooperation as it creates an environment for shirkers to flourish. Shirkers exploit shared reputations by using the reputation to gain donations without making any donations themselves, which is invoked by an agent i having a defective strategy (i.e., $s_i = u_i = d_i = 0$). By experimenting with different sharing structures, in terms of the extent to which traits are shared, we can assess how much cooperation can be sustained before shirking becomes a dominant strategy and disrupts the evolution of cooperation. Note that when reputation is carried by traits, and individuals carry traits, there are many different ways in which identity can be shared. The number of possible allocations of traits to agents is a combinatorial problem, and therefore it is prudent to use a strategy to explore different types of sharing.

Chapter 3 introduced a framework that enables the simulation of agents sharing reputations through traits within an indirect reciprocity scenario. The approach is based on the idea that agents can have traits that may be shared with others rather than the convention of traits held in isolation. Sharing can be established in various ways. In this chapter, we structure the different types of trait sharing that can take place, and we assess the effects of this on cooperation. This method enables us to assess how much traits can be shared before cooperation becomes infiltrated by shirkers and collapses. It also allows us to investigate the impact of different types of sharing.

The chapter is structured into subsections as follows. In Section 4.2, we introduce characterisation of sharing using definitions that define alternative types of sharing of traits. In Section 4.3, we highlight the relationship of our model to the concept of stereotyping in psychology by showing how stereotyping can be modelled through reputation. Agents share reputation by sharing of traits.

In Section 4.4, we present experimentation in which a single trait is shared with *multiple* other agents to address how sensitive cooperation is to the number of agents that share a single trait. In Subsection 4.4.1, we investigate how different structures of sharing can impact cooperation by organising agents into smaller clusters (groupings) while sharing the same number of traits.

While the previous sections focus on agents sharing single traits, Sections 4.5 and 4.5.1 focus on agents holding multiple traits in common. In Section 4.5, we show how introducing a single multi-trait agent can impact cooperation when sharing traits with single-trait agents. In Subsection 4.5.1, we introduce a second multi-trait agent to understand the effect of having multiple multi-trait agents. Finally, in Section 4.6, we discuss the implication of these results and their impact on cooperation.

4.2 Characterising the Sharing of Identity

In Chapter 3, we introduced our model of cooperation that adopts an indirect reciprocity model with the use of traits to address identity that is composed from multiple sources.

The model provides the option to implement many different types of reputation sharing, where reputation is carried by each of the traits. As the number of ways that agents may share traits is a combinatorial problem, an approach is needed to characterise different types of trait sharing. To achieve this we introduce the following definitions which capture how dependent agents are on each other in deriving their reputation; these definitions follow those introduced in Section 3.4.1. Additionally we introduce key parameters that are widely used in this chapter in Table 4.1.

Definition 7. For a trait t, the set of all agents that include t as a trait in their reputation is referred to as *group* t and denoted as $G_t = \{i \in A : t \in T_i\}$.

Definition 8. The set of dependent agents is denoted $N' = \{i : \exists j \in A - \{i\}\, with\ T_i \cap T_i \neq \emptyset\}$.

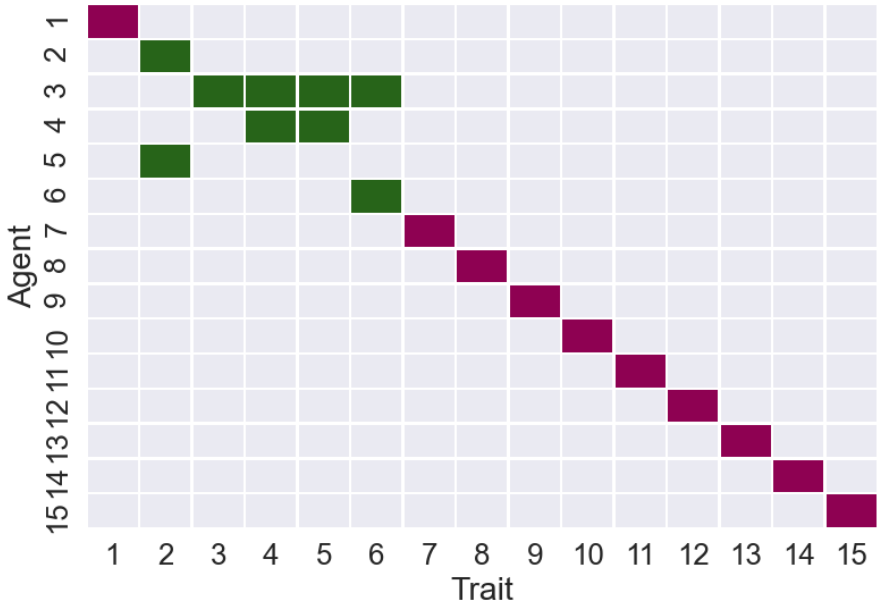


Figure 4.1: A simple visualisation of 15 agents showing the different ways that we can structure the sharing of traits among agents. Agent 1 is an independent single-trait agent. Agents 2 and 5 are single-trait dependent agents that have one trait which they share. Agent 3, which has multiple traits, shares two traits with agent 4 and and a single trait with agent 6, making agent 3 dependent. The remaining agents are independent.

In Figure 4.1, agents 1, 2, 5 and 6-15 are all single-trait agents. Note that if a single-trait agent is also dependent, then they present a risk to cooperation if they are a shirker.

Agents 3 and 4 are multi-trait agents. Multi-trait agents may also pose a threat to cooperation if they are a dependent agent that holds multiple shared traits. The green cells in the figure refer to dependent agents, and the red cells refer to the independent agents.

We note that potential cooperation may emerge with shirkers present if the population sustains a sufficient number of agents that maintain donation behaviour. However, the extent to which this is possible is not known, and how it relates to the number of traits held by a multi-trait agent is also unknown. To investigate this, for the purposes of this chapter we assume that an agent's traits remain fixed throughout the experiment, i.e. agents cannot change the traits that they have been assigned. This enables us to understand the potential for cooperation to be sustained in fixed groups, as defined in Definition 7. We run each experiment using the default parameters mentioned in Section 3.5.3. The experimental results are obtained from five runs, each with a random seed. Each agent participates in an average of 50 games per generation based on $A = 100$, $M = 100000$ and $m = 5000$.

Our experiments can be classified into two categories: firstly the effect of dependent single-trait agents on the evolution of cooperation (Section 4.4); secondly the effect of a dependent multi-trait agent on the evolution of cooperation (Section 4.5). In Section 4.4 we consider the number of agents able to hold a single trait between them before cooperation collapses. We also consider the effects of agents being divided into smaller sharing groups (Subsection 4.4.1) which allows more traits to be shared but by a smaller number of agents. In Section 4.5 we consider the effect of a single agent sharing multiple traits in the presence of single-trait agents on the evolution of cooperation. Finally in Section 4.5.1 we explore the case of multiple agents sharing multiple traits.

Parameter(s)	**Description**	**Role in model**	**Reference in chapter**
G_t	The set of agents that share trait t	Defines a group of agents sharing the same trait(s).	Section 4.4
T_i	The set of traits that belong to agent i	Agent reputation is built upon the traits that they have.	Section 4.4
N'	The set of dependent agents	Used to determine the number of agents that share at least one trait.	Subsection 4.4.1

Table 4.1: The Sharing Parameters Used in Chapters 4 to 7.

4.3 Stereotyping

As our model represents agents' reputation through shared traits, we note that it relates to concepts surrounding identity more widely. This section presents how our model provides a way of modelling *stereotyping* in a computational form and the relationship between stereotyping and cooperation.

As seen in Section 2.5.1, reputation is required for cooperation to emerge within an indirect reciprocity scenario. Within our model, traits are used as the basis for an individual's identity, and traits carry reputations in their own right. When agents share traits, they form groups based on features that they share or characteristics that they have, which allows agents to share reputations. Similarly, in psychology *social categorisation* refers to individuals being grouped based on features, traits, that they have regardless of their actions [183, 187]. Social categorisation naturally leads to stereotyping as it allows individuals to take short cuts when judging others, this can be useful when individuals are unable to recognise others based on their individuality but can identify them through their groups [74, 193, 194]. For example, teachers may not be familiar with all the students in their school but can recognise a student through their school uniform.

In our model traits are used as a proxy for indirectly assessing an individual's reputation, allowing for stereotyping to take place, thus allowing agents to disconnect their actions from their reputation. In the context of cooperation, this means that elements of an individual's reputation becomes dependent on the donation behaviour of others. In turn, this allows agents to deploy defective strategies: that is an agent can avoid paying the full costs of donation but receive donations based on the reputation aligning with its associated traits. Therefore, shirkers exploit reputations through stereotyping by being associated with other agents.

The rest of this chapter examines how both repeated sharing of the same trait, and sharing across multiple traits, affects the emergence of cooperation. The mechanism provides an assessment of the costs associated with shirking, in terms of the effect on cooperation. Specifically, we explore how shirking can exploit cooperation and consider different sharing structures that limit the effects of shirking while allowing cooperation to emerge.

4.4 Agents Sharing a Single Trait

In this section, we consider the effects of sharing a single trait in a set of single-trait agents. This experiment allows us to determine the number of agents that can share a single-trait while sustaining cooperation. We begin by assigning G_1 as the set of all agents i having $T_i = \{1\}$. Figure 4.2 presents an example schema for this arrangement of traits. Note that if all agents are single-trait and independent, their reputation is based entirely on their own past interactions and the results in [199] are replicated. At the other extreme, if all agents are dependent and share a single trait, then agents are (almost) entirely judged on the actions of others, and a greater incentive to defect can be expected. Additionally, in this experiment we vary parameter S (Section 3.5.1) to determine whether limiting the interaction of G_1's agents in playing the donation game with each other would present an effect.

Scenario 1. Input k, the desired size of G_1. k agents share a single trait and the $|A| - k$ remaining agents each have a single unique trait:

$$\begin{aligned} T_i &= \{1\} && \text{for } 1 \leq i \leq k \\ T_{k+1} &= \{2\} \\ T_{k+2} &= \{3\} \\ &\vdots \\ T_N &= \{|A| - k + 1\} \end{aligned}$$

Each agent is assigned an action rule at random from the eight possible.

Figure 4.3 presents the results of increasing the size of G_1 and varying parameter S. The size of G_1, in this case, is equal to $|N'|$. We use four different sizes to determine how many agents can share a trait before cooperation collapses these are $|N'| \in$

$\{10, 15, 20, 30\}$. Two patterns emerge: firstly cooperation declines rapidly when at least 15 dependent single-trait agents share a common trait. Secondly, the average cooperation declines as S increases.

This occurs because as dependent single-trait agents share their reputation with each other, they lack a distinguishable personal reputation. This disadvantages individuals who bear donation costs alone while the reputational benefits of donation are necessarily shared with others in the group G_1. This stereotyping effect provides an opportunity for defective strategies to take hold, where free-riders can benefit from enjoying a shared reputation without donating. However, this cannot be sustained at scale (e.g., beyond 15 agents), leading to the global collapse of cooperation, because *the increase in reputation of a shared trait results in greater opportunity for exploitation by free-riders.*

Recall from Section 3.5.3 that agents copy action rules based on their relative success as measured by f_i. In this scenario, when shirkers exploit reputations and incur no costs, they, in turn, gain a high f_i, resulting in a higher number of shirkers in the population. Figures 4.4, 4.5 and 4.6 highlight the key action rules (defection strategy $(0, 0, 0)$ and discrimination strategy $(1, 1, 0)$) that occur within populations considered as subsequent, but not consecutive, generations when $|G_1| = 10$. Figures 4.4, 4.5 and 4.6 present three different values of S $(0, 0.5, 1)$. It is known [199] that the discrimination strategy dominates when all agents carry their own unique reputation. *Prioritising interaction with those who share the same trait (i.e., high S) accelerates the collapse of cooperation further as the discriminative strategy directs donations towards agents with a similar reputation.* When S is low, dependent single-trait agents interact mainly with those who do not share their reputation as they are still incentivised to adopt cooperative strategies to maximise their fitness with a reduced risk of exploitation.

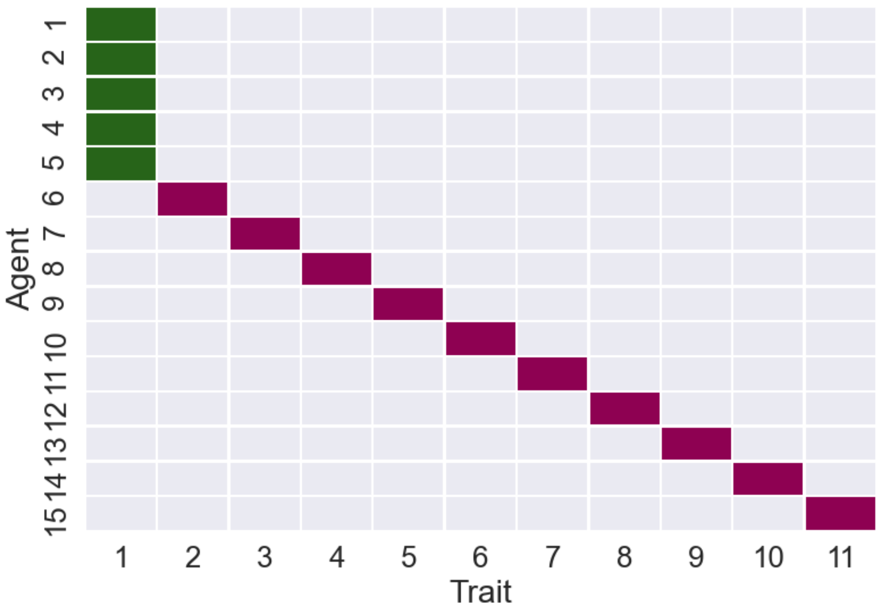


Figure 4.2: A simple visualisation of 15 agents showing single-trait agents sharing a single trait. In this example, five dependent agents share a single trait, i.e. $|G_t| = 5$ and $t = 1$. The green cells refer to dependent agents, and red cells refer to independent agents.

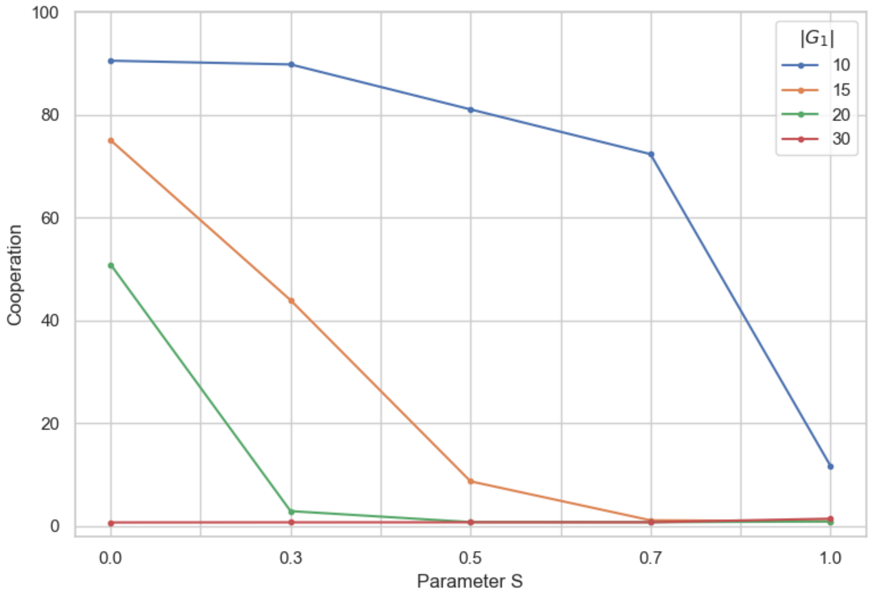


Figure 4.3: The average cooperation recorded as a result of agents sharing a single trait with different sizes of G_1 and with different implementation of parameter S.

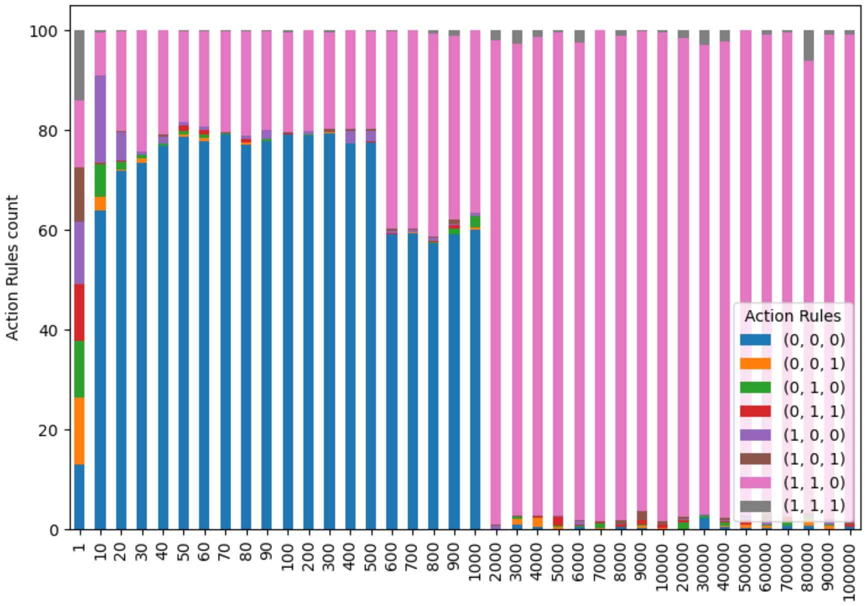


Figure 4.4: An example of the distribution of action rules by subsequent but not consecutive generations for the set of single-trait dependent agents G_1 with $|G_1| = 10$ and parameter S = 0 shows dominance by discriminators $(1, 1, 0)$ after the first 1000 generations.

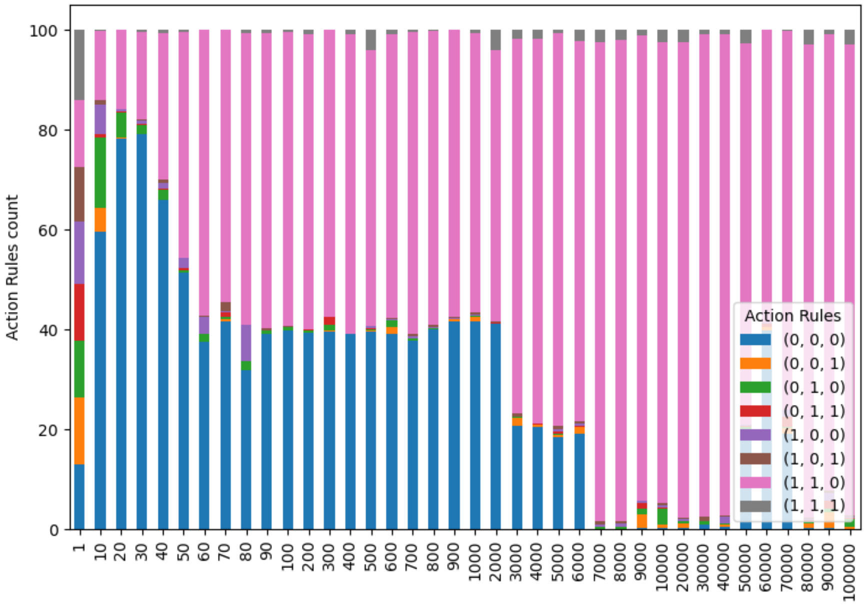


Figure 4.5: A snapshot of the distribution of action rules in subsequent but not consecutive generations for the sets of single-trait dependent agents G_1 with $|G_1| = 10$ and parameter S = 0.5 shows that discriminators $(1, 1, 0)$ dominate the population but only after 6000 generations.

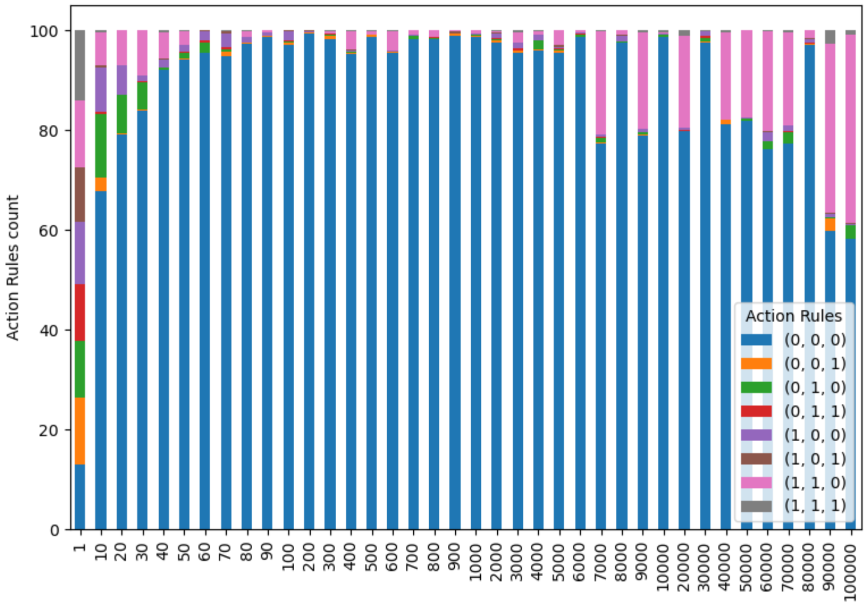


Figure 4.6: A snapshot of the distribution of action rules in specific subsequent but not consecutive generations for the sets of single-trait dependent agents G_1 with $|G_1| = 10$ and parameter S = 1.0 shows that defectors quickly dominate the population within the first 10 generations only as they are able to exploit sharing a reputation with other agents.

4.4.1 Multiple Sharing Groups

In this section, we consider the effects of dividing agents into smaller sharing groups as opposed to a single group. The results from Section 4.4 show that cooperation could be sustained when the number of agents in $|G_1|$ is small. This section considers the effect of sharing structures on cooperation by employing groups of equal size while maintaining the number of agents that share a trait from Section 4.4. For example, using this approach, we can consider the difference between a single group of 10 and two groups of five members or five groups of two members.

Recall that agents are referred to as either independent or dependent (see Definitions 1 and 2). Here we assign dependant agents to one of several groups in which they share a single trait with the other members of the group. Each group has the same size. In this experiment, two sizes are considered: groups of size 2 and groups of size 5. Therefore if we have ten individuals sharing reputations, $|N'| = 10$, they are divided into 5 groups of size 2 or 2 groups of size 5. Figures 4.7 and 4.8 visualises the distribution of agents into the groups. The total number of dependent agents is kept the same as in Section 4.4, i.e. $|N'| \in \{10, 15, 20, 30\}$[1].

Two key findings can be observed from the results in Figures 4.9, 4.10, 4.11, and 4.12. Firstly, when $|G_t|$ = 2, as in Figure 4.7, cooperation is maintained at a higher rate than when $|G_t|$ = 5, as in Figure 4.8. Interestingly, both cases yield better cooperation than when multiple agents share a single trait (Section 4.4). However, in Figure 4.12 both $|G_t|$ = 5 and the single group of agents sharing a single trait yield a cooperation below 1%. In other words, when traits are shared between a smaller number of agents, cooperation is sustained at a higher level. The results further reveal that when parameter S rises towards $S = 1$, cooperation decreases. *The results indicate that higher cooperation is obtained when the number of individuals sharing traits is small and when individuals interact outside of their shared trait.*

Scenario 2. Input k, the total number of dependent agents and g, the number of dependent groups. $n = \lfloor \frac{k}{g} \rfloor$ agents share each of the first g traits, and the $|A| - gn$ remaining agents each have a single unique trait:

$$\begin{aligned}
T_i &= \{1\} && \text{for } 1 \leq i \leq n \\
T_i &= \{2\} && \text{for } n+1 \leq i \leq 2n \\
&\vdots && \\
T_i &= \{g\} && \text{for } (g-1)n+1 \leq i \leq gn \\
T_{gn+1} &= \{g+1\} && \\
T_{gn+2} &= \{g+2\} && \\
&\vdots && \\
T_{|A|} &= \{g+(|A|-gn)\} &&
\end{aligned}$$

Each agent is assigned an action rule at random from the eight possible.

[1] Note in the case of $|G_t| = 2$ the range was $|N'| \in \{10, 14, 20, 30\}$

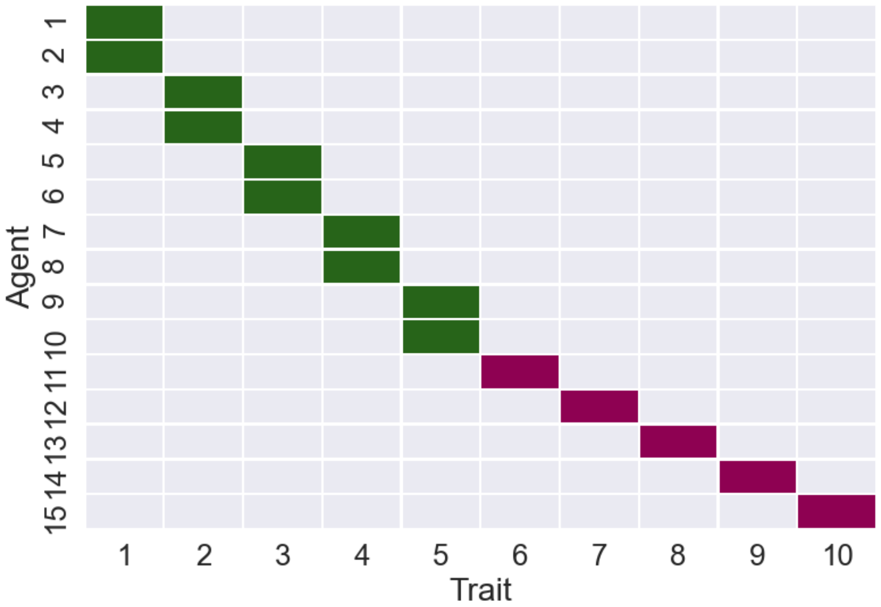


Figure 4.7: A simple visualisation of 15 single-trait agents in which traits are shared by groups of two. In this example, ten dependent single-trait agents, $|N'| = 10$, are divided into groups of two $|G_t| = 2$, where $t \in \{1, 2, 3, 4, 5\}$. The green cells refer to dependent agents, and red cells refer to independent agents.

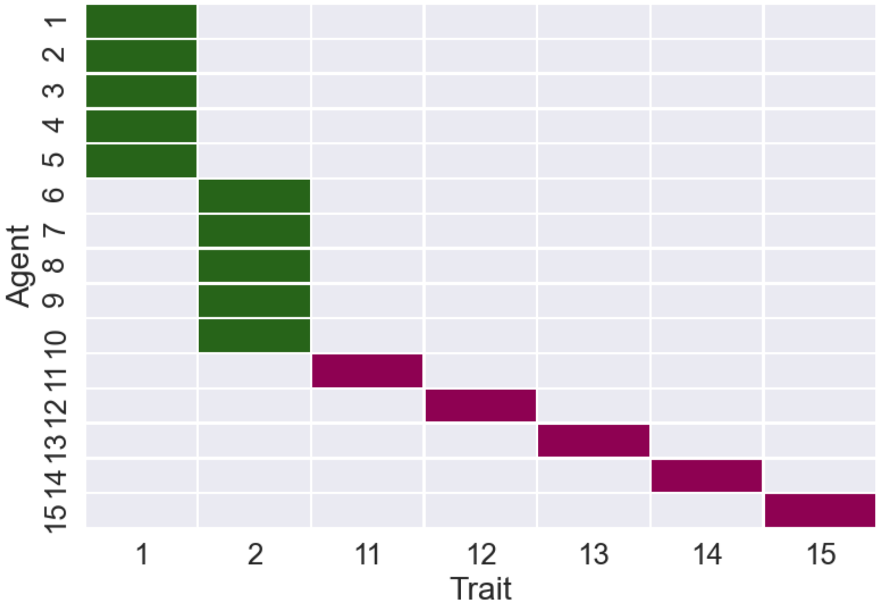


Figure 4.8: A simple visualisation of 15 single-trait agents in which traits are shared by groups of five. In this example, ten dependent single-trait agents, $|N'| = 10$, are divided into groups of five $|G_t| = 5$, where $t \in \{1, 2\}$. The green cells refer to dependent agents, and red cells refer to independent agents.

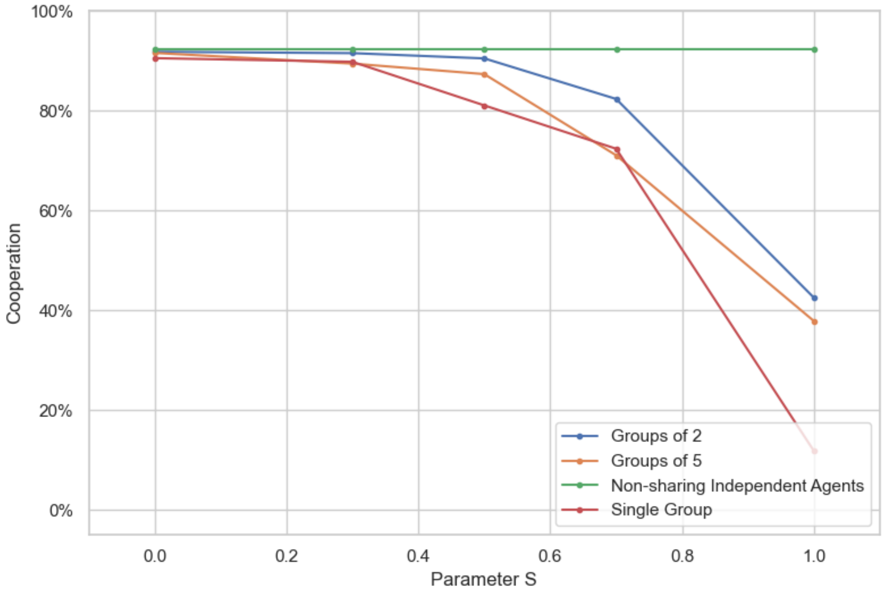


Figure 4.9: The average cooperation recorded as a result of 10 single-trait agents sharing a trait in different sharing groups using different values for parameter S. Cooperation decreases as parameter S gets closer to 1. The decrease in cooperation is reduced as the number of agents sharing the same trait decreases.

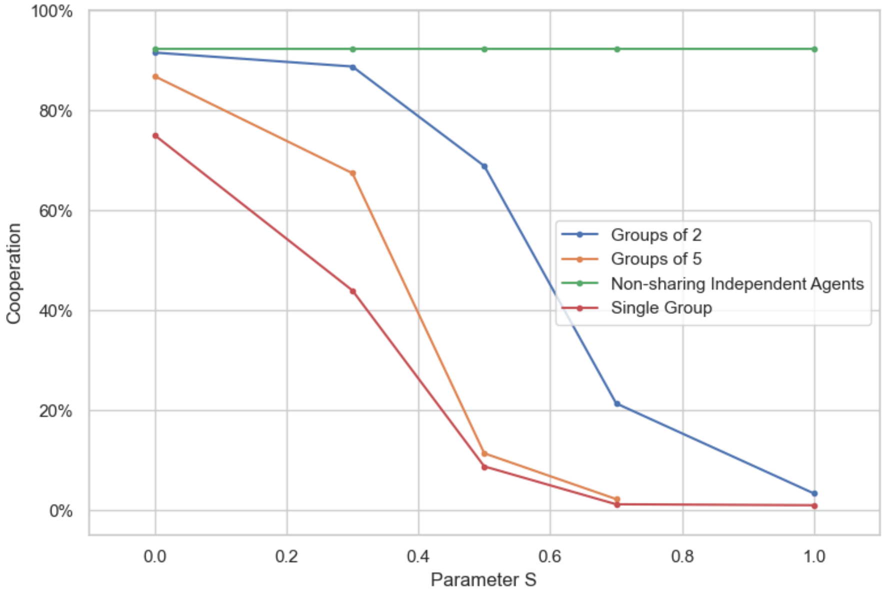


Figure 4.10: The average cooperation recorded as a result of 15 single-trait agents sharing a trait in different sharing groups using different values for parameter S. Cooperation decreases as parameter S gets closer to 1. The decrease in cooperation is reduced as the number of agents sharing the same trait decreases.

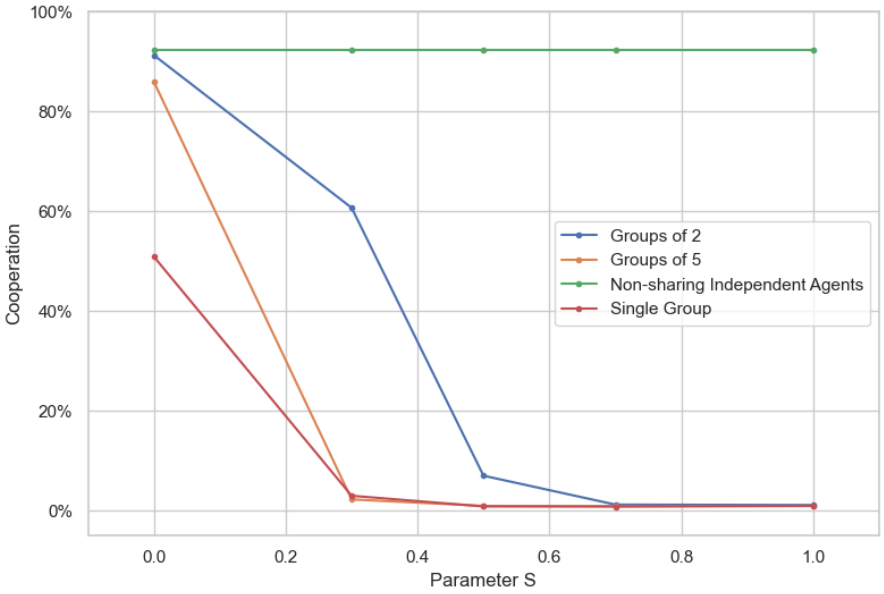


Figure 4.11: The average cooperation recorded as a result of 20 single-trait agents sharing a trait in different sharing groups using different values for parameter S. Cooperation decreases as parameter S gets closer to 1. The decrease in cooperation is reduced as the number of agents sharing the same trait decreases.

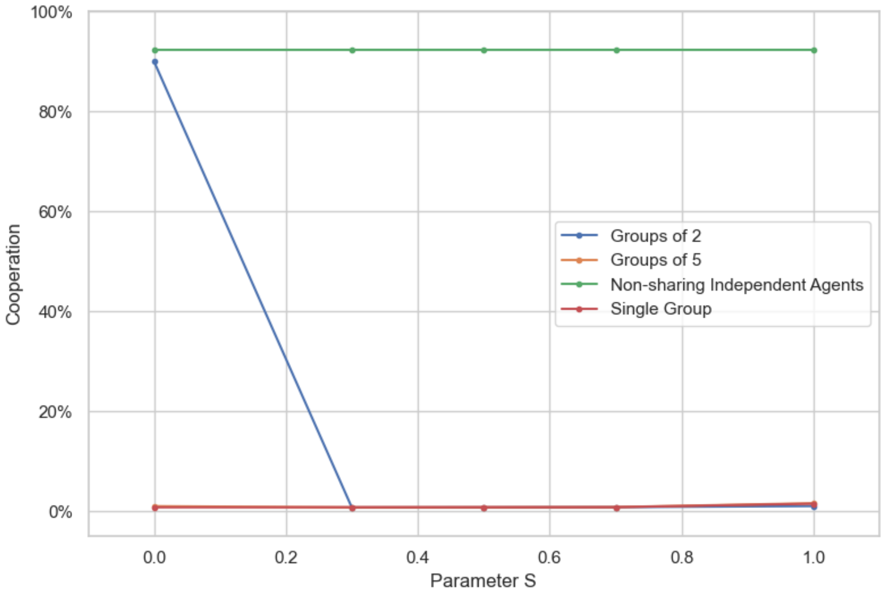


Figure 4.12: The average cooperation recorded as a result of 30 single-trait agents sharing a trait in different sharing groups using different values for parameter S. Cooperation decreases as parameter S gets closer to 1. The decrease in cooperation is reduced as the number of agents sharing the same trait decreases. Note that both the groups of 5 and the single group report under 1% cooperation, additionally when parameter $S \geq 0.3$, groups of 2 also record cooperation below 1%.

Similar to Section 4.4, the results show that cooperation emerges when group sizes are small (15 or less), and the number of individuals sharing traits is limited, i.e. the smaller $|N'|$ and $|G_t|$ are, the higher the cooperation. Next, we observe that when individuals are encouraged to interact with agents outside of their shared trait group, i.e. when $S = 0$, cooperation increases for the entire population. The observation holds true for all tested cases $|N'|$ = $10, 15, 20, 30$, as interactions get limited to only those agents who share a common trait, i.e. the closer S is to 1, the lower the cooperation. Finally, we noticed that as groups split, cooperation begins to evolve as can be seen in Figures 4.9, 4.10, 4.11, and 4.12.

There are several reasons why smaller groups enable the evolution of cooperation. Small groups have a limited number of individuals; this limits the number of different strategies (social comparison heuristics) that individuals carry. The discriminator strategy always dominates a cooperative population due to their ability to choose which individuals to donate to, which helps to eliminate defector strategies from the population. The reputation of a group depends on its members' actions. If all members of a group fail to donate, their reputation rapidly declines. When the reputation of a group is low, other individuals in the population stop donating to the group, leaving its members vulnerable as they are unable to build a fitness level, f, to propagate in future generations. This trend allows individuals to quickly identify defectors within groups but only in the case that the group has a small number of individuals, in comparison to the population, resulting in a rapid decline of the group's reputation. The trend limits the exploitation of reputation by defectors within a group that has cooperative members.

In contrast, larger groups are unable to sustain cooperation as defectors successfully take advantage of the group. Once a group maintains a high reputation, defectors begin to exploit it for their own gain. In order for defectors to exploit a reputation, a group must have a mix of strategies within its population. If a group consisted of only one strategy, then the strategy becomes visible, especially within smaller groups. Smaller groups are more transparent than bigger groups, where the strategies held by group members can be recognised quickly. Therefore if defectors wanted to exploit a group, they could take advantage of a bigger group far more easily than would be the case with smaller groups. Additionally, when individuals only play the game within their in-group, defectors exploit the fact that everyone would have the same reputation and rapidly increase their fitness (payoff).

4.5 Agents With Multiple Shared Traits

In this section, we consider the effects of introducing a single dependent multi-trait agent in a population of single-trait agents. By increasing the number of traits that a single agent shares with multiple agents, we can consider a more complex structure of sharing beyond the ones presented in Sections 4.4 and 4.4.1. Figure 4.13 presents the schema for this arrangement. In this section, $|T_1|$, see Table 4.1, is varied in a range between $(2, 100)$ allowing us to consider sharing traits between agent 1 and the whole population.

Scenario 3. Input k, the total number of dependent agents. Agent 1 holds the first k traits, the next $k - 1$ agents each have a single trait in common with agent 1 (and no others), and the $|A| - k$ remaining agents each have a single unique trait:

$$\begin{aligned} T_1 &= \{1, \ldots, k\} \\ T_i &= \{i\} \qquad \text{for } 2 \leq i \leq |A| \end{aligned}$$

Each agent is assigned an action rule at random from the eight possible.

The results (Figure 4.14) show that *as the number of traits held by agent 1 increases (i.e., $|T_1|$ increases), cooperation diminishes*. The trend occurs regardless of limiting interactions by parameter S, i.e., whether agent 1 interacts with those who have at least one trait in common. The reputation of the sharing agent, agent 1, is dispersed across dependent single-trait agents that between themselves have no trait in common. The dispersion of reputation helps to suppress the rise of defective action rules, as compared to the previous scenario (Section 4.4). In fact, $|T_1|$ can reach a considerable size (e.g., 30-45 traits) before cooperation starts to diminish significantly, i.e., the number of sharing agents is relatively higher than that of Section 4.4. Figure 4.15 shows how defective action rules are suppressed after dominating in the first 3000 generations as agent 1 holds 40 shared traits. Figure 4.14 shows a slight increase when $|T_1| = 40$, as a result of genetic drift, the fluctuation could be eliminated if the number of runs is increased from five.

In this scenario, single-trait dependent agents rely entirely on themselves and the multi-trait agent for their reputation. Each single-trait dependent agent can also free ride on the single multi-trait agent, and this opens the opportunity for defection to establish itself, although to a lesser extent than the case presented in Section 4.4. When the number of traits of the multi-trait agent is relatively small, the presence of free-riding dependent single-trait agents can be sustained without too much disruption to the reputation of the multi-trait agent. As $|T_1|$ increases, and the number of dependent single-trait agents increases, there is a greater opportunity for free-riding action rules to take hold (e.g., $(s_i, u_i, d_i) = (0, 0, 0)$). Figure 4.16 illustrates this trend as agent 1 holds 70 traits, and the defector action rule dominates the population.

As the number of dependent single-trait agents increases, the number of independent single-trait agents diminishes. This trend promotes the collapse of cooperation. As soon as a defective strategy takes hold across the population, it presents an opportunity for defective strategies to spread to other agents. Interestingly, S has relatively little impact on whether dependent agents prioritise interacting with those that have a common trait. However, they are less likely to have an equal reputation in this instance.

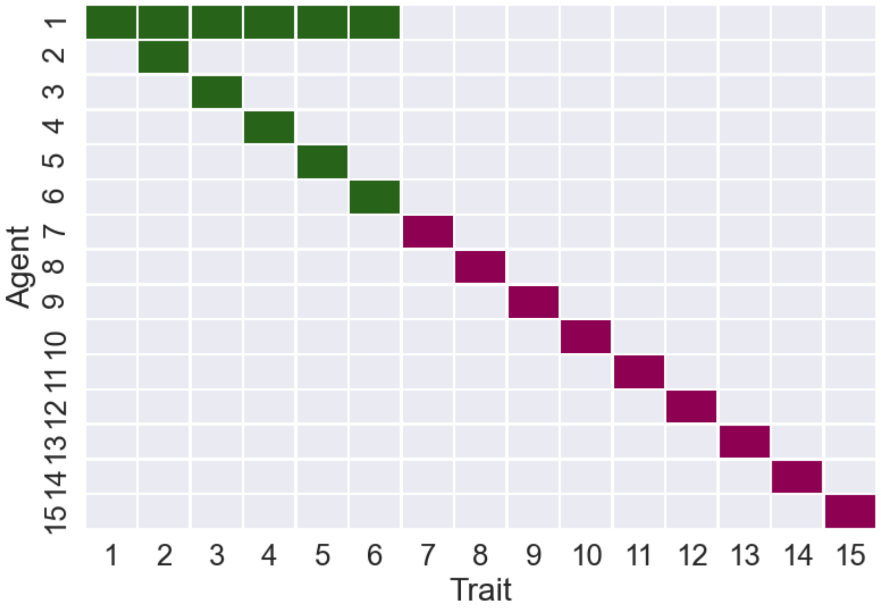


Figure 4.13: A simple visualisation of 15 agents, in which a multi-trait dependent agent shares traits with single-trait agents where $|N|' = 6$. The green cells refer to dependent agents, and the red cells refer to independent agents.

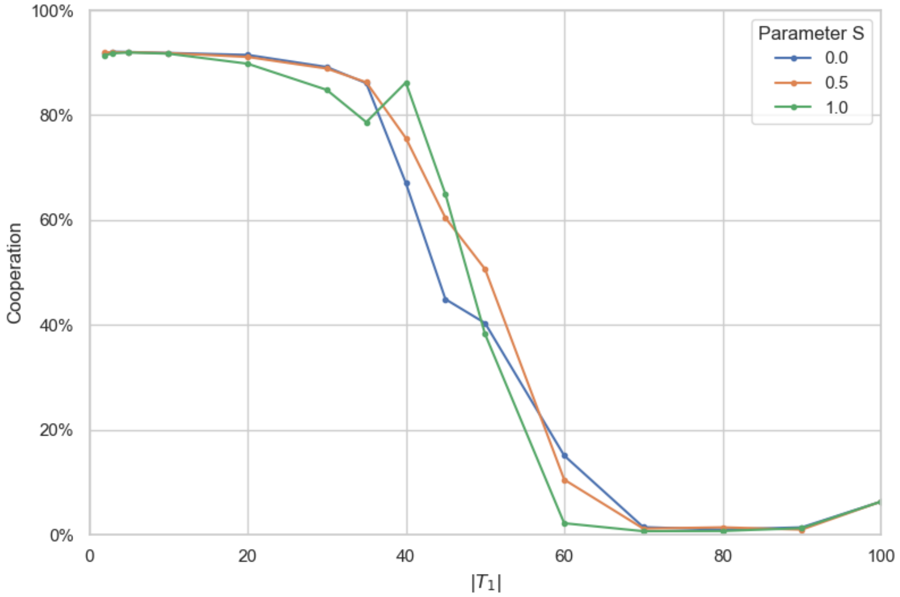


Figure 4.14: The average cooperation recorded as a result of a multi-trait agent, agent 1, sharing traits with single-trait agents using three different values for parameter S. The number of traits being shared by agent 1 is varied between 2 and 100 traits.

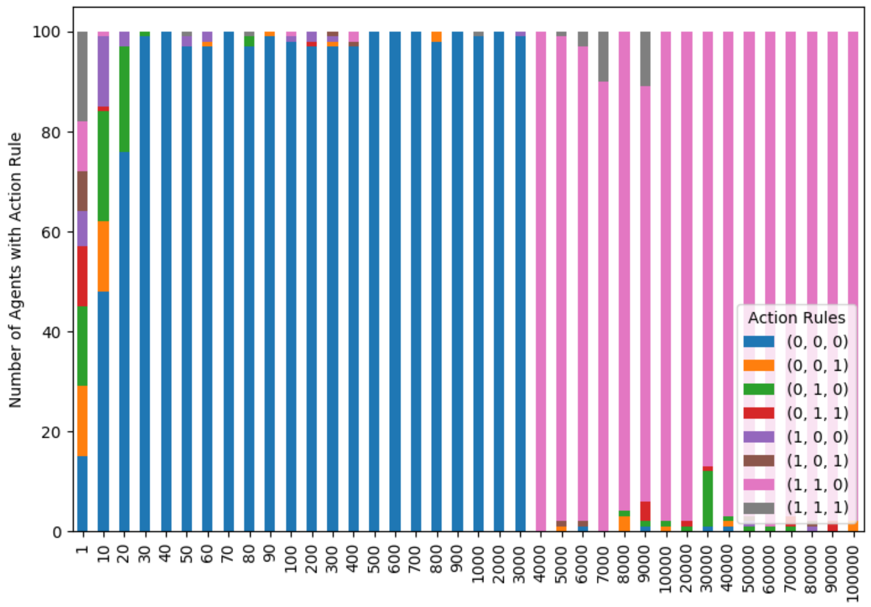


Figure 4.15: A snapshot of the distribution of action rules in subsequent but not consecutive generations when multi-trait agent 1 has 40 traits shared with single-trait agents and parameter $S = 0.5$.

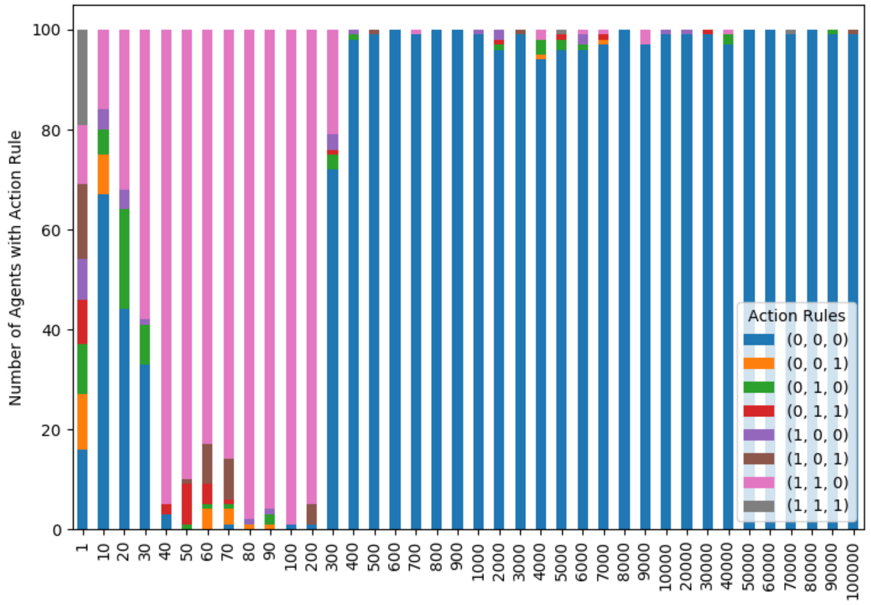


Figure 4.16: A snapshot of the distribution of action rules in subsequent but not consecutive generations when multi-trait agent 1 has 70 traits shared with single-trait agents and parameter $S = 0.5$.

4.5.1 Multiple Agents With Multiple Traits

In this subsection, a second multi-trait dependent agent is introduced to the population. The experiment examines what effect does the addition of a second multi-trait agent have on cooperation and specifically what happens when it overlaps with the first agent such as $T_2 \subseteq T_1$. Figure 4.17 shows a simple schema of this scenario. To achieve the overlap effect, we vary T_2 while fixing the number of traits in T_1 to examine how flexible this structure of sharing is before cooperation collapses. Three set of traits were investigated, $|T_1| = 25, 35, 50$. Note that in this experiment parameter $S = 0$, i.e., agents are not limited in their interactions to only those sharing a trait with them. Parameter $S = 0$ was chosen as the results in Section 4.5, showed that parameter S has no significance on the outcome of cooperation.

Figure 4.18 shows the effect of varying $|T_2 \cap T_1|$, that is the extent to which T_2 has the same traits as T_1. The results in the figure show that high proportions of shared traits through multi-trait agents undermine the reputation system. Because the second multi-trait agent can hold a large subset of the first agent's traits, it can heavily disrupt the first agent's reputation, by using defection as its action rule. The results are consistent with Figure 4.14 (Section 4.5) because as the percentage of overlap between agents 1 and 2 increases, cooperation correspondingly decreases. However, the results show that introducing the second agent decreases cooperation even further in all three tested cases. Note that in the case of $|T_1| = 25$ the decrease is not as rapid as $|T_1| = 35$ and $|T_1| = 50$. In the cases of $|T_1| = 25and35$, the decrease in cooperation is not monotonic as a result of only having five runs. Once the number of runs is increased, the fluctuation would lessen. This result indicates that the number of traits being shared is significant. The increase in the overlap between agents 1 and 2 is a factor here for the decrease in cooperation. The decline in cooperation is more pronounced than that of a dependent single-trait agent sharing reputation with the multi-trait dependent agent, and increases as $|T_2 \cap T_1|$ increases.

Scenario 4. Input k, the total number of dependent agents, and λ, representing the overlap in traits between agent 1 and 2. Agent 1 holds the first k traits, agent 2 holds the first $\lfloor \lambda k \rfloor$ traits, the next $k-2$ agents have a single trait in common with agent 1 and no traits in common with agents $3, 4, \ldots, |A|$, and the $|A| - k$ remaining agents each have a single unique trait:

$$\begin{aligned} T_1 &= \{1, \ldots, k\} \\ T_2 &= \{1, \ldots, \lfloor \lambda k \rfloor\} \\ T_i &= \{i\} \qquad \text{for } 3 \le i \le |A| \end{aligned}$$

Each agent is assigned an action rule at random from the eight possible.

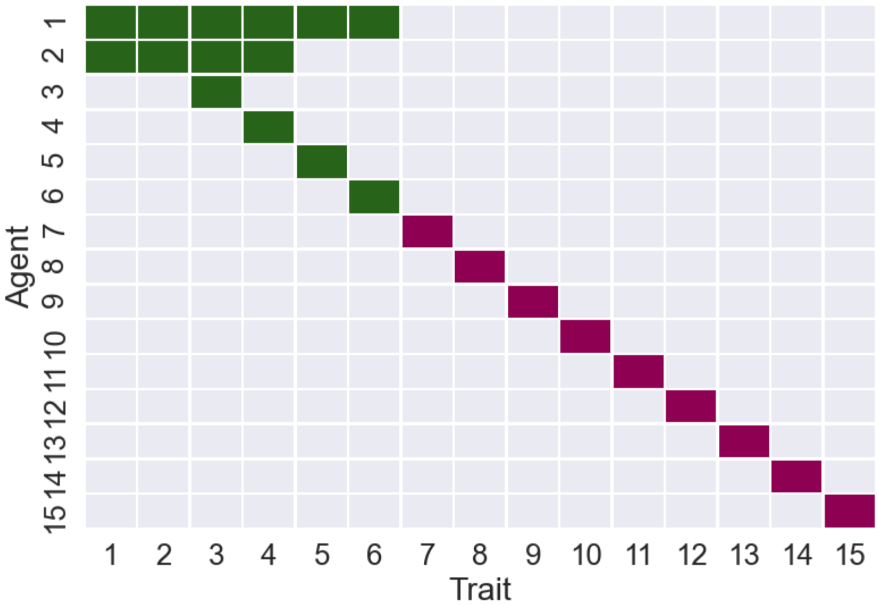


Figure 4.17: A simple visualisation of 15 agents, in which two multi-trait dependent agent share traits with single-trait agents. The green cells refer to dependent agents, and the red cells refer to independent agents.

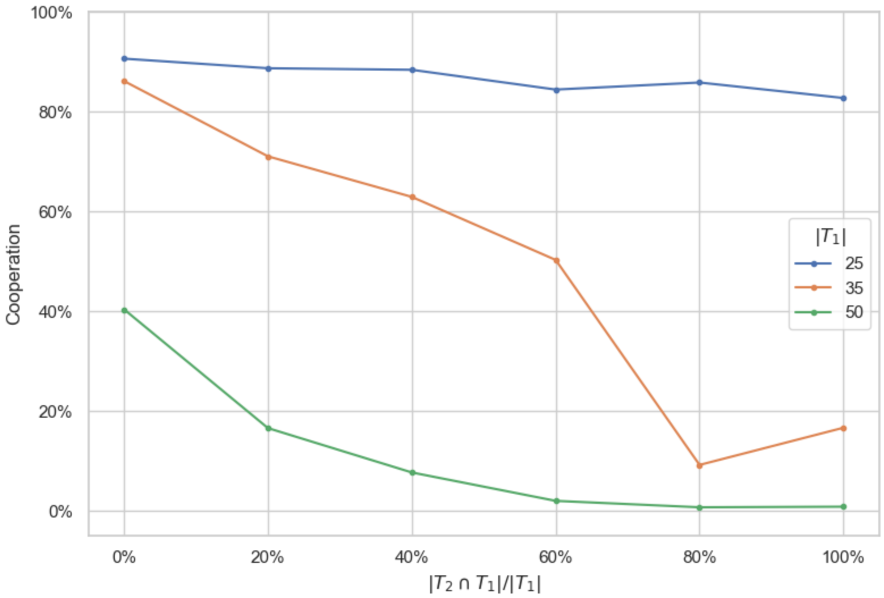


Figure 4.18: The average cooperation produced as a function of the size of the intersection between the sets belonging to multi-trait agents one and two for different values of $|T_1|$ where $S = 0$.

4.6 Discussion

The results in this chapter indicate that the sharing of reputation through common traits significantly disrupts reputation systems for cooperation. By using traits as proxies for indirectly assessing an individual's reputation, an opportunity is introduced for agents to disconnect their actions from their reputation. Agents can deploy defective strategies: that is an agent can avoid paying the full costs of donation but receive donations based on the reputation aligning with its associated traits. By exploiting reputations for their own gain, these agents spread their strategy by having a higher f_i than the rest of the population. This spread eventually causes cooperation to collapse.

Single-trait and multi-trait agents differentiate how other agents can share their traits. Single-trait agents have a reduced chance of others having a trait in common. However, when another agent shares their trait, their reputation becomes susceptible to the actions of a third party. As single-trait agents only have one trait, their reputation is at risk when shared with just one defective agent, as seen in Subsection 4.4.1. Further sharing increases the risk on their reputation, leading cooperation to collapse eventually. Similarly, when agents prioritise interaction with those who share the same trait (i.e., high S) the collapse of cooperation is accelerated.

In contrast, for multi-trait agents, increasing the number of traits can give them a chance to retain an element of unique personal identity, through traits that are not shared with others. However, cooperation does diminish as the number of traits held by the multi-trait agent increases. Moreover, for multi-trait agents, sharing can occur with many agents that have no dependency between them, in terms of common traits. On the other hand, parameter S did not show any influence on cooperation.

The results show that reasonable levels of cooperation can be sustained while there is a modest level of sharing of identity in the population, after which cooperation collapses. For example, a single trait being shared by 15 agents yields cooperation at about 70% and a single agent sharing traits with 40 other agents yields cooperation above 60%. These results highlight the importance of individual versus group identity in reputation systems.

4.7 Conclusions

Given the enormous number of possible ways in which traits can be shared, in this chapter, our focus has concerned assessing basic aspects of sharing, surrounding the number of traits held by an agent. The results show that reasonable levels of cooperation can be sustained while there is a modest level of sharing of identity in the population, after which cooperation collapses. The collapse in cooperation is attributed to free-riders who exploit the shared reputation for their own gain. However, their influence can be limited in several ways. Firstly, by changing the structure of sharing either by limiting the number of shared traits or by limiting the number of dependent agents. Secondly, by balancing the number of single-trait agents and multi-trait agents. Thirdly, by enabling dependent agents to interact with independent agents.

Furthermore, our results show that dependent single-trait agents have a strong influence on other agents' reputations as their reputation relies on a single trait. Similarly, the inability of agents to change their traits allows defective strategies to take hold on the population. This restriction on the trait can be deterred by allowing agents to copy other successful traits similar to the way that they copy successful action rules. In the next chapter, we investigate allowing agents to copy traits based on their level of success and its influence on cooperation. This relaxes the constraint that sharing structures for identity are externally fixed, and allows agents the freedom to pursue identity that maximises their payoff.

Chapter 5

Evolution of Identity through Traits and the Impact on Cooperation

5.1 Introduction

In the previous chapter, we assumed that agents held fixed traits, and these did not change when an agent's strategy was updated. In this chapter, we relax this assumption by introducing a new model that investigates different structures for sharing reputation through traits and their degree of impact on cooperation. Specifically, we investigate whether copying traits when agents update their cooperation strategy allows agents to sustain cooperation. Furthermore, we evaluate this approach's effect on combating shirkers.

As mentioned in Chapter 4, there are many ways in which agents can share reputations. In this thesis, agents share reputations through shared traits that represent identities. While the previous chapter allowed agents to update their action rules after every generation, this chapter investigates the impact of allowing agents to also update their traits after every generation. The evolution of traits offers a basis for an agent to change their identity by allowing agents to copy traits based on the relative success of agents who subscribe to that trait as measured by f_i. It also helps us to understand how and why behavioural strategies involving identity, such as whitewashing, function.

The chapter is structured into subsections as follows. In Section 5.2, we introduce the concept of inherited traits and explain how allowing agents to change their identities enables the influence of shirkers to be combated. In Section 5.3, we present experimentation where agents update both their identity and their action rules. This evolution allows agents to change their identity and their behaviour strategy by pursuing the traits and strategies of those deemed most successful. We track how this evolution affects cooperation.

While the previous section, 5.3, allowed agents to update their action rules, in Section 5.4, we limit evolution in an alternative way, by only allowing agents to evolve their traits without the action rules. Finally, in Section 5.5, we discuss the implications of these results and their impact on cooperation.

5.2 The Evolution of Identity

An agent's ability to change traits and pursue the traits of those deemed most successful is referred to here as the evolution of traits. This is an evolutionary form of 'whitewashing', where identity becomes a strategic component that is mutable in pursuit of payoff. Understanding how traits evolve helps us to understand how the structure of traits and the freedom of agents in changing them affects cooperation through indirect reciprocity.

Recall that traits are used as the basis for an individual's identity. Therefore the ability to change traits allows agents to change their identity. The option to change identity leads to opportunities for agents to gain an advantage. Whitewashing is a term that has been used to describe the action of agents who change their identities in order to avoid punishment from other agents [64]. The term has been mostly used to describe this action within peer-to-peer reputation systems where users have been able to replace their pseudonyms to escape from any punishment due to their bad reputation. Whitewashing or re-entry attacks enable free-riders to restore their reputation to gain some short-term payoff [90]. Only limited research has studied the subject within an evolutionary perspective with a view to gathering an understanding of whitewashing in cooperative situations [63]. Whitewashing reduces the opportunity for agents to accumulate a bad reputation and opens up opportunities for defection as a consequence.

In Chapter 4, we considered the evolution of behavioural action rules (strategies) while agent traits did not evolve and remained fixed. In this chapter, we consider the effects of also allowing agents to update their identity by changing their trait during the reproduction phase. To explore this scenario, we assume that all agents i are single-trait agents ($|T_i| = 1$). However, agents can share (and copy) the trait of another agent during reproduction. Specifically, agents can be independent or dependent without limitation, but all agents are single-trait agents. The probability of an agent i changing to the identity of another agent j is proportional to j's payoff relative to the whole population at the end of a generation. In Algorithm 2, we highlight how agents change their action rules and traits relative to the success of the previous generation.

This chapter uses the default parameters of $A = 100$, $M = 100000$ and $m = 5000$, which results in each agent participating in an average of 50 games per generation. At the end of each generation, the reproduction phase takes place. Recall from Section 3.5.3 that reproduction is when agents update their action rules. Similarly, to allow agents to update their traits, agents will probabilistically copy other members' traits based on their relative success as measured by fitness (f_i), i.e. the evolution of traits takes place.

An important step within the reproduction phase is mutation. Mutation changes the action rules or traits for an agent with a small probability. Specifically, after an agent is assigned a new set of action rules and a new trait, based on fitness, mutation is applied. In Chapter 4, we used mutation to change agents' action rules with a probability of 1%. In this chapter, we introduce trait mutation, μ_T, which randomly assigns a trait to agents other than the one that they inherited. Throughout the chapter, we use a trait mutation rate of $\mu_T = \frac{1}{10}$ unless otherwise stated (such as in Section 5.3.3). In this case, mutation randomly changes each agent i's trait into any other trait, $t \in T - \{t_i\}$. The mutation rate of 10% allows agents to move between traits invariably but not too often. At this rate, we found that cooperators have ample time to thrive without defectors taking advantage of the trait's reputation. This finding is consistent with findings by other researchers [71, 186] and is discussed further in Section 5.3 and in Figure 5.8

Algorithm 2 Algorithm for Indirect Reciprocity Based on the Reputation of Traits Where Agent Traits Evolve Using Payoff, as Fitness, and Mutation

```
Require: Number of iterations $m$; Number of generations $M$; set of agents $A$; set of traits
    $T$; set of binary action rules $AR = (s_i, u_i, d_i)$; cost $c$; benefit $b$; in-group interaction
    probability $S$; mutation rate of action rules $\mu_A$; The set of agents that share at least
    one trait with agent $j$ $N_j$; mutation rate of traits $\mu_T$;
for $M$ generations do                                          ▷ Perform evolutionary simulation
    Set $f_i = 0 \ \forall i \in A$ and $r_t = 0 \ \forall t \in T$          ▷ Reset fitness and reputation
    for $m$ iterations do
        $j \leftarrow U(A)$                                     ▷ Select recipient (see Section 3.5.1)
        $p \leftarrow U([0, 1])$
        if $p < S$ and $|N_j| > 0$ then
            $i \leftarrow U(N_j)$                               ▷ Select random in-group donor
        else if $p \geq S$ and $|\bar{N}_j| > 0$ then
            $i \leftarrow U(\bar{N}_j)$                         ▷ Select random out-group donor
        else
            $i \leftarrow U(A - \{j\})$                         ▷ Select random donor
        end if
        ▷ Apply action Rules (see Sections 3.4.3 and 3.5.2)
        if ($r^i$=$r^j$ and $s_i = 1$)                          ▷ Compare equal
              or ($r^i < r^j$ and $u_i = 1$)                    ▷ Compare upwards
              or ($r^i > r^j$ and $d_i = 1$) then               ▷ Compare downwards
            $r_t \leftarrow min(5, r_t + 1)$                    ▷ $i$ donates, increase reputation
            $f_i \leftarrow f_i - c$; $f_j \leftarrow f_j + b$  ▷ Update fitness
        else                                                    ▷ $i$ defects
            if $r^j \geq r^i$ then                              ▷ Detect unjustified defection.
                $r_t \leftarrow max(-5, r_t - 1)$               ▷ Decrease reputation
            end if
        end if
    end for
    ▷ Reproduction stage (see Sections 3.5.3 and 5.2)
    for $i \in A$ do
        $j \leftarrow R(A, f)$                                  ▷ Roulette wheel based on fitness
        $(s'_i, u'_i, d'_i) \leftarrow (s_j, u_j, d_j)$         ▷ $i$ copies $j$'s action rules
        $T'_i \leftarrow T_j$                                   ▷ $i$ copies $j$'s traits
        if $U([0, 1]) < \mu_A$ then
            $(s'_i, u'_i, d'_i) \leftarrow U(AR)$               ▷ Mutate action rules
        end if
        if $U([0, 1]) < \mu_T$ then
            $T'_i \leftarrow U(T)$                              ▷ Mutate traits
        end if
    end for
    for $i \in A$ do
        $(s_i, u_i, d_i) \leftarrow (s'_i, u'_i, d'_i)$         ▷ Update action rules for all agents
        $T_i \leftarrow T'_i$                                   ▷ Update traits for all agents
    end for
end for
```

5.3 The Evolution of Traits

In this section, we consider the effects of evolving agents' traits for a set of single-trait agents on cooperation and the structure of the population. This experiment allows us to determine whether the evolution of traits can sustain cooperation and mitgate against the effect of free-riding. We begin by assigning all agents i as independent single-trait agents with $|T| = |A|$. The completion of a generation initiates the reproduction phase where agents copy the action rules and traits of other agents based on their relative success as measured by f_i (lines 23-38 in Algorithm 2). Therefore, scenarios with different starting configurations are not considered as they have similar outcomes. Additionally, in this experiment, we vary parameter S (Section 3.5.1) to determine the effect of limiting the interaction of agents to their own trait group on cooperation.

The results in Figure 5.1 indicate that for the lowest values of S only limited cooperation is achieved, while it increases with higher rates of in-group interactions. Cooperation achieves an average of 78.06$\%$ when individuals only interact with those having the same trait ($S = 1$). When $S = 0$ cooperation never reaches a level above 10% on average over $100,000$ generations. The results are in contrast with the outcomes obtained where identity remained fixed throughout the simulation, for which increasing the proportion of in-group interactions produced a sharp decrease in cooperative behaviour (see Chapter 4).

In the following subsections, we analyse the results in further detail. In Subsection 5.3.1, we explore why cooperation is higher when interactions are limited to traits ($S = 1$). Similarly in Subsection 5.3.2, we investigate the lower cooperation achieved by agents when they exclusively interact with agents outside of their trait group ($S = 0$). Finally in Subsection 5.3.3, we examine the impact of mutation on traits and how it affects cooperation.

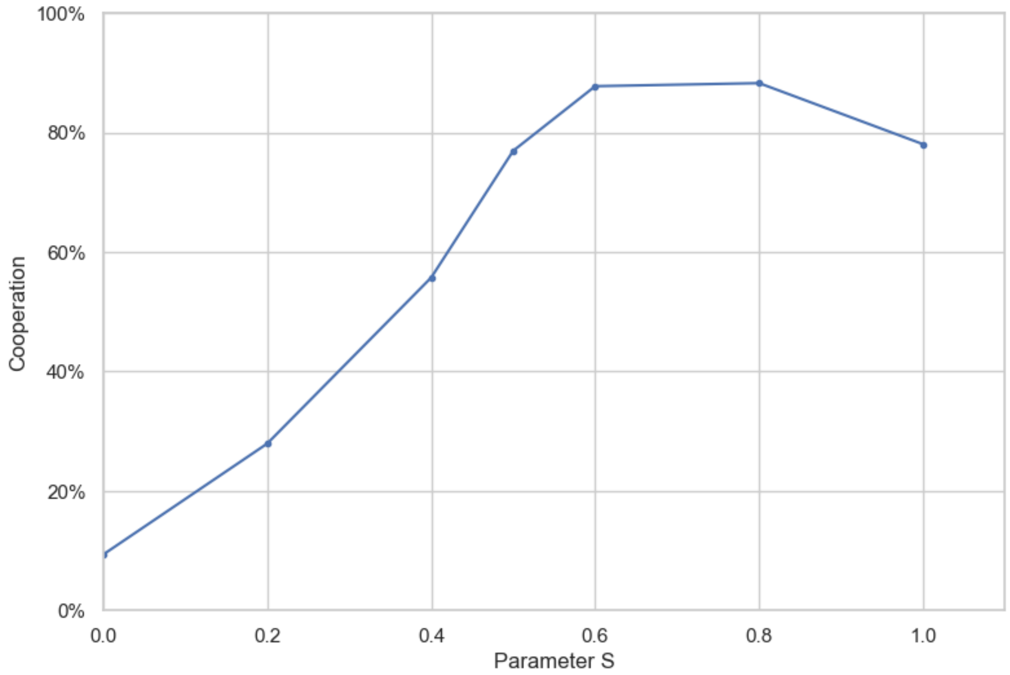


Figure 5.1: The average cooperation recorded as a result of the evolution of traits and a trait mutation of $\mu_T = 10\%$ with different values of parameter S. The result represents an average of five different seeded runs of the same experiment where all agents i are single-trait agents.

5.3.1 Limiting Interactions to Within Traits

In this subsection, we analyse a single run of an experiment where $S = 1$ and traits are allowed to evolve. The analysis allows us to explore the reasons behind the cooperation achieved. In this section, it is assumed that all agents i are single-trait agents ($|T_i| = 1$). Figure 5.1 shows that cooperation achieves an average of 78.06% when individuals only interact with those having the same trait ($S = 1$). To explore why cooperation sustains at such a high level, we need to examine the journey of agents and cooperation throughout each generation in a single experiment. Recall that when $S = 1$, interactions of dependent agents are limited to agents who share their trait. This limit leads to the reputation system becoming redundant because dependent agents only interact with those that have the same trait, and therefore the same reputation. This simplifies an agent's behaviour which becomes entirely dependent on s_i, and u_i and d_i are no longer used because all agents have the same reputation: simply an agent i cooperates if $s_i = 1$, otherwise it defects.

At the beginning of the experiment, each agent i is assigned an independent single trait not shared with others. As agents evolve, their traits may be shared, and the agents become dependent. Figure 5.2 shows that agents consolidate around a small number of traits in each generation. Additionally, the figure shows that a trait can be shared with up to 30 agents before cooperation collapses, which is in line with previous experimentation where cooperation cannot be sustained when several agents share the same trait (see Figure 4.3). The collapse in cooperation is attributed to defectors infiltrating a trait group, that is as a trait expands in size there is a higher chance of defectors joining and exploiting the trait's reputation. The infiltration of traits by defectors leads to agents alternating between the cooperative strategy and the defector strategy. The struggle for dominance between the cooperators ($s_i = 1$) and defectors ($s_i = 0$) is seen in Figure 5.3. Here cooperative agents establish themselves in common traits and are then disrupted by defectors who adopt the same identity before they mutate to a new trait. Figure 5.4 shows that cooperative agents dominate the most shared traits. Therefore, the evolution of traits in the case of $S = 1$ favours cooperators as it allows them to protect their trait from defectors by mutating once a defector agent is identified in the shared trait.

The pattern of agents alternating between strategies in each generation leads to a fluctuating cooperation. Figure 5.5 shows how that cooperation reaches high levels in some generations followed by lower levels of cooperation, leading to an average of above 70%. This result is directly attributed to the agents' strategies of that generation. This is noted specifically on the 9000th generation where defectors dominate the population in Figure 5.3.

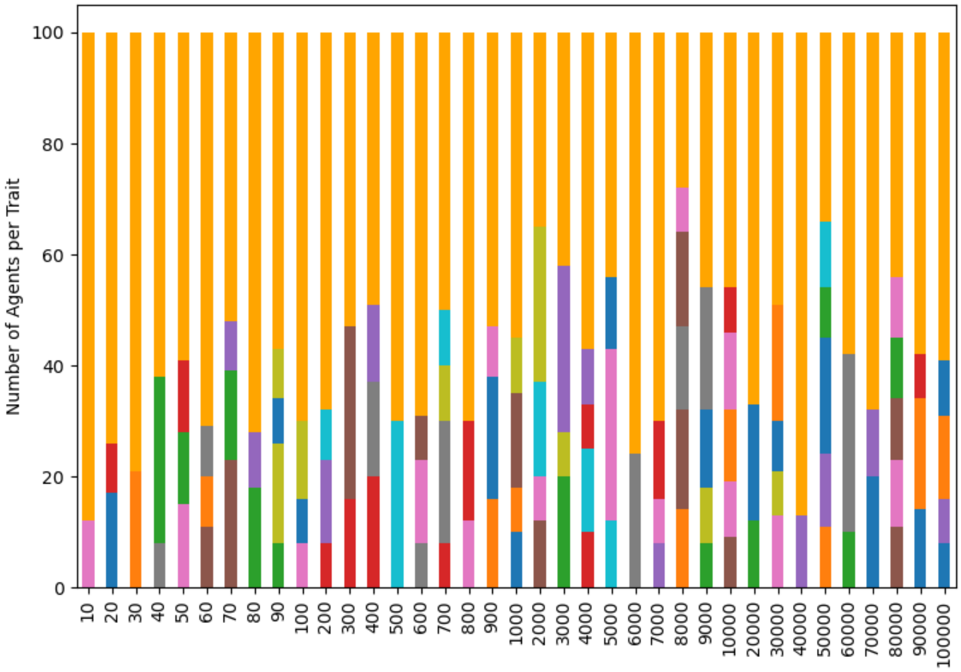


Figure 5.2: An example of the distribution of traits by subsequent but not consecutive generations as agents evolve their traits where parameter $S = 1$. In the first generation, all agents i are assigned as independent single-trait agents, and at the completion of each generation, agents evolve their traits and action rules and apply a mutation of 10%. Each colour represents a single trait with a minimum of eight agents. The yellow bar in each generation designates traits with a smaller number of agents, indicating that agents consolidate around a small number of traits per generation.

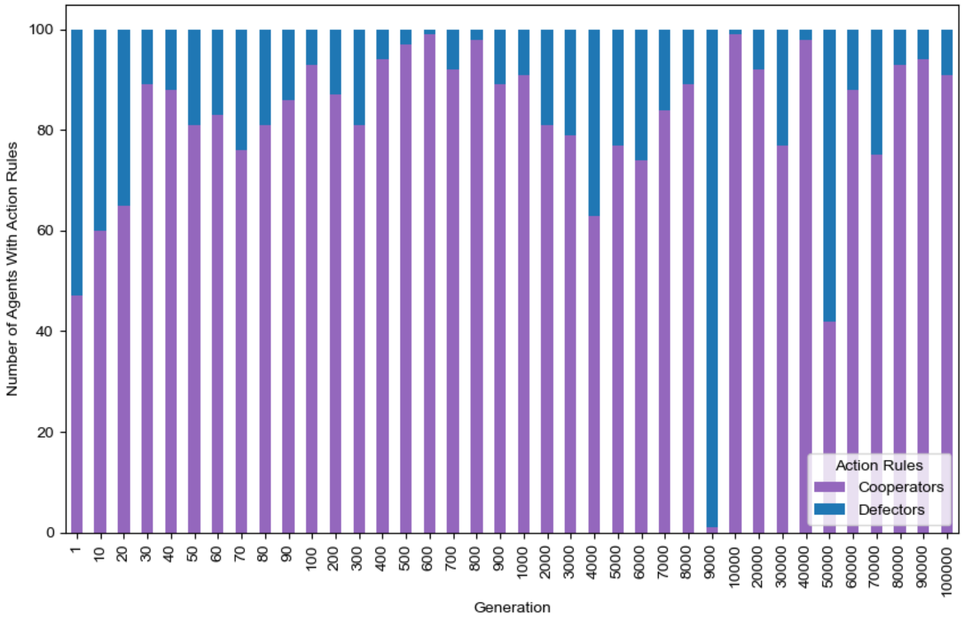


Figure 5.3: A snapshot of the distribution of action rules in subsequent but not consecutive generations, where $S = 1$ and agents evolve their traits, shows that cooperators ($s_i = 1$) are a more dominant strategy than defectors. However, defectors ($s_i = 0$) continue to infiltrate the population and can topple cooperators in some generations.

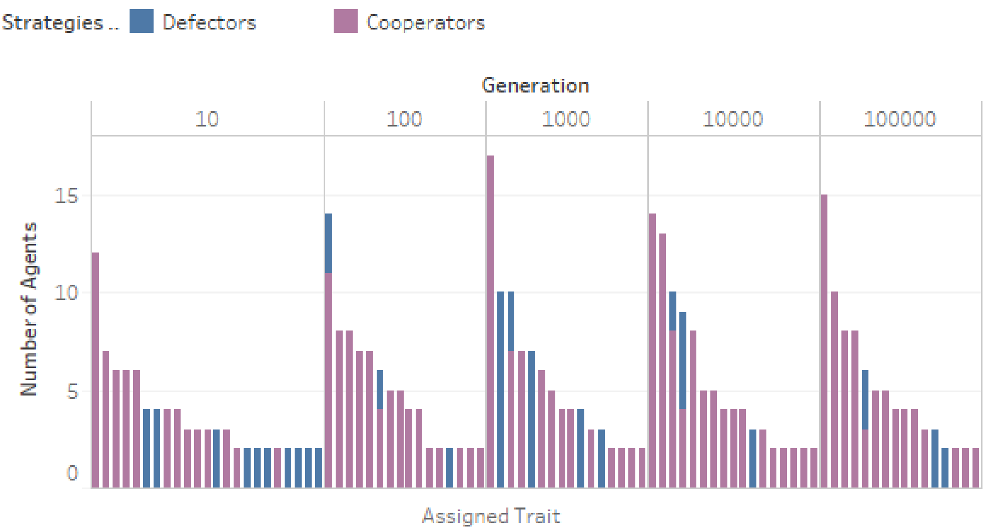


Figure 5.4: A snapshot of the distribution of action rules within the most shared traits, at fixed generations $(10$ to $100,000)$, when agents evolve their traits and parameter $S = 1$. Traits with a higher number of agents tend to favour being cooperators $(s_i = 1)$ rather than being defectors $(s_i = 0)$.

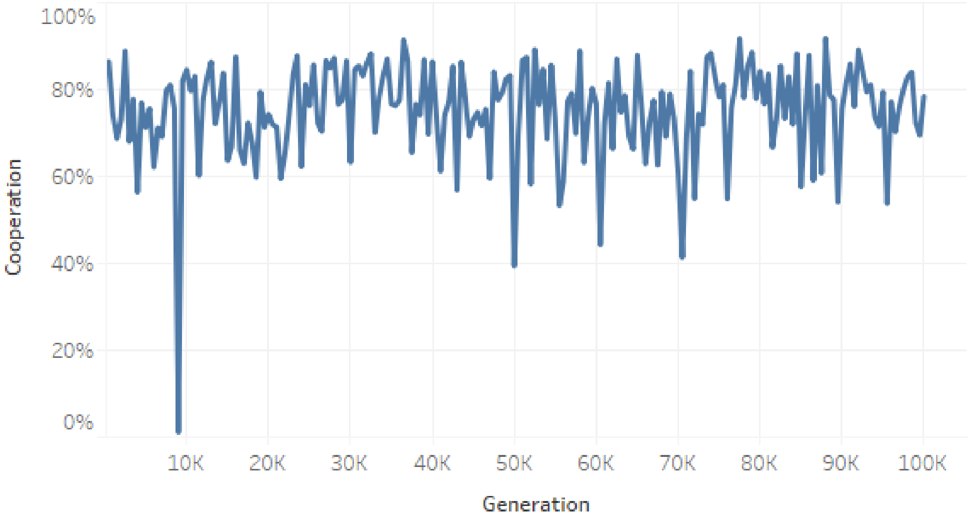


Figure 5.5: The cooperation recorded as a result of a single run where agents evolve their action rules and traits with a trait mutation of $\mu_T = 10\%$ and parameter $S = 1$. The single run shows that cooperation has a fluctuating trend throughout the $100,000$ generations when evolving both identity and action rules. The fluctuating trend is attributed to the struggle between cooperators and defectors and this can be seen in Figure 5.3.

To summarise, we found that the set of interacting cooperators sharing the same trait maximises payoff and, as a result, attracts other agents, increasing their number. The increase of agents having a trait takes place as long as their trait group does not include defectors (i.e., agents i with $s_i = 0$) that take advantage of shared reputations without donating. This provides opportunities for defective strategies to take hold and cooperation collapses. In this context, trait mutation allows cooperators to escape from defectors and move to an alternative trait. Similar behaviour is observed for high proportions of in-group interaction e.g. when $S = 0.5, 0.6, 0.8$.

5.3.2 Allowing Agents to Interact With Out of Trait Agents

Following on from Subsection 5.3.1, in this subsection, we analyse a single run of an experiment where $S = 0$ and traits are allowed to evolve. We adopt the same assumption in this experiment, where all agents i are single-trait agents ($|T_i| = 1$). Figure 5.1 shows the average cooperation of five different runs achieving less than 10% when individuals interact with any agent outside of their trait ($S = 0$).

To understand why cooperation achieves such a low rate, we examine a single run of the experiment. From Figure 5.6 it can be deduced that the defective strategy ($s_i, u_i, d_i = (0, 0, 0)$) is the dominant strategy in most generations. As agents are only allowed to interact with others outside their shared trait, donor agents cannot recognise individual agents. As such, defectors are able to exploit reputations that they share with others. In turn, defectors expand their payoff, allowing them to evolve within the population through reproduction, and eventually, they take over the population. This pattern is similar to the one discussed in Section 4.4, where defectors were able to exploit traits that had more than 15 members.

Although the defector strategy ($s_i, u_i, d_i = (0, 0, 0)$) is noted as the dominant strategy in most generations a fluctuating cooperation was still produced. In Figure 5.7 cooperation fluctuated at lower rates between generations leading to the low average of 10%. The fluctuation shows that cooperation was achieved at a higher rate in some generations before defective agents were able to exploit the trait.

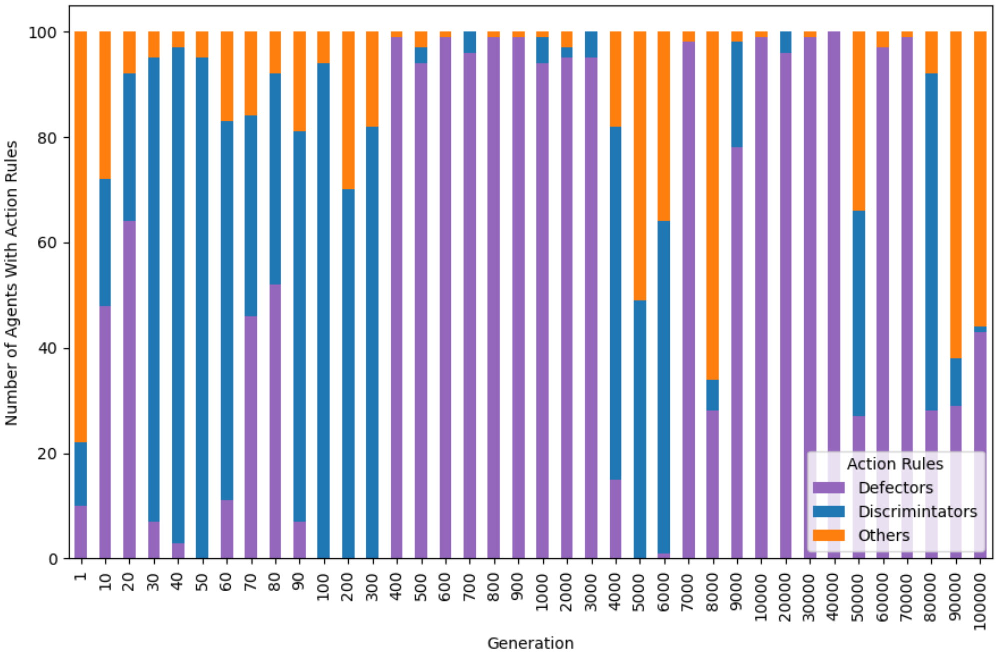


Figure 5.6: A snapshot of the distribution of action rules in subsequent but not consecutive generations when agents evolve their traits and $S = 0$ shows the most frequent strategy within the population is the defector strategy $(0, 0, 0)$. Early generations were dominated by the discriminator strategy, $(1, 1, 0)$. The remaining action rules, see Subsection 3.4.3, also appear in the population. This explains the lack of cooperation within the population that is displayed in Figure 5.7.

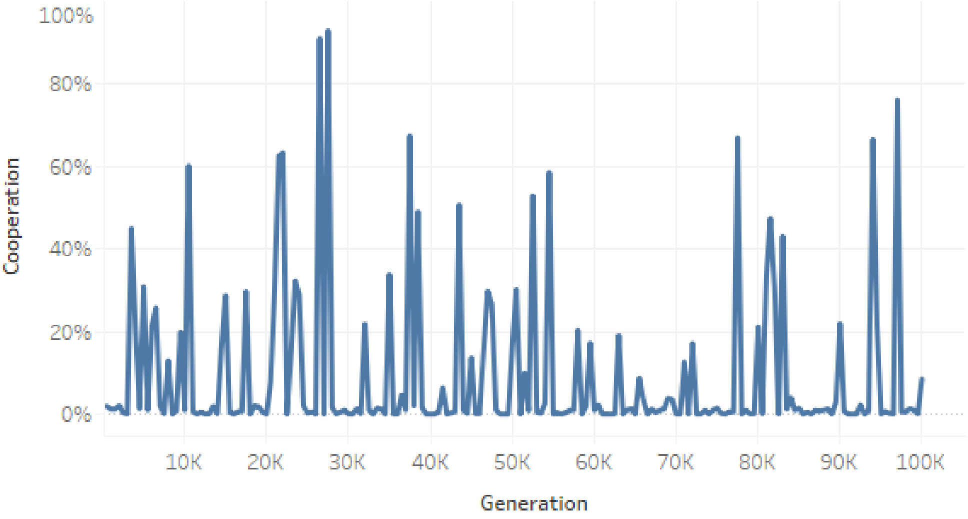


Figure 5.7: The cooperation recorded as a result of a single run of agents evolving their action rules and traits with a trait mutation of $\mu_T = 10\%$ and parameter $S = 0$. The run shows that cooperation fluctuates throughout the $100,000$ generations when $S = 0$ producing a cooperation with an average below 10% when evolving both identity and action rules.

5.3.3 Sensitivity of Mutation

In this subsection, we explore the influence of mutation on cooperation when agents are allowed to evolve both their traits and their action rules. In Subsections 5.3.1 and 5.3.2 a mutation of $\mu_T = 10\%$ was used on traits. Therefore, agents had a probability of 10% to mutate to a different trait at the end of a generation. This subsection, explores different mutation rates and their effect on both ends of parameter S, i.e. when $S = 0$ and $S = 1$. Note that the mutation used here is applied to traits and not on action rules.

To understand the criticality of mutation, we apply a variety of rates to both $S = 0$ and $S = 1$. The effect of mutation is seen in Figure 5.8. The Figure presents an average of 5 runs for each mutation rate applied on both parameter rates. The mutation rates applied are $\mu_T = \{0\%, 0.5\%, 1\%, 5\%, 11\%, 13\%, 15\%, 20\%\}$. When no mutation is applied, i.e. $\mu_T = 0$, an agent's trait does not evolve, and therefore agents are not able to escape from any defectors that share the trait with them. $\mu_T = 0$ is used to demonstrate the effect that mutation can have. In contrast, a mutation occurs more frequently when applied at a rate higher than 20%, resulting in fluctuating cooperation.

When the mutation rate is modestly increased (e.g., 1%), cooperative agents can change traits and rebuild a network of cooperative peers. The increase in mutation leads to more cooperation in both cases of parameter S; interestingly mutation has a higher effect on $S = 1$. However, when the mutation rate increases significantly, mutation impedes cooperation because agents are rapidly mixing, increasing the chances of defectors and cooperators to share a trait (e.g., mutation of $\mu_T = 100\%$ is pure chance). This mechanism underlies the results in Figure 5.4, which shows how traits are shared by agents, with a few traits achieving a large amount of sharing by cooperators. These results align with the conclusions of [71] and [186], where a limited rate of trait mutation allows cooperators to rebuild, albeit improving on the levels of cooperation achieved, this is discussed further in Section 5.5.1. Similar techniques aimed to deter defectors as a mean to promote cooperation in the absence of reputation, punishment, or other ostracising mechanisms have also had relevance in the literature [3].

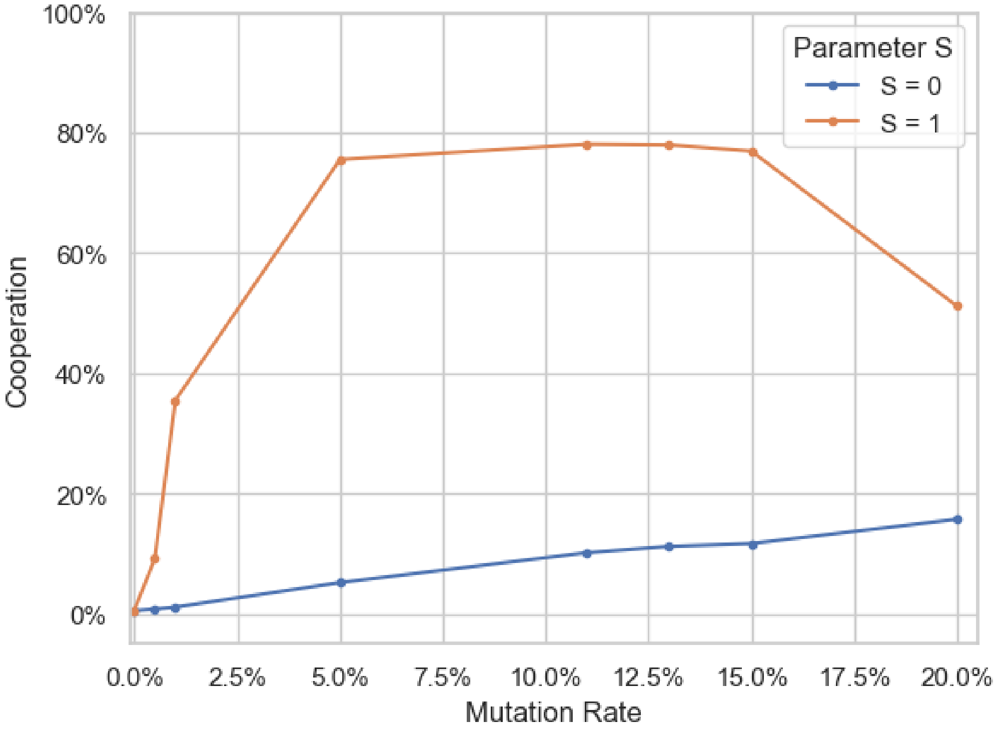


Figure 5.8: The application of different rates of mutation on traits as agents evolve their traits, Section 5.3.3, affects cooperation when $S = 1$ depending on the mutation rate applied. However, when $S = 0$ mutation has a lower effect.

5.4 Evolving Identity Rather Than Action Rules

In this section, we consider the effects of evolving agents' traits without evolving action rules. For all agents, action rules remain fixed throughout and are not subject to mutation. This experiment allows us to determine whether restricting agents from evolving their actions impacts cooperation. Additionally, we vary parameter S to assess whether limiting the interactions of agents within their trait group has another effect on cooperation. We begin by assigning all agents i as independent single-trait agents similar to Section 5.3. However, at the reproduction stage agents will only be able to change their traits based on the relative success of other agents without changing their action rules.

At the beginning of each experiment, agents are assigned randomly one of the eight action rules identified in Section 3.4.3. The random distribution of action rules results in about 12 agents per action rule. By forbidding agents from evolving their action rules, we limit the number of defectors and cooperators in the whole population.

As agents do not change their action rules at the end of each generation, parameter S has a low impact on cooperation, as can be seen in Figure 5.9. Specifically, when agents interact solely with those who do not share a trait with them, i.e. $S = 0$, limited cooperation is achieved with an average of 45% (SD = $0.02 - 0.04$) from five different runs across all parameter S values tested. Cooperation increases modestly with higher rates of in-group interactions, i.e. when S gets closer to 1 with an average of 50% of cooperation when $S = 1$.

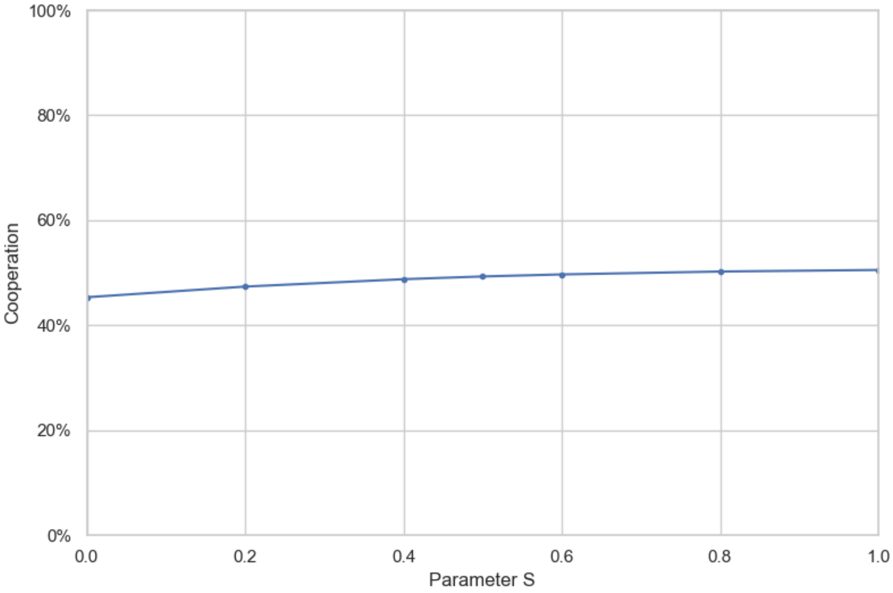


Figure 5.9: The average cooperation, of five runs, recorded as a result of agents only evolving their identity (traits) without evolving their action rules with a trait mutation of $\mu_T = 10\%$ while varying parameter S.

As cooperation does not alter much between different parameter rates, we analyse our results using a single run of the experiment where $S = 1$. The trait groups compose a different number of agents in each generation, with a different combination of traits. Therefore, changing identity does not offer protection against those holding defective strategies, as success equally attracts both defectors and cooperators, see Figure 5.10.

Accordingly, as the evolution of traits allows agents to copy others based on their success as indicated by agents' payoff, defectors exploit traits that have a healthier payoff. However, as agents are unable to copy action rules, the payoff achieved cannot exceed high levels and as such, produces a stable level of cooperation. Figure 5.11 shows a pattern of stable cooperation. The pattern reveals that as agents move between trait groups cooperation levels do not fluctuate much remaining constantly between $40 - 60\%$.

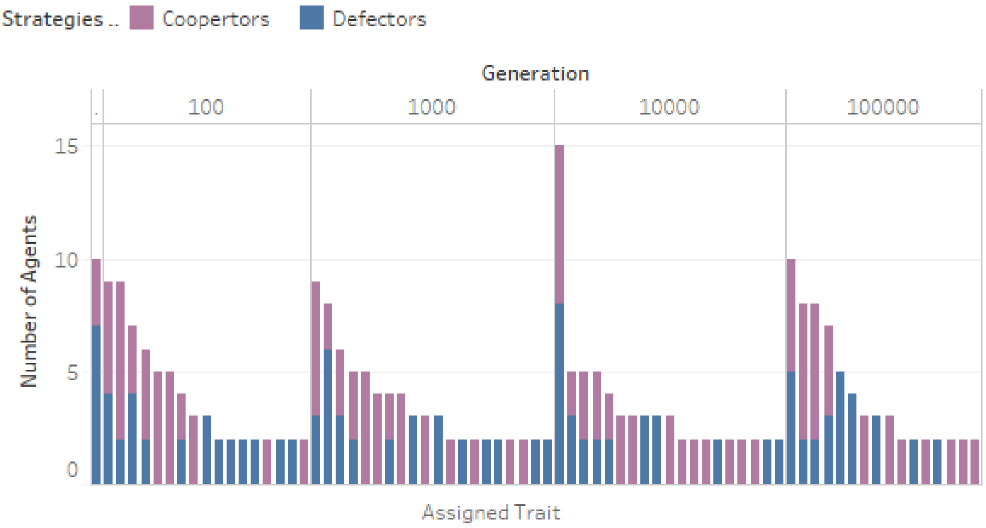


Figure 5.10: A snapshot of the distribution of action rules within the most shared traits, at fixed generations $(10$ to $100,000)$, when agents evolve their traits without evolving their action rules and parameter $S = 1$. Cooperators represent agents with action rule $s_i = 1$ and defectors represent agents with action rule $s_i = 0$, as agents do not evolve their action rules no dominant strategy emerges.

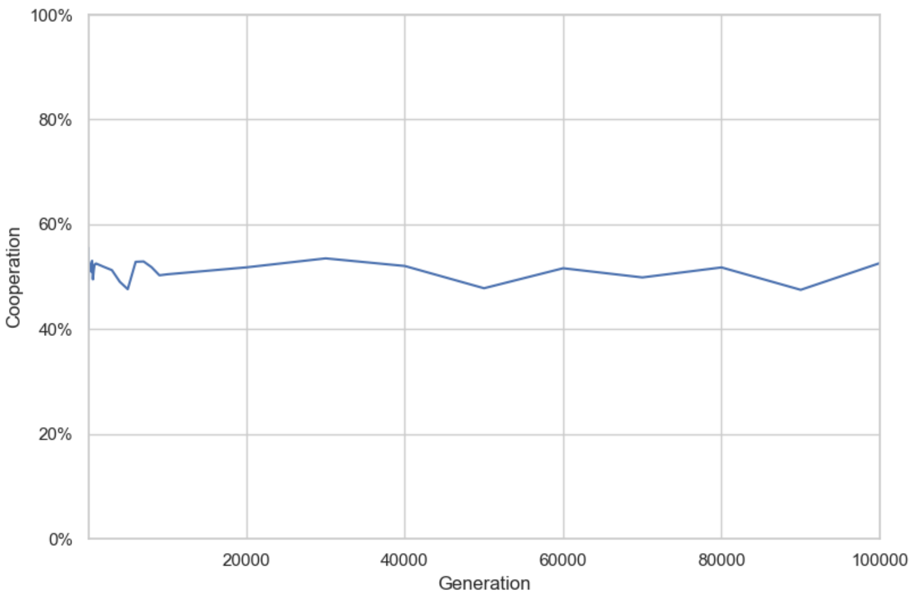


Figure 5.11: The cooperation recorded as a result of a single run of agents evolving their traits with a trait mutation of $\mu_T = 10\%$ without evolving their action rules where parameter $S = 1$ producing an average of 50%, when $S = 1$.

5.5 Discussion

In this section, we discuss our findings and their relations to the wider literature. Our results have shown that allowing agents to change their identity can be damaging to the emergence of cooperation. By allowing agents to “whitewash” their identity, an opportunity is introduced for defector agents to infiltrate cooperative trait groups by effectively resetting their reputation. The evolution of an agent’s traits allows it to adopt defective strategies that appear advantageous because they do not incur costs, but which cause cooperation to collapse. However, cooperation emerges when agents primarily interact with those having the same trait in common. We review this in more detail below.

In Section 5.4, our results showed that cooperation stabilises between 45% and 50% when agents are permitted to *evolve their traits without evolving their action rules*. This occurs with very small influence from parameter S. Because agents were unable to change their action rules, high levels of cooperation were capped by the extent of defective strategies, causing cooperation to stabilise. In other words, the inability for defective behaviour to be replaced imposes a fundamental limit on the widespread achievement of cooperation across the population.

When we provide agents with the freedom to evolve both their action rules and traits (Section 5.3) very different results were observed. The results showed that cooperation does not evolve when agents solely interact with agents having different traits, in contrast to the results when agents' traits do not evolve (Section 4.4). This occurs when S is low - specifically, in Section 5.3.2, our results showed that *cooperation could not emerge above* 10%. However, within this scenario, varying the rate of mutation on traits had a positive effect and cooperation increased as the rate of mutation increased, yet the increase in cooperation was relatively small as seen in Figure 5.8.

Interestingly however, the results showed that when agents favour same-trait interactions (i.e., $S = 1$), as seen in Section 5.3.1, and evolve both their action rules and their traits, cooperation reaches an average above 70%. As agents interact with others who share the same trait, and therefore the same reputation, the reputation system collapses as it is ineffective in differentiating between individual agents. Furthermore, due to agents effectively being in sets based on their trait, *donation decisions are made on a single component of an agent's action rule, namely* s_i which in this context identifies whether or not two agents have the same reputation. Agents are therefore either defectors ($s_i = 0$) or cooperators ($s_i = 1$), based on s_i, while d_i and u_i are redundant members of the action rule. Finally, we observe from the results that *higher rates of mutation promote cooperation* in comparison with lower levels of mutation, as can be seen in Figure 5.8. However, it appears that the presence of defectors still caps the overall cooperation level to around 70%, which is still a respectable level. This is a surprising finding that has led us to question how such high levels of cooperation are sustained, while also noting that in this case, the reputation system collapses due to shared identity. This indicates that alternative dynamics must come into play that enable cooperation to function independently from a shared identity and without a reputation system. This is unexpected and represents a significant new observation that we examine further in the following section.

5.5.1 Links to Evolutionary Set Theory

Evolutionary set theory takes its inspiration from observation of human society [186], where both cooperation and the presence of “sets” (e.g., groups) are abundant in the human population. The concept is based on the hypothesis that the population structure itself is a consequence of an evolutionary process, and one which also results in cooperation. The assumption here is that within a population, individuals are distributed across a range of sets, and the sets give a boundary within which the individuals can interact in an evolutionary game. The game then allows agents to evolve the set to which they belong, as well as their strategy for cooperation. This results in a simple representation where the structure of the population is a product of evolution, and where agents are not concerned with sustaining their reputation as a means to secure future payoff. As such this represents a distinctive alternative paradigm for cooperation that has minimal cognitive requirements and is independent from reputation systems.

Being one of the first alternatives to a reputation system for cooperation, the concept of evolutionary set theory (Tarnita et al. [186]) focuses on agents’ set memberships and their movements between sets. This is consistent with human sensitivity to “in-group” membership [31]. Their model is also in line with the social identity theory discussed in Section 2.3 as it “suggests that preferential cooperation with group members exists”.

In more detail, the evolutionary set theory model divides agents into multiple sets that may overlap. Agent interaction is limited to agents that have sets in common with each other. This limitation results in agents adopting one of two action rules - simply either to cooperate or to defect, and no complex action rules are applied. Agents update their cooperation strategy and set membership during asexual reproduction where fitness is weighted by payoff. Mutation is an important part of the model, allowing for random perturbation.

Through this approach, it has been reported that moderate levels of cooperation (e.g., $50\% - 70\%$) can be sustained with limited mutation on group membership [122, 185]. The evolutionary set theory model has been used to investigate in-group favouritism [71, 154], allowing the authors to further consider out-group interactions [71].

Interestingly, we note that our model induces a set structure onto the population through traits - in other words, each agent i is a member of set t if and only if agent i identifies with trait t for its reputation. Note also that i belongs to multiple sets when its reputation involves more than one trait. Equally, as described in Section 5.3.1, the action rules collapse to the choice between two strategies, cooperate or defect, based on s_i. Consequently, we observe that *our model provides a bridge between cooperation based on reputation (indirect reciprocity) and cooperation based on the evolution of sets (evolutionary set theory) through traits.* This is a significant observation that relates two alternative approaches to sustaining cooperation which have previously been considered unrelated, and indeed potentially orthogonal in the sense that these relate to different communities of interest (i.e., reputation systems and set theory). We believe this is the first work that has been able to show the relationship between these different types of models as we have not observed it in other works.

To observe the differences and similarities between cooperation based on reputation and cooperation based on the evolution of sets, we summarise the key points in Table 5.1. Indirect reciprocity relies on reputation, as agents make donation decisions they utilise other agents' reputation before making that decision. Explicitly in this thesis, reputation is based on the traits carried by agents, which provides the basis for groups of agents to be identified via their common traits. Equally, in studies of evolutionary set theory, agents are divided into sets based on commonality and agents interact with others who share the same sets as them. High levels of trait sharing effectively leads to the collapse of the reputation system, because the agents' reputations cannot be distinguished from each other when agents interact "in-group". However, at this point the reputation system collapses and evolutionary set theory takes hold, provided $S = 1$, as described above.

This is enabled because in both systems, agents reproduce proportionally to payoff and use the Wright-Fisher model [66] as a basis for evolution, allowing inheritance of the most successful strategies from the previous generation to propagate.

	Indirect Reciprocity without Sharing of Traits [199]	**Evolutionary Set Theory [186]**	**Indirect Reciprocity with Sharing of Traits**
Cooperation based on;	Agent's reputations	Set membership	Reputation based on traits which can be shared with others
Agents evolve;	Action rules (strategies)	Action rules (strategies) and set membership	Action rules (strategies) and traits
Agents reproduce proportional to;	Payoff (fitness)	Payoff (fitness)	Payoff (fitness)
Number of strategies;	Eight	One: either cooperation or defection. [186] Two: One for in-group and one for out-group [71]	Eight
Type of mutations;	Strategy mutation	1.Strategy mutation 2. Set mutation	1. Strategy mutation 2. Trait mutation
Interaction	Mixed interaction without limits	Limited to agents who share the same set. Mixed interaction without limits	Parameter S is the probability of interaction with agents that have a shared trait(s).

Table 5.1: Comparison of Cooperation Models Based On: Indirect Reciprocity Without the Sharing of Reputation, Evolutionary Set Theory, and Indirect Reciprocity Based on Traits.

5.6 Conclusions

The aim of this chapter was to study the evolution of traits and its impact on cooperation. We experimented with a form of whitewashing where we enabled agents to adapt their identities by changing traits through evolution. As agents adopted new identities, without constraints on population mixing, defectors were able to gain an advantage as they could benefit from new reputations as they benefited from sharing traits with cooperators.

However our results have also shown that cooperation can emerge depending on several factors. Notably, we find that cooperation maintains an average of 50% when agents evolve their traits without evolving their action rules (Section 5.4). The result is attributed to the inability of agents to change their action rules, creating a struggle of dominance between defectors and cooperators.

Furthermore, when agents are allowed to evolve both action rules and traits, surprisingly we observe that cooperation reaches an average above 70%, but only when in-trait interactions are favoured over out-of-trait interactions (Section 5.3.1). Note that when interactions are in-trait that reputation system collapses. This is sensitive to the effects of mutation on the emergence of cooperation (Section 5.3.3) optimally requiring a mutation rate around 5-15%. However we have established that this finding is entirely rational, and occurs as a consequence of evolutionary set theory (Subsection 5.5.1). In doing so, we have established the point at which the collapse of the reputation system, due to shared identity, is replaced by the structure of the population evolving to favour cooperation. This is a new finding that relates previously disparate approaches to explain indirect reciprocity.

Overall the results of this chapter show that cooperation emerges when agents evolve their identities, but cooperation can also be limited due to defectors being able to exploit shared reputations. When agents share a trait, they also share the trait's reputation with potentially many other agents. As such, traits provide an opportunity for defectors to exploit other agents. In the next chapter, we allow agents to hold both a personal and a shared trait, investigating the consequences of allowing both a unique and shared identity to coexist. By providing a personal trait, agents can compose their identity to reflect both personal and group elements, which has not been previously explored, to the best of our knowledge.

Chapter 6

Blended Identity and its Impact on Cooperation

6.1 Introduction

In this chapter, we consider identity that is based on a combination of both a personal trait (i.e., not shared with others) and a shared trait. We refer to this as *blended identity*. This extends our representation of traits in the previous chapters and presents a general framework where an agent's individual and shared reputations coexist.

In this framework, a shared trait can be thought to represent an identity of the group that an agent belongs to. In contrast, a personal trait represents a unique identity that is only available to an individual agent. We extend our model from Chapter 3 to further explore blended identities through this combination of traits. Specifically, we allow an agent to express a balance where its identity is a combination of group identity and its own personal (i.e., unique) identity.

To investigate this, we include a new parameter that enables agents to weight the combination of a group trait and an individual trait that together represent their identity. By providing agents access to both types of traits, agents can compose their identity to reflect different levels of personal and group elements, which has not been previously systematically explored. The investigation of blended identity is interesting because it allows us to understand to what extent individual identity can coexist with group identity while sustaining cooperation. This represents the focus of our work.

The chapter is structured into subsections as follows. In Section 6.2, we introduce the concept of a blended identity based on a combination of a personal trait and a shared trait and discuss its relation to the literature on identity in the psychology, as introduced in Section 2.3. In Section 6.3, we formally extend our model to accommodate the concept of blended identity. We address how reputations are updated in Subsection 6.3.1.

The presented experiments in this chapter are divided into two main sections. In Section 6.4 the experiments rely on agents that do not evolve their blended identity; these are referred to as *exogenous agents*, keeping their identities constant throughout multiple generations. This section is divided into subsections as follows: Subsection 6.4.1 applies the extended framework and is a benchmark experiment that determines how blended identities may impact cooperation as compared to identity without blending. Subsection 6.4.2 reintroduces parameter S to allow agents to control who they prefer to interact with depending on the extent of their shared identity. Subsection 6.4.3 focuses on the reproductive step, and uses the agent's blending levels to allow agents more flexibility in determining who they copy for reproduction of the next generation. Finally in Section 6.5, we allow agents to evolve their identities by evolving their blending levels. This allows agents greater flexibility during the reproduction phase, which aligns with their identity. The chapter concludes with a discussion in Section 6.6 where we present overall findings regarding the evolution of cooperation and blended identity.

6.2 Blended Identities in Psychology

Identity is a fundamental topic of high interest within social psychology; specifically, researchers have been interested in the concept of social identity and how groups impact individuals to form a part of their identity. In Section 2.3, we highlighted how psychologists consider groups to be vital to any individual's identity. In this section, we continue our investigation of shared identities based on an individual's group membership and it is useful to recap the most important points. Generally, identities are a combination of personal and shared values [33], which is a basis for our investigation, which we call *blended identity*. Personal identity defines unique characteristics of individuals that distinguishes them from others. In contrast, shared (group) identity represents the relationship that an individual has with groups to which they affiliate. Our notion of blended identity combines personal and group identities for the individual, which is a fundamental element of social psychology.

Humans are naturally social beings and belong to multiple groups which allow them to be identified (in part) through their group affiliations [180]. For example, university students can be identified through the university which they attend. Outside of the university setting the students may be alternatively known through their personal identity or affiliation with other groups. This provides a way through which individuals can express themselves in differing contexts using alternative elements of their identity. It also reflects that humans are strongly pro-social, and therefore their identity is influenced by the groups to which they belong [195].

At the same time, individuality is a key part of identity as humans seek to be distinct and to be unique [191]. This might seem to be a slight conflict, but it can be resolved by balancing personal identity and the identity that is shared. Humans seek to be recognised for their individuality and want freedom in self-expression; this forms their personal identity. On the other hand, humans are group-oriented as they share a part of their values with others and want to be in groups that help to express their beliefs and reputation; this forms their shared identity [97]. Our approach and assumption in this chapter is an abstraction of this scenario - agents are able to maintain both individual and group identities which is balanced for each individual agent by blending their two identities.

We also note that this simultaneous representation of both personal and group identities aligns with the concept of *identity fusion* [175, 177]. Identity fusion describes personal and group identities working in tandem, with individuals retaining a strong sense of self-identity, and with that identity "overlapping" with that of a group, creating self-driven individuals that act with strong group-level devotion. In the human world, this can result in the self-less empowerment of the individual to take extreme action in response to group opposition and threats. Here, in the absence of human cognition, we are examining the structural implications of this combined identity, using indirect reciprocity. This allows us to determine how cooperation becomes disrupted by blended identity.

6.3 Extending the Trait Based Model for Blending Identity

In this section, we modify the framework developed in Chapter 3. The changes are introduced to incorporate agents holding both a personal and a shared trait, resulting in a blended identity based on the two traits, as can be seen in Figure 6.1. The underlying model, as introduced in Chapter 3, uses the evolutionary game theory framework for indirect reciprocity with reputations based on traits. Within this model, agents compare reputations and use action rules to determine their donation decisions.

To accommodate blended identity, we extend the model as introduced in Chapter 3 in several ways: In Section 6.3.1, we introduce a blending parameter which allows agents to balance their reputation between their personal and their shared trait, a similar parameter is used in [197] to balance an agent's reputation between their personal identity and their fusion level. The remaining components and parameters of the model have not been modified; these include the use of parameter S to determine the interacting agents (Section 3.5.1), the usage of action rules in the model (Section 3.4.3), and the evolutionary simulation and the application of mutation to action rules (Section 3.5.3). This allows us to specifically observe the effects of the blending parameter on the model. In Section 6.3.2 we highlight the changes to the underlying model from Chapter 3, as presented through Algorithm 3.

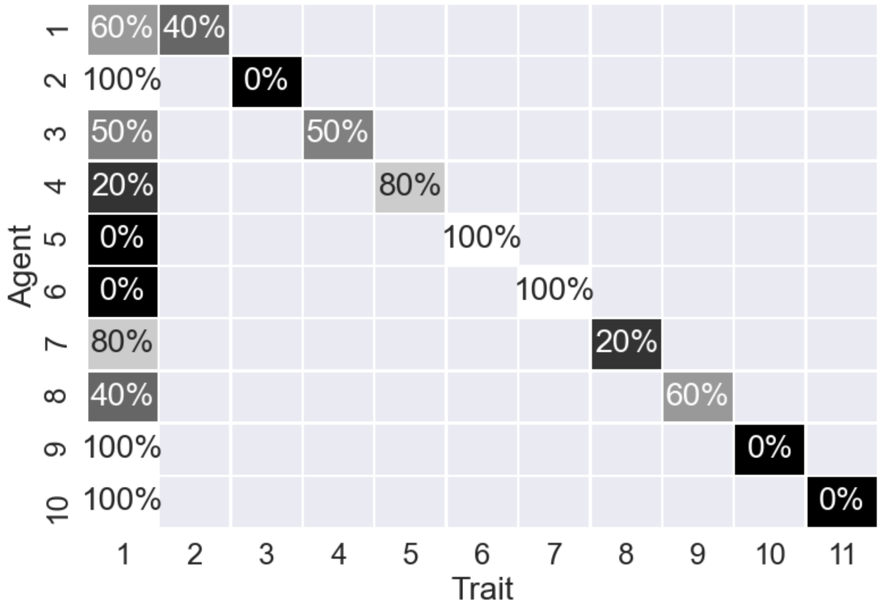


Figure 6.1: A simple visualisation of 10 agents where all agents i have an identity composed of a personal trait and a shared trait. Each agent i has a personal trait t_i and a shared trait g as can be seen in the diagram. However, each agent i has a different blending level, w_i, which determines how their blended identity is composed of the two traits. For example, agent 1 has trait 2 as a personal trait and shares trait 1 with all other agents i. Agent 1 also has $w_1 = 60\%$ while agent 2 has $w_2 = 100\%$, therefore agent 2 is blended to the group at a higher level than agent 1. Note that w_i is bounded between 0% and 100% where $1 == 100\%$.

6.3.1 Updating Reputation for Blending Identities

Reputation is the score given to individuals in response to interacting with others. Reputations are built on past interactions of individuals or through their associations. In terms of our model, reputations are used in determining the donation decisions taken by agents. These are based on the traits that each agent subscribes to. In this chapter, our assumption is that each agent i has a personal trait and a shared trait. To combine these, all agents i are assigned a blending factor referred to as parameter w_i, where $0 \leq w_i \leq 1$ (see Figure 6.1).

Specifically, each agent i carries a shared trait g and a personal trait t_i and a blending level w_i. The shared trait, g, is carried by all agents, while the personal trait, t_i, is unique to each agent i, i.e. we assume that personal traits are all unique. Each trait carries its own reputation, denoted $r(g)$ and $r(t_i)$ respectively. Traits are used by the individual i to represent the elements by which i derives its identity and its reputation r^i. The blending of the two traits, through w_i, allows the individual to express the extent to which their personal-self is composed of a group's identity (g) as opposed to an individual identity in isolation (t_i). This structures the individual agent's personal reputation, expressing the extent to which this is derived from the blend with the group's reputation, as opposed to its individual behaviour. Then i's reputation is defined as:

$$r^i = (1 - w_i)r(t_i) + w_i r(g) \tag{6.1}$$

where $0 \leq w_i \leq 1$. When w_i is high, i's identity is highly blended to the group, with i's reputation r^i being derived mainly from that source. Conversely, when w_i is low, the individual predominantly derives their own personal identity - the group's identity plays a lesser role in the agent's reputation.

At each interaction, the reputation that i holds is updated based on i's donation decision in respect of a potential recipient j (lines 13-22 in Algorithm 3). We use the general concept of standing [171], as in Chapter 3 Subsection 3.5.2, to update i's reputation. Because this is composed of two traits, both the group reputation, $r(g)$, and i's personal trait's reputation $r(t_i)$ are updated. Specifically, if i donates, then $r(g)$ is incremented if and only if w_i is non-zero, and $r(t_i)$ is incremented if and only if $w_i \neq 1$. If $r^j \geq r^i$ and i defects then $r(g)$ is decremented if and only if f_i is non-zero, and $r(t_i)$ is decremented if and only if $w_i \neq 1$. Note that this updating approach ensures that a reduction in reputations does not occur when i fails to donate and j is of a lesser reputation, providing a defence against shirkers, consistent with the concept of standing. We allow the value of each trait to vary in the integer range [-5,5]. This approach to standing is based on the original approach by [171] as modified for indirect reciprocity based on social comparison [199].

6.3.2 Algorithm for Evolution With Blended Identity

In this section, we present a framework through pseudocode. In Algorithm 3, we present the parameters of the model, which is based on 1 to accommodate blended identity and, as such, focuses on the addition of parameter w_i. The parameter enables us to calculate agents' reputations in the model by weighting the extent that an individual's reputation is drawn from the shared (i.e., group) trait. As discussed in Section 6.3, the changes are a result of agents having a blended identity made up from both a personal and a shared trait. The remaining parameters of the model remain unchanged, as compared to Algorithm 1.

To begin, this pseudocode sets the number of agents in the population, the number of generations and the number of rounds per generation. Once the population is created, we assign each agent i a set of traits T (line 2). Unique to this chapter, all agents have been additionally assigned a blending parameter w_i. In some experiments (Section 6.4), the distribution of the blending parameter is uniform, that is all agents i have the same blending level. In other experiments (Section 6.5), the distribution of the blending level is assigned randomly to each agent. The range of blending levels is bounded between 0, and 1 and in all cases w_i belongs to the set $W = \{0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0\}$.

Once the initialisation of the algorithm takes place, each round begins by picking a recipient (line 4) and a donor (lines 5 - 12). The process is informed through the use of parameter S in some experiments (e.g., Sections 6.4.2, 6.4.3 and 6.5), which biases interaction towards those with similar characteristics (i.e., in-group bias). The nature of these experiments (Sections 6.4.2-6.5) allow agents to use parameter S to control whether agents prioritise interaction with other blended agents (lines 6 - 7) or with the independent agents (lines 8 - 9). In the absence of parameter S, the interaction pair is chosen at random (lines 10-11). The process of picking the interaction pair is highlighted in Algorithm 3 in lines 4 - 12.

The next step in the pseudocode is the donation decision taking by the donor agent i (lines 13 - 22), which is consistent with the basic model from Chapter 3. Each agent i determines their donation decision through their action rules which dictate whether agents donate or defect by comparing the donor agent i's reputation with that of the recipient agent j (line 15). Once the decision has been taken the reputation of agent i is updated to reflect its decision using the standing assessment rule (lines 15 and 19). The blending parameter, w_i, is used to determine each agent's reputation as a combination of the reputations held by the shared trait g and their personal trait t_i. Both traits from which agent i's reputation is composed are updated once the donor agent implements its donation decision.

Once a generation is complete, the reproduction process begins, see Section 3.5.3 and lines 24 - 34. This stage determines how agents evolve between generations by enabling agents to inherit, or effectively copy, the action rules of previous generations (line 27).Agents that do not evolve their blended identity are referred to as *exogenous agents*, keeping their identities constant throughout each generation. In Section 6.4.3, we use parameter w_i to determine if blended agents will inherit solely from other blended agents or the population as a whole.

Algorithm 3 Algorithm for Indirect Reciprocity Based on the Reputation of Traits and Blended Identities

Require: Number of iterations m; number of generations M; set of agents A; set of traits T; set of traits T; set of binary action rules $AR = (s_i, u_i, d_i)$; set of blending levels W; cost c; benefit b; in-group interaction probability S; mutation rate of action rules μ_A; the set of agents that share at least one trait with agent j N_j;

```
for $M$ generations do                                        ▷ Evolutionary simulation
    Set $f_i = 0\ \forall i \in A$ and $r_t = 0\ \forall t \in T$
    for $m$ iterations do
        $j \leftarrow U(A)$                                   ▷ Select recipient (see Section 3.5.1)
        $p \leftarrow U(0,1)$
        if $p < S$ and $|N_j| > 0$ then
            $i \leftarrow U(N_j)$                             ▷ Select random in-group donor
        else if $p \geq S$ and $|\bar{N}_j| > 0$ then
            $i \leftarrow U(\bar{N}_j)$                       ▷ Select random out-group donor
        else
            $i \leftarrow U(A - \{j\})$                       ▷ Select random donor
        end if
        ▷ Apply action Rules (see Sections 3.4.3 and 3.5.2)
        $r^i = (1 - w_i)r(t_i) + w_i r(g)$                    ▷ $i$'s reputation is derived from $w_i$
        if ($r^i$=$r^j$ and $s_i = 1$)                         ▷ Compare equal
                or ($r^i < r^j$ and $u_i = 1$)                 ▷ Compare upwards
                or ($r^i > r^j$ and $d_i = 1$) then            ▷ Compare downwards
            $r_t \leftarrow min(5, r_t + 1)$                  ▷ $i$ donates, increase reputation
            $f_i \leftarrow f_i - c$; $f_j \leftarrow f_j + b$   ▷ Update fitness
        else                                                  ▷ $i$ defects
            if $r^j \geq r^i$ then                            ▷ Detect unjustified defection.
                $r_t \leftarrow max(-5, r_t - 1)$             ▷ Decrease reputation
            end if
        end if
    end for
    ▷ Reproduction stage (see Section 3.5.3)
    for $i \in A$ do
        $j \leftarrow R(A, f)$                                ▷ Roulette wheel based on fitness
        $(s'_i, u'_i, d'_i) \leftarrow (s_j, u_j, d_j)$       ▷ $i$ copies $j$'s action rules
        if $U([0,1]) < \mu_A$ then
            $(s'_i, u'_i, d'_i) \leftarrow U(AR)$             ▷ Mutate action rules
        end if
    end for
    for $i \in A$ do
        $(s_i, u_i, d_i) \leftarrow (s'_i, u'_i, d'_i)$       ▷ Update action rules for all agents
    end for
end for
```

6.4 Exogenous Blended Agents

In this section, we consider the effects of exogenous agents having a blended identity. Exogenous agents are agents that have a constant blending level throughout an experiment and only evolve their action rules (i.e., blending levels remain fixed). Having a blended identity allows cooperative agents the opportunity to reduce the impact of free-riders exploiting shared reputations. An essential element of blended identity is the blending parameter w_i, which controls the extent that an agents' reputation is dependent on the shared trait g. Additionally, we consider the effects of agents using the blending level parameter to influence their interaction partner aligned to parameter S, as in Subsection 6.4.2.

All experiments in this section share some common assumptions. We begin by assigning all agents the shared trait g and a personal trait t_i. A group of agents are designated as blended agents and are each assigned a uniform blending level $w_i > 0$: therefore all blended agents i have the same blending level w_i, which does not change during the simulation. In other words, we divide the population into two sets, one that has a blended identity $w_i > 0$, and one that is independent and has $w_i = 0$. The number of agents that have a blended identity remains fixed throughout an experiment. We vary the number of agents sharing trait g in each experiment, using group sizes of $10, 20, 40, 50, 80, 90$ and 100. Throughout we apply blending levels from the set W (see Section 6.3.2). All agents evolve their action rules and a mutation of action rules is applied with a probability of 1%, based on previous experimentation in Section 3.5.3 and also in [199], where it was found to be sufficient to trigger a change in the structure of the population. The mutation applied changes each element of the action rules binary vector of an agent (Section 3.4.3).

Note that throughout this Section, agents do not evolve their blending levels as these are assumed to be exogenously (i.e., externally) controlled. Note that if an agent i has $w_i = 0$, it is considered to be independent of the group (i.e., doesn't share trait g), and its reputation would be based entirely on their own past interactions. In this case the results from [199] are replicated. Furthermore, if an agent i has $w_i = 1$, they are considered fully dependent, and their reputation would be entirely shared and based on all agents sharing the trait g, replicating the results in Section 4.4.

To explore the impact of blended identities on cooperation, we divide our experiments into subsections that address different blending assumptions. These are designed to explore the extent that blending levels may promote cooperation while deterring free-riders. In Subsection 6.4.1, we establish a benchmark experiment that allows us to test the model's validity with the newly updated components. In Subsection 6.4.2, we use parameter S to allow blended agents to prioritise interacting with other blended agents (i.e., in-group effects). In Subsection 6.4.3, we examine a probability based approach, where blended agents decide whether to inherit from blended agents or the whole population based on the blending level w_i.

6.4.1 Cooperation From Agents With Blended Identities

In this section, we consider the cooperative effects of blended agents. By allowing all agents to hold both a personal trait and a shared trait, we can determine the effect of a blended identity on reputation and subsequently cooperation. This gives us a benchmark that enables us to compare our results with previous findings and subsequent changes. In particular, in Section 4.4, we classified single-trait agents as either independent or dependent. In the extended model presented in this chapter, independent agents have a blending level of $w_i = 0$ while dependent agents have a blending level w_i from the set $\{0.2, 0.4, 0.5, 0.6, 0.8, 1\}$.

In this experiment, we analyse how agents with different blending levels impact on cooperation. The experiment in this Section adopts the assumptions identified in Section 6.4 - that is all blended agents are assigned the same blending level w_i and these remain fixed and are not subject to evolution. Additionally, parameter S is not used when agents choose a potential partner for donation, allowing agents to interact without any restriction (i.e., uniform random pairing of interacting agents). We observe the impact of free-riders as blending levels are changed, alongside the number of agents with a non-zero blending level.

Figure 6.2 shows that an increase in blending levels and group size (i.e., number of agents with a non-zero blending level) leads to a decrease in cooperation. We note that as cooperation declines defectors quickly dominate the population, as shown in Figure 6.3. The dominance of defectors in the population is attributed to the high f_i produced due to their high blending levels, which results in the defector strategy being copied. However, when cooperation yields higher results the discriminator action rule $(s_i, u_i, d_i = (1, 1, 0))$ is more prevalent in the population. The results show that *once the weight distribution shifts towards a sharing trait and become more dependent, i.e., w_i tends towards 1, cooperation decreases.* In contrast, when agents favour their personal trait and are more independent (i.e., w_i tends towards 0) cooperation rises. The result is expected as agents rely on their shared trait to cooperate. This result is also consistent with the results from Section 4.4, where cooperation decreased as more agents shared traits. This also provides confidence in the function of the extended model that is the focus of this chapter.

When the whole population (100 agents) share some elements of a trait ($w_i < 1$), agents sustain cooperation as long as they do not share 100% of the trait, as shown in Figure 6.2. This result is attributed to the pair of donor and recipient agents sharing the same proportion of the reputation which forces agents to rely on their personal trait to interact with other agents regardless of their blending level. Once agents are 100% blended, they have to rely solely on the shared trait's reputation, which causes their cooperation to decrease. In summary, these results show that cooperation can be sustained as long as agents do not fully share their trait ($w_i = 1$) and therefore avoid fully sharing their reputation. This shows reasonable resilience overall, even in the context of considerable trait sharing.

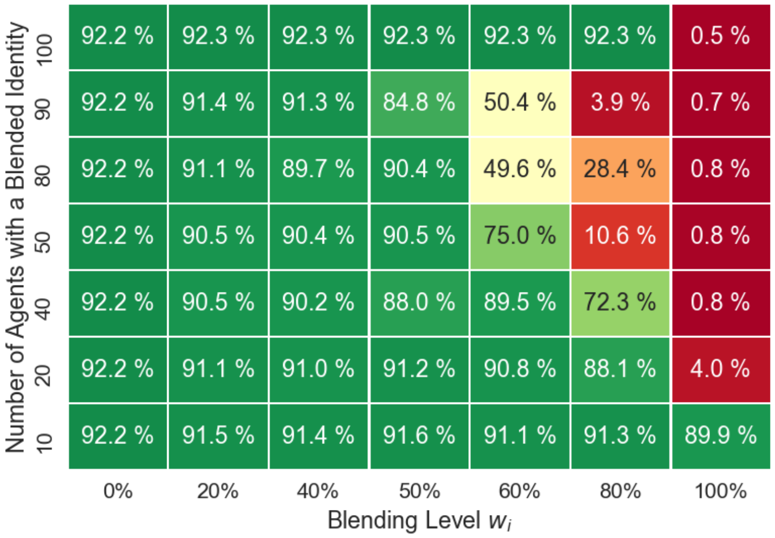


Figure 6.2: The average cooperation recorded over generations when agents have a blended identity with different starting configurations. The results indicate that cooperation can be achieved at high levels when agents have blended levels lower than 80%. Once agents become more devoted to their group, cooperation declines as agents become more reliable on their shared reputation allowing defectors to exploit it. Similarly, an increase in the number of agents sharing a trait causes cooperation to decline as the large number of agents allows defectors to take advantage of the group's reputation.

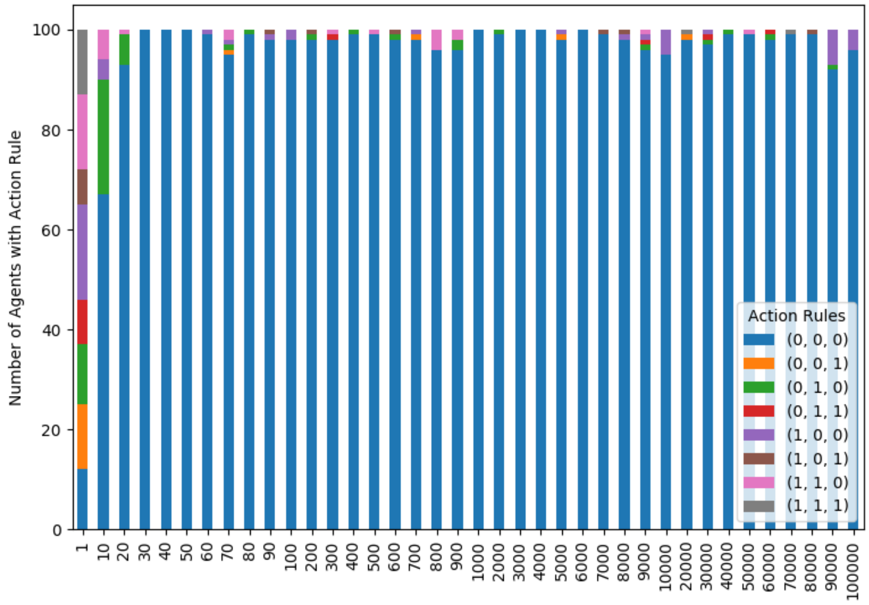


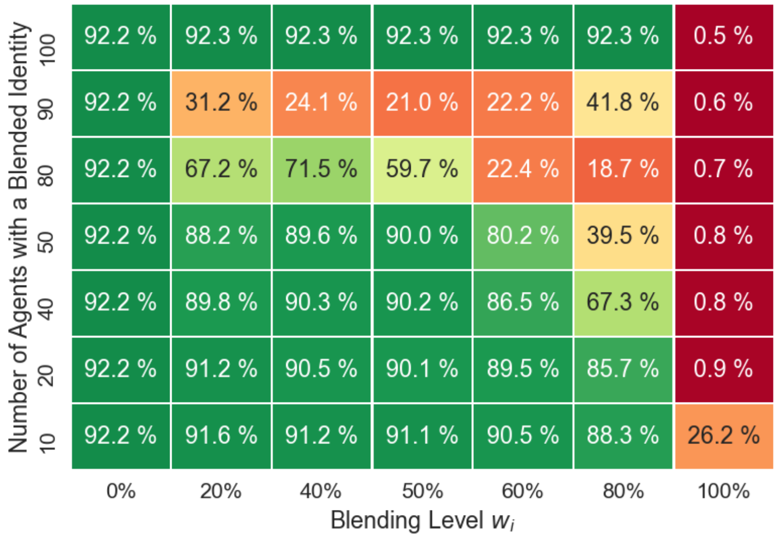
Figure 6.3: An example of the distribution of action rules by subsequent, but not consecutive, generations for the set of 50 agents having a blending level of $w_i = 80\%$ shows that defectors, $(0, 0, 0)$, quickly dominate the population within the first 20 generations, resulting in an average cooperation of 10.6%. This is a result of defectors exploiting the shared reputation.

6.4.2 Controlling the Probability of Blended Agents Interacting With Other Blended Agents

In this section, parameter S is considered to investigate in-group interactions of blended agents. Parameter S (originally introduced in Section 3.5.1) can be applied to govern the probability of a blended agent interacting with another blended agent (i.e., another agent with some identity drawn from the group). This aligns with the concept of in-group attraction, where agents with something in common have homophilic attraction (see Section 3.5.1). We study this while maintaining the assumptions of Section 6.4.1 - that is we continue to assign agents with the same blending levels and these blending levels do not evolve (i.e., exogenous blended agents). The experimental results allow us to analyse the impact of parameter S on blended identities and how biasing the scope of interactions can impact cooperation.

In this experiment, we re-define parameter S to control the extent that in-group interactions occur, using blending as the basis for the in-group. Parameter S controls the interactions of blended agents using a probability equal to their blending level. That is, when the recipient agent j is blended (i.e., $w_j > 0$), agent j attempts to select the potential donor agent i from the set of blended agents $\{k : w_k > 0\}$ with $k \neq j$ and with a probability $S = w_j$. With probability $1 - S = 1 - w_j$, agent j attempts to select the potential donor agent i from the set of non-blended agents $\{k : w_k = 0\}$. When the recipient agent j is non-blended ($w_j = 0$), agent j selects the donor agent i from the population at random with disregard to their blending level.

Figure 6.4 shows that the introduction of parameter S lowers cooperation overall and is particularly evident when comparing the results of 90 agents sharing a trait, as contrasted with the analogous scenario in Figure 6.2. The lower cooperation can be attributed to parameter S because S increases the interaction between the number of agents who have co-dependency on the same shared trait. This opens up opportunities for shirkers which eventually deters cooperation as observed in Section 4.4. Even when the number of sharing agents is low but S is high, cooperation decreases significantly, ultimately achieving just 0.9% when the number of blended agents is 20 and $S = w_i = 1$, i.e. when limiting the interaction of trait sharing agents to only other trait sharing agents.



	0%	20%	40%	50%	60%	80%	100%
100	92.2 %	92.3 %	92.3 %	92.3 %	92.3 %	92.3 %	0.5 %
90	92.2 %	31.2 %	24.1 %	21.0 %	22.2 %	41.8 %	0.6 %
80	92.2 %	67.2 %	71.5 %	59.7 %	22.4 %	18.7 %	0.7 %
50	92.2 %	88.2 %	89.6 %	90.0 %	80.2 %	39.5 %	0.8 %
40	92.2 %	89.8 %	90.3 %	90.2 %	86.5 %	67.3 %	0.8 %
20	92.2 %	91.2 %	90.5 %	90.1 %	89.5 %	85.7 %	0.9 %
10	92.2 %	91.6 %	91.2 %	91.1 %	90.5 %	88.3 %	26.2 %

Figure 6.4: The average cooperation recorded over generations recorded as a result of parameter $S = w_i$ with different starting configurations. In this scenario, S is based on the current blending level of the blended agents. Parameter S causes a decrease in cooperation overall when compared to Section 6.4.1.

6.4.3 Influencing Reproduction for Blended Agents

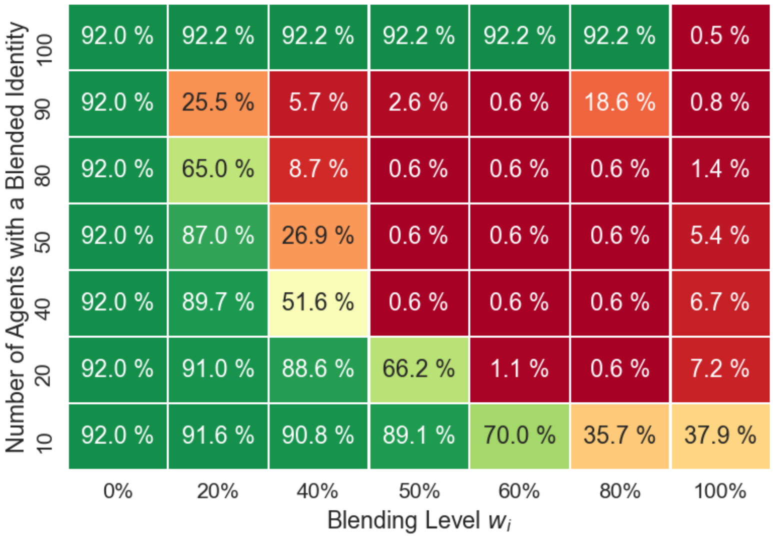
In this Section, we analyse influencing reproduction through an agent's blending level. This experiment allows us to determine whether limiting reproduction within the group of blended agents sustains cooperation and deters free-riders. Reproduction is the step that allows agents to socially learn from others by copying their action rules (see Section 3.5.3). In the following experiment, at the reproduction step agents either copy their action rules from other blended agents (with probability $P = w_i$) or from the wider population (with probability $P = 1 - w_i$). Therefore parameter P controls whether agents copy other blended agents or the wider population. In other words, we allow an agent's blending level, w_i, to control who that agent copies from. Agents have a higher chance of copying independent (non-blended) agents when P is low and this probability diminishes as P

increases. Again, this assumption aligns with the possibility that agents could be more likely to be influenced by their “in-group”, that is others who have shared identity with the group.

In addition to controlling reproduction through blending, this experiment maintains the assumptions of Section 6.4.1 - that is we continue to assign agents’ with the same blending levels and these blending levels do not evolve. Additionally, parameter S is used in the same way introduced in Subsection 6.4.2, in which S controls the probability that the pair chosen for an interaction are both blended agents. This arrangement allows us to examine the progressive effects of adding different model components aligned with in-group identity.

Figure 6.5 shows that when blended agents restrict their interactions and inheritance to other blended agents, cooperation decreases. Particularly noteworthy is the additional effect that reproduction has on decreasing cooperation when it is influenced by blending. This is seen by comparing Figures 6.5 and 6.4.

It appears that when agents are influenced by their blending for selection of action rules, the agents with high blending levels begin to reinforce defective (i.e., shirking) strategies by being disproportionately disposed to copying such strategies. This has the effect of propagating defective strategies which limit the overall prospects for high levels of cooperation. The contrast between Figures 6.5 and 6.4 is quite stark in this regard, particularly when blending levels are 50% or greater. Modest levels of mutation are insufficient to counter the spread of defective strategies, leaving the system effectively trapped with a preference for in-group and defective interactions. As evidenced by Figure 6.6 the rapid spread of defective strategies is not countered as they dominate the population.



Number of Agents with a Blended Identity	0%	20%	40%	50%	60%	80%	100%
100	92.0 %	92.2 %	92.2 %	92.2 %	92.2 %	92.2 %	0.5 %
90	92.0 %	25.5 %	5.7 %	2.6 %	0.6 %	18.6 %	0.8 %
80	92.0 %	65.0 %	8.7 %	0.6 %	0.6 %	0.6 %	1.4 %
50	92.0 %	87.0 %	26.9 %	0.6 %	0.6 %	0.6 %	5.4 %
40	92.0 %	89.7 %	51.6 %	0.6 %	0.6 %	0.6 %	6.7 %
20	92.0 %	91.0 %	88.6 %	66.2 %	1.1 %	0.6 %	7.2 %
10	92.0 %	91.6 %	90.8 %	89.1 %	70.0 %	35.7 %	37.9 %

Figure 6.5: The average cooperation recorded over generations as a result of agents reproducing based on their blending levels while varying both the starting configuration and the number of agents with a blended identity as described in Section 6.4. The results indicate that cooperation decreases as a result of restricting blended agents interactions and inheritance. Note that in this figure parameter $S = P = w_i$.

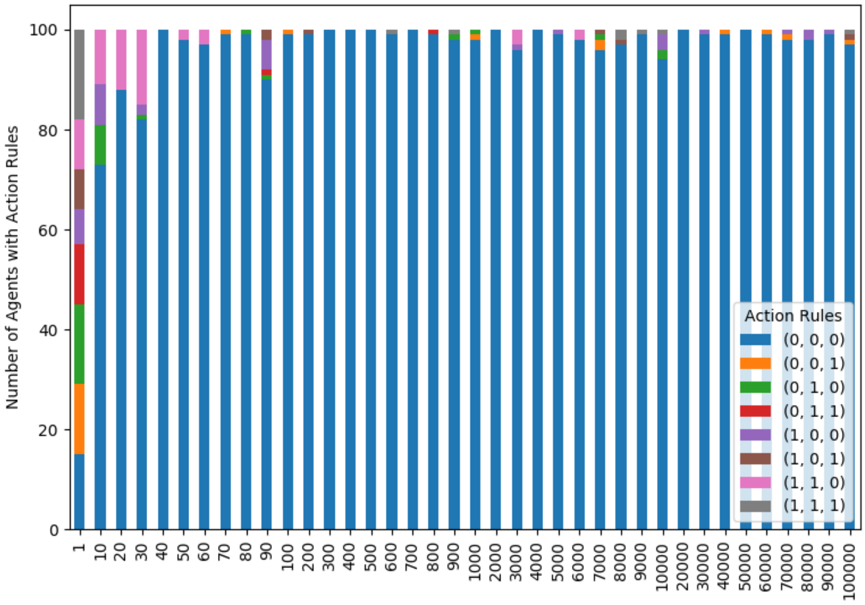


Figure 6.6: An example of the distribution of action rules by subsequent, but not consecutive, generations for the set of 20 agents having a blending level of $w_i = 80\%$ shows that defectors $(0, 0, 0)$ quickly dominate the population within a few generations, causing cooperation to collapse. The result is an effect of $S = 0.8$ and the probability of blending agents copying from other blending agents being $P = 80\%$ as well. This leads to defectors quickly exploiting the shared reputation.

6.5 Allowing Agents to Evolve Their Blended Identity

In this section, we consider the effects of allowing agents to evolve their blending levels. In contrast to Section 6.4 this allows the structure of the population to change, in terms of the extent of group involvement. By enabling agents to change their blending level, we can determine the effect of the agents growing or diminishing their allegiance to the group (or specifically the trait representing the group). At the end of every generation, we assume that the agents inherit both the action rules and the blending levels of their selected agents. Furthermore, a mutation on action rules and on blending level is applied with a probability of 1% as discussed in Section 6.4. Algorithm 4, presents a framework of how the model works and how blending levels evolve after every generation.

This section makes several assumptions. To begin, we divide the population into blended agents ($w_i > 0$) and independent agents. We assign each blended agent i a blending level, w_i, at random from the levels identified in Subsection 6.3.2. A subset of agents are designated as independent and are assigned $w_i = 0$. Additionally each agent i is assigned an action rule, (s_i, u_i, d_i) at random, as is our convention for all experiments in this thesis. The assigned action rules and blending levels remain constant within a generation for each agent. During each round within a generation, the agents interact making donation decisions using their action rules as usual (i.e., as discussed in Chapters 3), updating their reputation and payoff as outlined in Subsection 6.3. Each choice of agent pair in any interaction is controlled by parameter S.

Additionally, we analyse the effects of allowing agents to evolve their blending level, w_i, at the reproduction step (lines 24-39 in Algorithm 4). Once a generation of interactions is complete, agents socially learn from others at the reproduction step based on their relative success as measured by their payoff. Once an agent updates their action rules (line 27) and their blending level (line 28), mutation takes place. Mutation changes an agent's action rules with probability $\mu_A = \frac{1}{100}$ (lines 29-31) and blending level with a probability of $\mu_B = \frac{1}{100}$ (lines 32-34), the rate being sufficient to alter the population's structure.

Mutation randomly assigns agents one of seven blending levels identified in Subsection 6.3.2, including $w_i = 0$, and one of the eight action rules specified in Subsection 3.4.3. Note that mutation is applied to both independent and blended agents.

Algorithm 4 Algorithm for Indirect Reciprocity Based on the Reputation of Traits and Evolving Blended Identities

```
Require: Number of iterations $m$; number of generations $M$; set of agents $A$; set of traits
    $T$; set of traits $T$; set of binary action rules $AR = (s_i, u_i, d_i)$; set of blending levels
    $W$; cost $c$; benefit $b$; in-group interaction probability $S$; mutation rate of action rules
    $\mu_A$; mutation rate of blending levels $\mu_B$; the set of agents that share at least one trait
    with agent $j$ $N_j$;
for $M$ generations do                                        ▷ Evolutionary simulation
    Set $f_i = 0\ \forall i \in A$ and $r_t = 0\ \forall t \in T$
    for $m$ iterations do
        $j \leftarrow U(A)$                                   ▷ Select recipient (see Section 3.5.1)
        $p \leftarrow U(0,1)$
        if $p < S$ and $|N_j| > 0$ then
            $i \leftarrow U(N_j)$                             ▷ Select random in-group donor
        else if $p \geq S$ and $|\bar{N}_j| > 0$ then
            $i \leftarrow U(\bar{N}_j)$                       ▷ Select random out-group donor
        else
            $i \leftarrow U(A - \{j\})$                       ▷ Select random donor
        end if
        ▷ Apply action Rules (see Sections 3.4.3 and 3.5.2)
        $r^i = (1 - w_i)r(t_i) + w_i r(g)$                    ▷ $i$'s reputation is derived from $w_i$
        if ($r^i$=$r^j$ and $s_i = 1$)                        ▷ Compare equal
                or ($r^i < r^j$ and $u_i = 1$)                ▷ Compare upwards
                or ($r^i > r^j$ and $d_i = 1$) then           ▷ Compare downwards
            $r_t \leftarrow min(5, r_t + 1)$                  ▷ $i$ donates, increase reputation
            $f_i \leftarrow f_i - c$; $f_j \leftarrow f_j + b$    ▷ Update fitness
        else                                                  ▷ $i$ defects
            if $r^j \geq r^i$ then                            ▷ Detect unjustified defection.
                $r_t \leftarrow max(-5, r_t - 1)$             ▷ Decrease reputation
            end if
        end if
    end for
    ▷ Reproduction stage (see Section 3.5.3)
    for $i \in A$ do
        $j \leftarrow R(A, f)$                                ▷ Roulette wheel based on fitness
        $(s'_i, u'_i, d'_i) \leftarrow (s_j, u_j, d_j)$       ▷ $i$ copies $j$'s action rules
        $w'_i \leftarrow w_j$                                 ▷ $i$ copies $j$'s blending level
        if $U([0,1]) < \mu_A$ then
            $(s'_i, u'_i, d'_i) \leftarrow U(AR)$             ▷ Mutate action rules
        end if
        if $U([0,1]) < \mu_B$ then
            $w'_i \leftarrow U(W)$                            ▷ Mutate blending level
        end if
    end for
    for $i \in A$ do
        $(s_i, u_i, d_i) \leftarrow (s'_i, u'_i, d'_i)$       ▷ Update action rules for all agents
        $w_i \leftarrow w'_i$                                 ▷ Update blending levels for all agents
    end for
end for
```

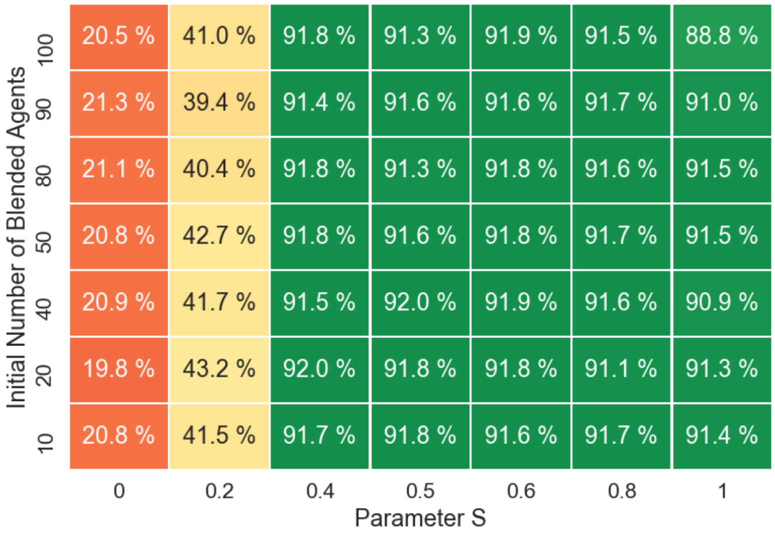
Figure 6.7 shows that an increase in parameter S leads to an increase in cooperation. We note that S is the probability that a donor is a blended agent in the case that the recipient is a blended agent, i.e. if an agent is dependent, then S is the probability that their donor is also a dependent agent. When S is low (i.e., 0 & 0.2), cooperation averages less than 50% regardless of the initial number of blended agents. In contrast, as S increases or the number of interactions between blended agents increases, so does cooperation. Both these results are attributed to the level of blending in the population. In particular, *when blending levels are low, cooperation averages high levels, and when blending levels increases cooperation decreases* as can been seen in Figure 6.11.

Low cooperation is in line with that of Figure 4.3 in Section 4.4, where cooperation declines as the number of dependent agents increases. In Section 4.4, the number of dependent agents is much lower than 10% for high S whereas for $S = 0$ it can be as high as 20%. However, in this section, agents are not restricted to $w_i = 0$ or $w_i = 1$, and the number of agents that are blended is not fixed as agents may evolve their blending levels through copying the parent agent's level or mutation, as such a higher number of dependent agents may exist in the population.

The high cooperation recorded when S is high is attributed to the population's low average blending level, as shown in Figure 6.10. Initially, the evolution of blending presents an opportunity for shirkers to exploit reputations that they share with highly blended cooperators. As agents adopt defective action rules, as can be seen in Figure 6.12, cooperation collapses and payoff reaches low levels. Consequently, agents begin to evolve towards low blending level achieving $w_i = 0$, as can be seen in Figure 6.13. As agents evolve towards low levels of blending, they become independent agents, and as such, they cease to share reputation with other agents. Interactions between independent agents quickly expose any shirkers, virtually eliminating them within a few generations, as shown in Figure 6.12, which in turn promotes cooperation.

With lower levels of S, a cycling behaviour occurs where initially a larger number of agents can increase their blending level while sustaining cooperative action rules, in line with the findings of Figure 4.3 in Section 4.4. As the number of agents increases, so does the number of shirkers leading cooperation to collapse, leading in sequence to agents becoming uncooperative and later adopting a blending level of $w_i = 0$. This cycle is repeated as can be seen in the example illustrated by Figures 6.8 and 6.9. As a result of this cycling behaviour where defectors are reintroduced to the population, cooperation records between $20\% - 43\%$ on average as can be seen in Figure 6.7.

These results indicate a negative correlation between blending levels and cooperation, which shows an increase in blending levels, even if minimal, deters cooperation, $r(49) = -.98$, $p < .001$. In contrast, the lack of blending leads to higher cooperation, as seen in Figure 6.11. These results are in line with those of Subsection 6.4.1, in which higher blending levels among blended agents caused cooperation to drop.



	0	0.2	0.4	0.5	0.6	0.8	1
100	20.5 %	41.0 %	91.8 %	91.3 %	91.9 %	91.5 %	88.8 %
90	21.3 %	39.4 %	91.4 %	91.6 %	91.6 %	91.7 %	91.0 %
80	21.1 %	40.4 %	91.8 %	91.3 %	91.8 %	91.6 %	91.5 %
50	20.8 %	42.7 %	91.8 %	91.6 %	91.8 %	91.7 %	91.5 %
40	20.9 %	41.7 %	91.5 %	92.0 %	91.9 %	91.6 %	90.9 %
20	19.8 %	43.2 %	92.0 %	91.8 %	91.8 %	91.1 %	91.3 %
10	20.8 %	41.5 %	91.7 %	91.8 %	91.6 %	91.7 %	91.4 %

Figure 6.7: The average cooperation recorded over generations as agents evolve their blending levels. Cooperation is impacted by agents evolving their blending levels while enforcing parameter S, particularly when parameter S is low, such as when $S = 0$ and 0.2. Note that the initial number of blended agents is randomly assigned a blending level w_i.

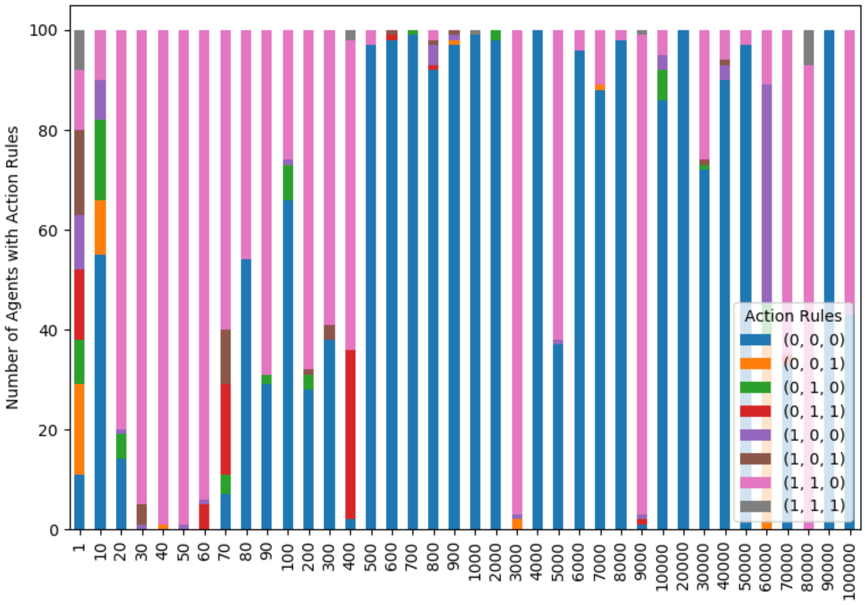


Figure 6.8: An example of the distribution of action rules by subsequent, but not consecutive, generations where agents are allowed to evolve their blending levels and $S = 0$. The scenario starts with 10 blended agents and 90 independent agents and shows a struggle between discriminators $(1, 1, 0)$ and defectors $(0, 0, 0)$ resulting in 20.8% cooperation as can be seen in Figure 6.7.

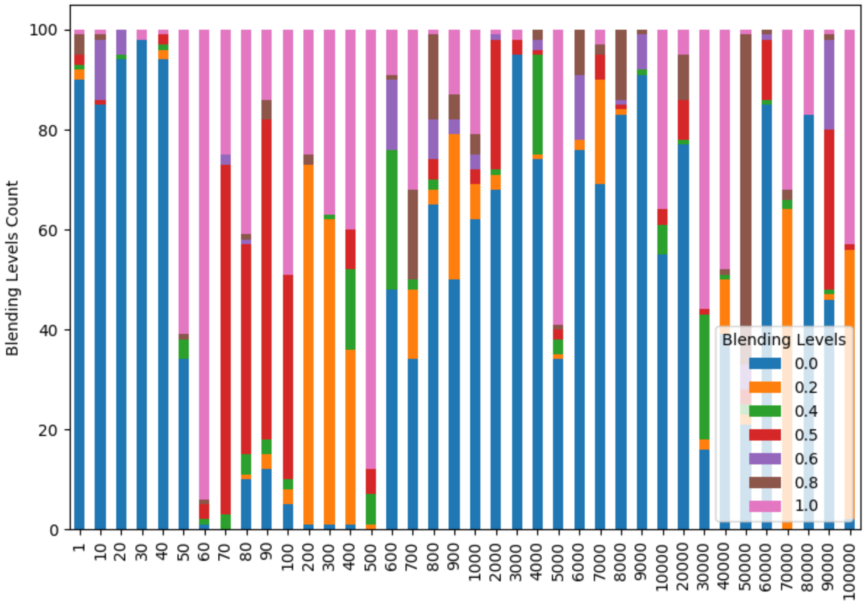
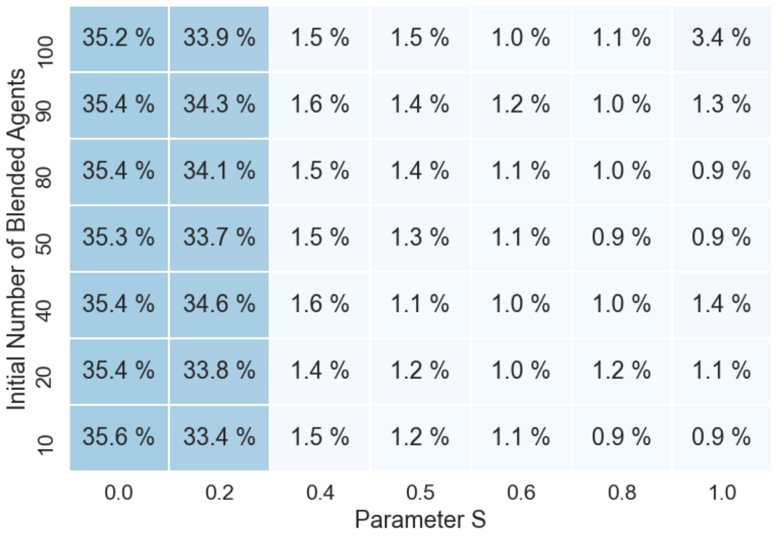


Figure 6.9: A snapshot of the distribution of blending levels by subsequent, but not consecutive, generations where agents are allowed to evolve their blending levels and $S = 0$. The scenario starts with 10 blended agents and 90 independent agents and shows a cycling behaviour between low and high blending levels.



Initial Number of Blended Agents	0.0	0.2	0.4	0.5	0.6	0.8	1.0
100	35.2 %	33.9 %	1.5 %	1.5 %	1.0 %	1.1 %	3.4 %
90	35.4 %	34.3 %	1.6 %	1.4 %	1.2 %	1.0 %	1.3 %
80	35.4 %	34.1 %	1.5 %	1.4 %	1.1 %	1.0 %	0.9 %
50	35.3 %	33.7 %	1.5 %	1.3 %	1.1 %	0.9 %	0.9 %
40	35.4 %	34.6 %	1.6 %	1.1 %	1.0 %	1.0 %	1.4 %
20	35.4 %	33.8 %	1.4 %	1.2 %	1.0 %	1.2 %	1.1 %
10	35.6 %	33.4 %	1.5 %	1.2 %	1.1 %	0.9 %	0.9 %

Figure 6.10: The average blending levels recorded over generations as agents evolve their blending levels. Blending levels are impacted by parameter S as lower levels of S allow for higher levels of blending.

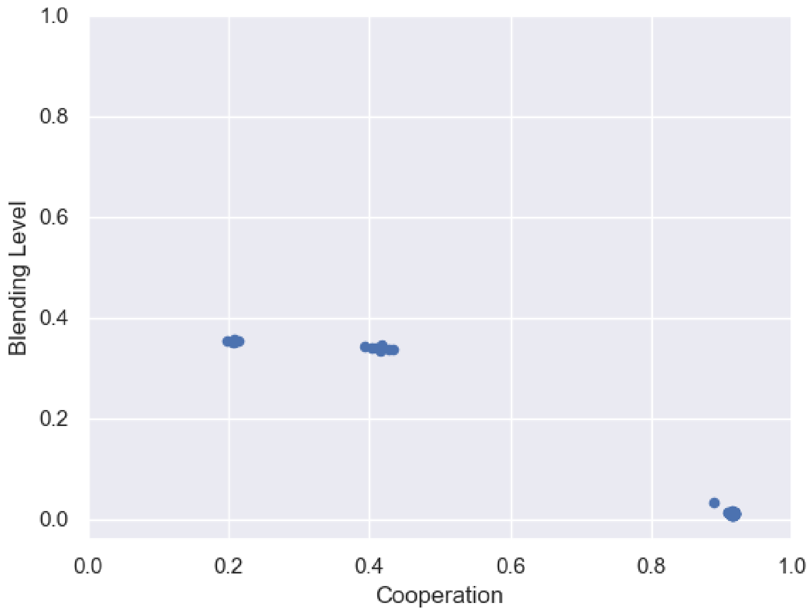


Figure 6.11: The correlation of blending levels and cooperation when agents evolve their blending levels as described in Section 6.5. The correlation represents the cooperation recorded in Figure 6.7 with the blending levels recorded in Figure 6.10. Each point represents a pair of (blending level, cooperation) from the two Figures.

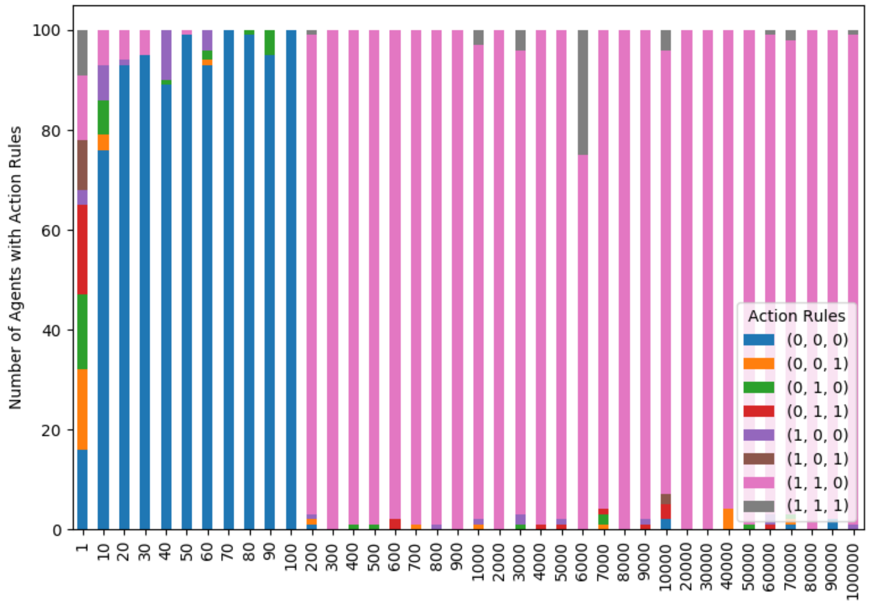


Figure 6.12: A snapshot of the distribution of action rules by subsequent, but not consecutive, generations where agents are allowed to evolve their blending levels and $S = 1$. The scenario starts with 10 blended agents and 90 independent agents and shows a a dominance for discriminators $(1, 1, 0)$ resulting in 91.4% cooperation when $S = 1$ as can be seen in Figure 6.7.

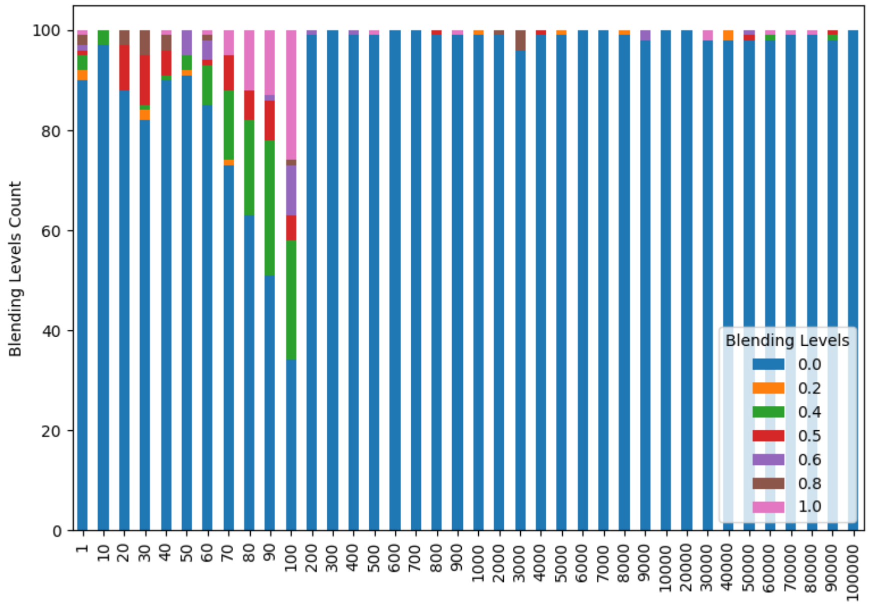


Figure 6.13: A snapshot of the distribution of blending levels by subsequent, but not consecutive, generations where agents are allowed to evolve their blending levels and $S = 1$. The scenario starts with 10 blended agents and 90 independent agents and shows that agents evolve towards $w_i = 0$.

6.6 Discussion

This chapter's focus on how identities can blend together is a shift from previous chapters where the focus was on how individuals may share identities through traits. In this chapter, we broke down identity further and looked at the makeup of an individual's identity. Alternatively, in previous chapters sharing a trait meant that an agent fully shares their reputation with others, in this chapter agents blended their identity to only share a part of a trait.

Our results have shown that allowing agents to blend their personal and shared identities can be disruptive to cooperation. Our results in Section 6.4 showed that while cooperation is achieved when blending levels are low, once blending levels increase cooperation plummets. The outcome can be attributed to two main reasons. Firstly, low cooperation results from defector agents infiltrating the population and taking advantage of the shared reputation. Secondly the current reproduction process allows for the defector strategy, $(s_i, u_i, d_i) = (0, 0, 0)$, to take hold as it encourages agents to gain a high payoff. In Subsection 6.4.1, the increase in the number of blended agents and the increase in their blending level allowed defectors to exploit the shared reputation, which caused cooperation to plummet in return. In Subsection 6.4.2, the introduction of parameter S caused a further collapse to cooperation as blended agents could not interact with independent agents. This phenomenon was proven further in Subsection 6.4.3 when agents only inherited from other agents who were blended as can be seen in Figure 6.5. As blended agents bound their interactions and their inheritance to other blended agents, their cooperation decreased dramatically. These results are limited as agents were not permitted to evolve their blending levels.

In Section 6.5, we allowed agents to evolve their blending levels. Our results in Section 6.5 show that cooperation recovers from the levels achieved in Section 6.4. The outcome was credited to agents being allowed to evolve their blending level. Cooperation was impacted by parameter S, in particular, higher values of S produced high cooperation, and in correlation, lower levels of blending were recorded. However, with lower values of S high, blending levels were recorded, allowing defectors to take advantage of the shared reputations. The presence of defectors caused cooperation to achieve an average of below 43%.

6.7 Conclusions

This chapter has introduced the blending of personal and shared identities to form an individual's identity. A novel method to calculate reputation was assessed, as a result of the blended identity. To accommodate these concepts, we updated the model of indirect reciprocity. The updated model allowed individual agents to form their identity by partially subscribing to a shared trait. This concept contrasts with previous chapters, where single or multiple traits fully formed an individual's identity.

The introduction of a blended identity allowed us to create agents with a single identity formed of both a personal and a shared trait. A balance between the two traits enabled agents to hold unique elements that do not get compromised by their shared characteristics. This chapter's results are consistent with those of previous chapters where the increase in the number of agents sharing a trait caused cooperation to decline. Additionally, the increase in in-group interactions caused cooperation to fall. A balance between the number of agents in a group and the number of in-group interactions was needed to maintain decent cooperation levels (above 40%). Similarly, in this chapter, we found that high blending levels cause cooperation to decrease, and an increase in parameter S has the same effect. However, when agents have freedom to evolve both their action rules and blending levels they can maintain reasonable cooperation levels and do so by reducing the extent of identity sharing. This chapter reaffirms that balance between identities is needed for cooperation to emerge and be sustained.

We note that these findings on identity assume that economic factors drive the evolutionary process. Specifically, reproduction at the end of each generation restructures the population based on weighting candidates to be copied based on their payoff. This represents measuring success as an agent's accrued benefits less their costs. This approach proclaims that economic wealth is the primary precursor to survival. Indeed, this is common practice in evolutionary games since the works of John Maynard Smith [162] [163]. However, one may question this assumption when identity comes in to play, because identity can provide intrinsic value to an individual (e.g., being seen to belong to a prestigious group). This is reaffirmed by identity fusion [175, 177], where seemingly counter-intuitive selfless individuals acts can emerge because individuals value and fuse with the group's identity. Therefore it is prudent to further explore the impact of identity and how it might influence an individuals preferences for selection.

Accordingly, in the following chapter, we allow agents to reproduce taking into account their preferences for the extent of group identity that they may see in others. This novel motivation allows us to investigate how agents cooperate when they are motivated by an agent's value in respect of an identity rather than economic benefits. This motivation represents a shift from the conventional view that success is measured entirely by an individual's monetary wealth (payoff) to valuing their social success and how they are identified in society. The concept is therefore significant and has not been previously modelled, to the best of our knowledge.

Chapter 7

Modelling Intrinsic Incentives - Fusion Motivation

7.1 Introduction

This chapter introduces identity as a personal motivation for decision making at the reproduction step and refers to this incentive as *fusion motivation*. This captures the idea that individuals who derive some identity themselves from a group may place value or respect in those who are doing the same. It is aligned with homophily (i.e., attraction to those with similar characteristics) and contrasts with current conventions where purely economically rational decisions are typically modelled.

Our modelling to this point in the thesis has allowed agents to evolve using selection criteria based on economic success. In other words, an agent is deemed successful if they have a high payoff, i.e. having received higher benefits and/or lower costs due to their choice of action rules (i.e., strategy). To evolve, agents copy the action rules of those deemed economically successful using payoff. One of the problems with this assumption is that it presupposes that agents are perfectly rational [40] and pursue the choice that will optimise their future resources. In reality many experiments show human behaviour deviates substantially from this, and that for example, human decision making has bounded rationality [112]. This may be due to external factors (e.g., time to assess everything in full to make a decision), or internal factors, where personal motivations or world view shape the decision-making criteria of the individual.

In the case of indirect reciprocity, our argument for considering identity as an alternative to perfectly rational behaviour stems from internal motivations. Identity becomes a significant internal issue for an agent when it derives a personal reputation from alignment with a group's identity, and where the group's identity is shared with others. The most significant evidence for this being a plausible assumption lies with the relatively recent concept of identity fusion [177], that presents a framework through which an individual values a group's identity as a component of their own personal identity. This makes the group more important to the individual, and it effectively becomes an extension of the individual. The chapter is structured into subsections as follows. In Section 7.2, we introduce the concept of fusion motivation based on an agent's identity and discuss its relation to the literature on identity in psychology, as mentioned in Section 2.3. Section 7.3 compares the traditional reproduction approach of economic motivation, previously introduced in Sections 3.4 and 3.4.2, with the proposed approach that is driven by the individual's blended identity and is referred to as fusion motivation. Subsection 7.3.1 introduces the chapter's experiments by presenting different scenarios where fusion motivation can be applied and overviews the experiments' assumptions.

The first scenario presented in Section 7.4 applies fusion motivation to all agents. The scenario enables us to determine how agents cooperate when their motivation is not payoff but agents' devotion to their group as measured by blending level. The scenario is similar to that of Section 6.5 but utilises fusion motivation for all agents in place of economic motivation. In Section 7.5, we allow agents' blending level to determine their motivation based on a probability. The scenario is the first that allows agents to their motivations. By equating an agent's motivation to their blending level, we allow identity to be a determinant factor for agents' reproduction. Section 7.6 introduces evolution to motivation by allowing agents to probabilistically copy the motivation of others in the same manner that they copy their action rules and blending levels. We summarise the chapter findings in a discussion in Section 7.7.

7.2 Identity-Driven Motivation in Psychology

As mentioned in Chapters 2 and 3, traditionally indirect reciprocity models utilise payoff (the difference between costs and benefits) as fitness to determine which agents reproduce. This chapter challenges that view by aligning fitness to group identity, for those agents who have a case to align with such a disposition through their dependency on the group for their personal identity. This builds on the structure of identity introduced in Chapter 6, where agents combine a shared group identity with an individual identity using the idea of a blending level. When the blending level is high, the agent has more connection to the group (and their strategies) due to greater dependency on their identity being shared with others. Here we explore what happens when strength of the shared group identity presents itself as an incentive to prioritise copying another agent at the reproduction step.

This incentive relates to the concept of identity driven motivations from psychology [177]. As described in Section 2.3, psychologists have suggested that identity is an important driver for prosocial behaviour, with it providing powerful intrinsic motivations [7]. Intrinsic motivation refers to the personal value that the individual associates with a particular activity; this could be the feeling of contentment and fulfilment that they associate with an act, for example [157]. In our context, intrinsic motivation means that a reward's value may not be represented by explicit factors which are received by an individual undertaking a task [19]. Payoff, as represented in previous chapters, is an example of an extrinsic reward. In contrast, identity fusion [177] presents an intrinsic motivation back to the individual. This occurs through the identity of the individual overlapping with the identity of the group, which also incentivises the individual to act for the group. In qualitative terms, the concept of identity fusion has also been described in terms of feelings that influence pro-group decision-making, such as providing a "visceral sense of oneness with a group" [176] that in-turn reaffirms supporting the group's behaviour. For the models of this chapter, identity fusion can be interpreted as supporting or promoting the action rules of influential (i.e., highly blended) group members.

While it is a challenge to model visceral feelings, it is possible to model the hypothetical consequences of those feelings in the decisions that agents take at the reproduction step. From this perspective agents will not reproduce only based on an agent's economic value but rather by taking into account their identity and how devoted they are to the group identity. We extend the evolutionary game theoretic framework to incorporate prioritisation of copying the action rules of other group members, alongside the economic motivation of copying the action rules of those that are most successful in terms of payoff. This is described in detail in the following sections.

7.3 Implementing Fusion Motivation

In this section, we highlight how fusion motivation can extend the general framework as presented in Chapter 6. We assume that fusion motivation affects the selection process within the reproduction phase of the framework. Reproduction is the social learning step that occurs at the end of every generation. In this step, agents probabilistically copy the action rules of other agents, taking into account the success of other agents, as perceived by an agent at the end of every generation. This is the element of the model that we adjust.

We refer to the two alternative motivations at the selection stage as *economic motivation* and *fusion motivation*. Economic motivation prioritises having more benefits than costs or having the highest payoff. This represents an extrinsic motivation. Agents with this motivation probabilistically copy others strategies based on the differences between their benefits and costs (i.e., payoff). Alternatively we assume that *fusion motivation* prioritises copying agents based on the extent of their commitment towards the shared identity of the group, as measured by their blending level. This represents an intrinsic motivation, providing positive feedback from copying a blended when the agent is strongly aligned to the group, in terms of identity. Agents with this motivation copy other agents' strategies, probabilistically based on the extent of other agent's blending level towards the shared trait.

7.3.1 Fusion Motivation Scenarios

In the following sections, we consider different fusion motivation scenarios. The scenarios allow us to understand to what extent fusion motivation disrupts population structures and its impact on identity and cooperation. We focus on three different scenarios that implement fusion motivation. The chosen scenarios complement experiments in past chapters to draw on their results for comparison.

The scenarios are divided into sections as follows.

- In Section 7.4, we apply fusion motivation to all agents in the population enabling us to determine how agents cooperate when their reproduction is not determined through payoff but through the agents' regard for fusion to the group, as measured by blending levels. The scenario compliments Section 6.5 of Chapter 6 as it uses the same assumptions except for the motivation used in the reproduction step.
- In Section 7.5, we allow the agents' blending levels to determine their motivation based on a probability equal to their level. This scenario allows for both motivations to exist in the population structure and explores how the combination impacts cooperation and identity.
- In Section 7.6 we introduce evolution to motivation, allowing agents to change their motivation after every generation as they are copied from others who are deemed successful under their current motivation. In this scenario, motivations are based on probability rather than being a binary choice between economic and fusion motivation.

These experiments share several assumptions. Each experiment uses the default parameters mentioned in Section 3.5.3. That is all the experiment results are obtained from five runs, each with a random seed. Each agent participates in an average of 50 games per generation based on the following parameters $A = 100$, with $M = 100000$ and $m = 5000$

where A is the population size, M is the number of generations, and m is the number of rounds per generation.

To begin, we divide the population into blended agents and independent agents. We assign each blended agent i a blended level, w_i, at random from the levels identified in Subsection 6.3.2. Additionally each agent i is assigned an action rule, (s_i, u_i, d_i) at random. In other words, a set of agents are designated as independent and will be assigned $0 < w_i = 0$, while the remaining agents are randomly assigned a blending level where $w_i > 0$. The assigned action rules and blending levels remain constant within a generation for each agent. During each round within a generation, the agents will interact, taking donation decisions using their action rules as usual and update their reputation and payoff as outlined in Subsection 6.3. Each agent pair chosen in any interaction is controlled by parameter S, which is the probability that the recipient j attempts to choose a blended agent i as donor (i.e. $1 > w_i > 0$). $1 - S$ represents the probability that the chosen donor is an independent agent (i.e. $w_i = 0$). Non-blended agents, ($w_i = 0$), select donors at random from the population with disregard to their blending levels.

Once a generation is completed, reproduction takes place. The following experiments have different approaches for that phase which are detailed in their specific sections. The different approach to reproduction is necessary to examine how identity and fusion motivation affects different scenarios.

7.4 Evolving All Agents With Fusion Motivation

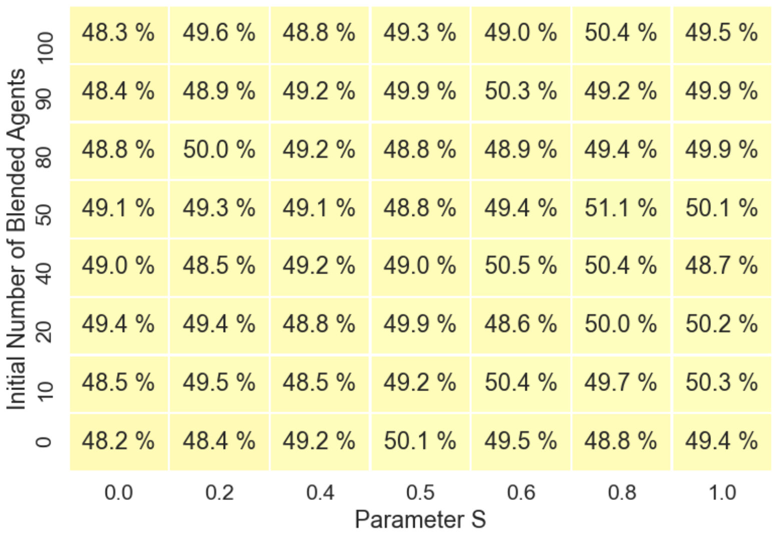
In this Section, we consider the effects of allowing agents to evolve their blending levels while using only fusion motivation. Similar to Section 6.5, this allows the population structure to change in terms of the extent of group involvement. The use of fusion motivation enables agents to reproduce based on the highest blending agent without considering payoff (economic motivation). Fusion motivation is applied on all agents regardless of their initial blending level (lines 24-34 in Algorithm 5). At the end of every generation, we assume that the agents inherit both the action rules (line 26 in Algorithm 5) and the blending levels of successful agents based on that generation's highest blended agents (line 27 in Algorithm 5). Furthermore, a mutation on action rules and on blending level is applied with a probability of 1% as discussed in Section 6.4.

This experiment is similar to that of Section 6.5; however, in this experiment, fusion motivation is used instead of economic motivation. In addition to adopting the assumptions identified in Section 7.3.1, specific parameters are adjusted for this scenario. These begin at the reproduction stage, in which agents socially learn based on fusion motivation. In other words, agents learn from others based on their relative success measured by their blending level. Once an agent updates their action rules and their blending level, mutation takes place. Mutation changes an agent's action rules and blending level with a probability of $\frac{1}{100}$, the rate being sufficient to alter the population's structure as demonstrated in [198] and [199] and used in previous experiments in this chapter. Blending level mutation randomly assigns agents one of seven blending levels identified in Subsection 6.3.2, including $w_i = 0$. Action rule mutation randomly assigns one of the eight action rules specified in Subsection 3.4.3. Note that mutation is applied to both independent and blended agents.

Figure 7.1 shows that fusion motivation heavily impacts cooperation regardless of the initial configuration of agents resulting in an average cooperation of $48 - 51\%$ (SD =$0.00 - 0.03$) . In this scenario, all agents are motivated to adopt high blending levels, resulting in a decrease in independent agents, as shown in Figure 7.2 (SD = 0.0001-0.027). Recall that independent agents, in this scenario, are those with a blending level of $w_i = 0$. As all agents adopt high blending levels and independent agents cease to exist in the population, agents can no longer interact with any independent (out-group) agents; therefore, parameter S has no influence on agents' interactions.

Additionally, as agents with fusion motivation base success on how highly an agent is blended to the shared trait, they do not consider payoff as an attribute for success. As such, agents adopt the action rules and blending levels of others who have high blending levels.

In turn, this results in a population with various action rules with no dominant strategy, as shown in Figure 7.3, causing cooperation to achieve an average of $48 - 51\%$. This represents near uniform random chance of cooperation. Note that even when the initial number of blended agents is 10 or, in other words, the number of independent agents is 90, agents quickly adopt high blending levels as recorded in Figure 7.4.



Initial Number of Blended Agents	0.0	0.2	0.4	0.5	0.6	0.8	1.0
100	48.3 %	49.6 %	48.8 %	49.3 %	49.0 %	50.4 %	49.5 %
90	48.4 %	48.9 %	49.2 %	49.9 %	50.3 %	49.2 %	49.9 %
80	48.8 %	50.0 %	49.2 %	48.8 %	48.9 %	49.4 %	49.9 %
50	49.1 %	49.3 %	49.1 %	48.8 %	49.4 %	51.1 %	50.1 %
40	49.0 %	48.5 %	49.2 %	49.0 %	50.5 %	50.4 %	48.7 %
20	49.4 %	49.4 %	48.8 %	49.9 %	48.6 %	50.0 %	50.2 %
10	48.5 %	49.5 %	48.5 %	49.2 %	50.4 %	49.7 %	50.3 %
0	48.2 %	48.4 %	49.2 %	50.1 %	49.5 %	48.8 %	49.4 %

Figure 7.1: Average cooperation is sustained in the range of $48\% - 51\%$ when agents use fusion motivation, and blending levels evolve and mutate. The effect of fusion motivation on the population renders Parameter S ineffective because all agents evolve their blending levels and become devoted to the group, leaving no agent outside the group. This trend occurs regardless of the starting number of blending agents, even with different implementations of S.

Algorithm 5 Algorithm for Indirect Reciprocity Based on the Reputation of Traits and Fusion Motivation

```
Require: Number of iterations $m$; number of generations $M$; set of agents $A$; set of traits
    $T$; set of traits $T$; set of binary action rules $AR = (s_i, u_i, d_i)$; set of blending levels
    $W$; cost $c$; benefit $b$; in-group interaction probability $S$; mutation rate of action rules
    $\mu_A$; mutation rate of blending levels $\mu_B$; the set of agents that share at least one trait
    with agent $j$ $N_j$;
for $M$ generations do                                              ▷ Evolutionary simulation
    Set $r_t = 0 \ \forall t \in T$
    for $m$ iterations do
        $j \leftarrow U(A)$                                 ▷ Select recipient (see Section 3.5.1)
        $p \leftarrow U(0,1)$
        if $p < S$ and $|N_j| > 0$ then
            $i \leftarrow U(N_j)$                                ▷ Select random in-group donor
        else if $p \geq S$ and $|\bar{N_j}| > 0$ then
            $i \leftarrow U(\bar{N_j})$                           ▷ Select random out-group donor
        else
            $i \leftarrow U(A - \{j\})$                                   ▷ Select random donor
        end if
        ▷ Apply action Rules (see Sections 3.4.3 and 3.5.2)
        $r^i = (1 - w_i)r(t_i) + w_i r(g)$                 ▷ $i$'s reputation is derived from $w_i$
        if ($r^i$=$r^j$ and $s_i = 1$)                                       ▷ Compare equal
              or ($r^i < r^j$ and $u_i = 1$)                               ▷ Compare upwards
              or ($r^i > r^j$ and $d_i = 1$) then                        ▷ Compare downwards
            $r_t \leftarrow min(5, r_t + 1)$                   ▷ $i$ donates, increase reputation
        else                                                                  ▷ $i$ defects
            if $r^j \geq r^i$ then                              ▷ Detect unjustified defection.
                $r_t \leftarrow max(-5, r_t - 1)$                      ▷ Decrease reputation
            end if
        end if
    end for
    ▷ Reproduction stage (see Sections 3.5.3 and 7.3)
    for $i \in A$ do                                    ▷ Determine new values for all agents
        $j \leftarrow R(A, w)$                        ▷ Roulette wheel based on blending level
        $(s'_i, u'_i, d'_i) \leftarrow (s_j, u_j, d_j)$                 ▷ $i$ copies $j$'s action rules
        $w'_i \leftarrow w_j$                                  ▷ $i$ copies $j$'s blending level
        if $U([0,1]) < \mu_A$ then
            $(s'_i, u'_i, d'_i) \leftarrow U(AR)$                            ▷ Mutate action rules
        end if
        if $U([0,1]) < \mu_B$ then
            $w'_i \leftarrow U(W)$                                     ▷ Mutate blending level
        end if
    end for
    for $i \in A$ do
        $(s_i, u_i, d_i) \leftarrow (s'_i, u'_i, d'_i)$          ▷ Update action rules for all agents
        $w_i \leftarrow w'_i$                                  ▷ Update blending for all agents
    end for
end for
```

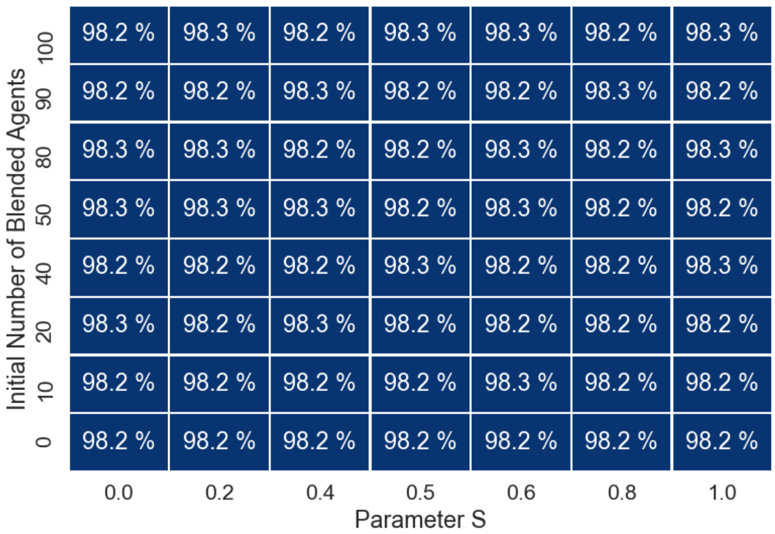


Figure 7.2: Average blending levels are above 98% as a result of all agents having fusion motivation while their blending levels evolve and mutate. Parameter S does not affect blending levels in this case, regardless of the initial number of blended agents. High blending levels are an expected result of agents having a fusion motivation.

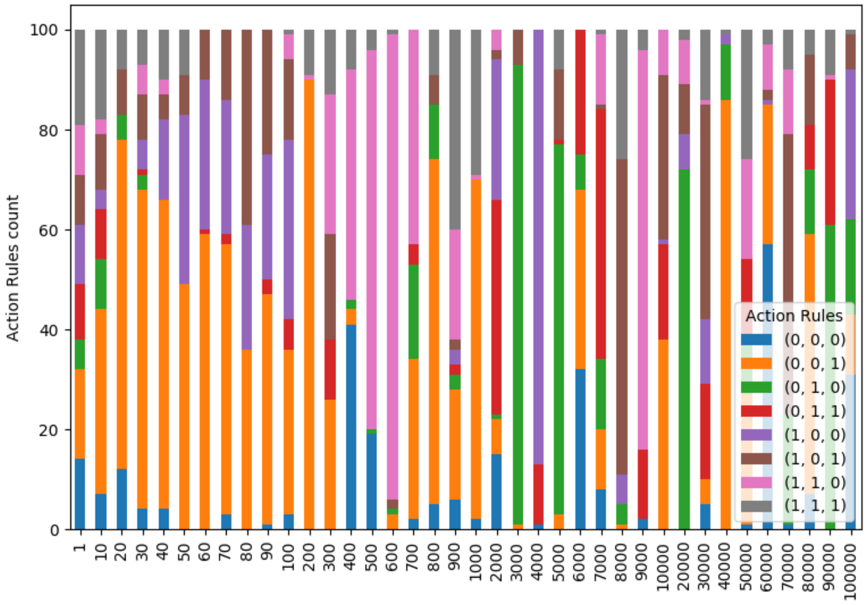


Figure 7.3: A snapshot of the distribution of the action rules produced, in subsequent but not consecutive generations, due to fusion motivation shows that fusion motivation does not allow a dominant action rule to emerge. This pattern results from agents seeking to gain high blending levels to show their devotion to their trait without regard for payoff or self-gain. In this figure the initial configuration was 50 blending agents with $S = 0.5$.

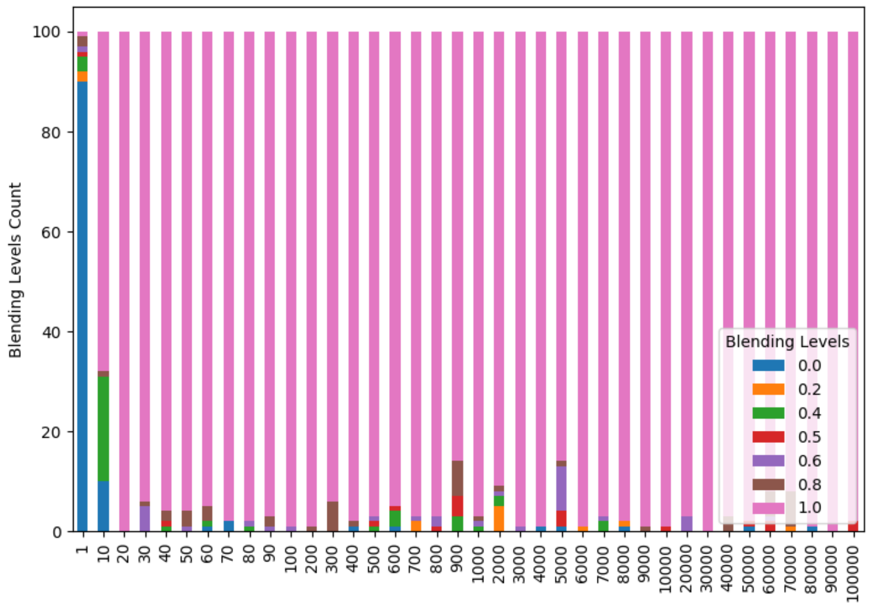


Figure 7.4: A snapshot of the distribution of the blending levels produced as a result of all agents having fusion motivation shows that agents adopt high blending levels within a couple of generations regardless of the starting configuration. In this figures, 90 agents were assigned as independent, i.e $w_i = 0$.

7.5 Using the Blending Level to Determine an Agent's Motivation

In this Section, we investigate the use of agents' blending levels, w_i, to determine an agent's motivation. The decision on whether economic motivation or fusion motivation is used by agent i for selection is governed by a probability equal to that of an agent's blending level w_i (line 25 in Algorithm 6). Specifically, fusion motivation is applied with probability w_i, and economic motivation is applied with probability $1 - w_i$. The motivation chosen is then actioned, and the agent's action rules and blending level are updated accordingly, on the basis of copying. Algorithm 6 illustrates the change in reproduction that is relevant to this scenario.

In this experiment we continue to apply mutation on action rules and on blending levels with a probability of 1% as discussed in Section 6.4. Additionally, we use different probabilities of S that enable us to determine the effects of restricting interactions between blended (dependent) agents and non-blended (independent) agents. Note that if all agents adopted economic motivation, then the results of Section 6.5 would be repeated. Similarly, if all agents adopt fusion motivation, then the results of Section 7.4 are repeated. The experiment in this section adopts the assumptions identified in Section 7.3.1.

Figure 7.6 shows that allowing agents to determine their motivation through blending levels favours fusion motivation, as the average blending level of the population is higher than 78% except where the initial number of blending agents is $0, 10$ and 20. These findings result in low cooperation overall except for the mentioned cases, as shown in Figure 7.5. The low cooperation recorded is attributed to the following. Firstly agents with high blending levels quickly adopt fusion motivation as their probability dictates them to through their blending level. This trend allows for agents with fusion motivation to copy any set of action rules blindly, which causes lower cooperation, as was seen in Section 7.4. Secondly, agents with low blending levels, and therefore with economic motivation, will copy others based on their payoff. Defector agents, within economic motivation scenarios, tend to be the agents with the highest payoff. As such, other agents with economic motivation will inherit high blending levels and defector strategies. For these reasons, fusion motivation dominates the population, which drives blending levels to high levels.

When the initial number of blended agents is low, and parameter S is higher than 0.5, the same trend is found, albeit at a slower rate. In this case, the combination of the smaller number of initial agents with a blending level greater than $w_i > 0$ and parameter $S \geq 0.5$ does not allow for blending levels to evolve to higher levels as quickly as the other cases, as observed in Section 6.5. In turn, this allows for economic motivation to exist in the population, albeit at a slower diminishing rate than when agents adopt high blending levels. In light of observing slower levels of evolution here, as compared to earlier experiments, we have extended substantially the number of generations. Figures 7.8 and 7.9 show the average cooperation and average blending levels, respectively, when the number of generations is permitted to increase from the default $100,000$ generations to a million generations. These figures suggest that the trend of higher blending levels and lower cooperation would be reached if the number of generations is allowed to increase. This also reaffirms the slowing of evolution for this scenario, as compared to previous experiments.

Figure 7.11 shows how blending levels increase when the number of generations is increased. The figure illustrates that if the number of generations is increased for all scenarios examined, fusion motivation will dominate the population, causing a high average rate of blending levels and cooperation to drop. Cooperation drops due to agents adopting action rules at random by focusing only on inheriting agents with high blending levels as can be seen in Figure 7.10.

In the first $200,000$ generations, the discriminator strategy $(1,1,0)$ is dominant as blending levels are low and agents use economic motivation. However, once agents adopt higher blending levels, triggered through mutation, agents have a higher probability of using fusion motivation which leads to agents copying random action rules, eventually reducing cooperation. In contrast, when agents evolve to lower blending levels and adopt economic motivation, cooperation records a higher average. The results are expected and are in line with those of Sections 6.5 and 7.4. Furthermore, the result reaffirms our theory that cooperation moves negatively to the average blending level, $r(49) = -.99, p < .001$, as shown in Figure 7.7.

Algorithm 6 Algorithm for Indirect Reciprocity Based on the Reputation of Traits Where Blending Levels Determine Motivation

```
Require: Number of iterations $m$; number of generations $M$; set of agents $A$; set of traits
    $T$; set of traits $T$; set of binary action rules $AR = (s_i, u_i, d_i)$; set of blending levels
    $W$; cost $c$; benefit $b$; in-group interaction probability $S$; mutation rate of action rules
    $\mu_A$; mutation rate of blending levels $\mu_B$; the set of agents that share at least one trait
    with agent $j$ $N_j$;
for $M$ generations do                                        ▷ Evolutionary simulation
    Set $r_t = 0\ \forall t \in T$
    for $m$ iterations do
        $j \leftarrow U(A)$                         ▷ Select recipient (see Section 3.5.1)
        $p \leftarrow U(0, 1)$
        if $p < S$ and $|N_j| > 0$ then
            $i \leftarrow U(N_j)$                         ▷ Select random in-group donor
        else if $p \geq S$ and $|\bar{N}_j| > 0$ then
            $i \leftarrow U(\bar{N}_j)$                      ▷ Select random out-group donor
        else
            $i \leftarrow U(A - \{j\})$                         ▷ Select random donor
        end if
        ▷ Apply action Rules (see Sections 3.4.3 and 3.5.2)
        $r^i = (1 - w_i)r(t_i) + w_i r(g)$           ▷ $i$'s reputation is derived from $w_i$
        if ($r^i$=$r^j$ and $s_i = 1$)                              ▷ Compare equal
                or ($r^i < r^j$ and $u_i = 1$)                     ▷ Compare upwards
                or ($r^i > r^j$ and $d_i = 1$) then              ▷ Compare downwards
            $r_t \leftarrow min(5, r_t + 1)$             ▷ $i$ donates, increase reputation
        else                                                          ▷ $i$ defects
            if $r^j \geq r^i$ then                    ▷ Detect unjustified defection.
                $r_t \leftarrow max(-5, r_t - 1)$                 ▷ Decrease reputation
            end if
        end if
    end for
    ▷ Reproduction stage (see Sections 3.5.3 and 7.3)
    for $i \in A$ do                              ▷ Determine new values for all agents
        if $U([0, 1]) < w_i$ then
            $j \leftarrow R(A, w)$               ▷ Roulette wheel based on blending level
        else
            $j \leftarrow R(A, f)$                      ▷ Roulette wheel based on fitness
        end if
        $(s'_i, u'_i, d'_i) \leftarrow (s_j, u_j, d_j)$              ▷ $i$ copies $j$'s action rules
        $w'_i \leftarrow w_j$                             ▷ $i$ copies $j$'s blending level
        if $U([0, 1]) < \mu_A$ then
            $(s'_i, u'_i, d'_i) \leftarrow U(AR)$                    ▷ Mutate action rules
        end if
        if $U([0, 1]) < \mu_B$ then
            $w'_i \leftarrow U(W)$                               ▷ Mutate blending level
        end if
    end for
    for $i \in A$ do
        $(s_i, u_i, d_i) \leftarrow (s'_i, u'_i, d'_i)$      ▷ Update action rules for all agents
        $w_i \leftarrow w'_i$                        ▷ Update blending level for all agents
    end for
end for
```

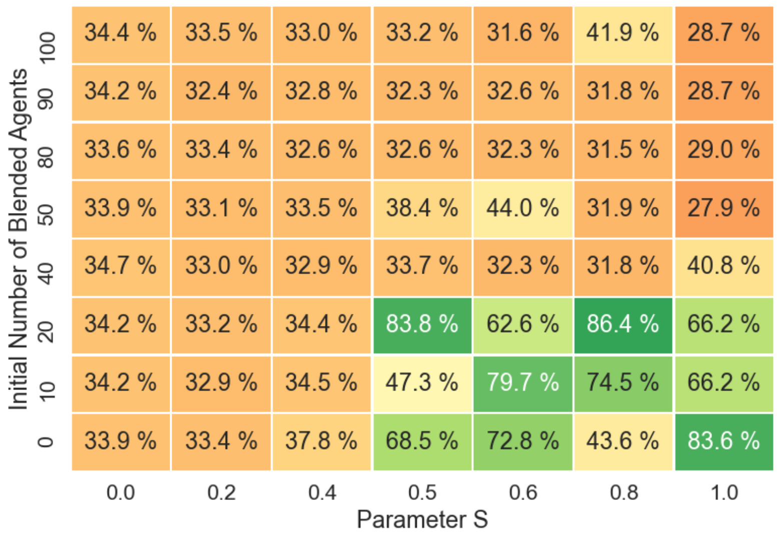


Figure 7.5: The average cooperation produced over generations when blending levels determine an agent's motivation. Parameter S imposes that when $S = 1$, blended agents cannot receive donations from independents agents, and when $S = 0$, blended agents can only receive donations from independent agents. Overall the cooperation recorded is low as agents evolve their action rules and blending levels with the exception of 10 and 20 when $S \geq 0.5$.

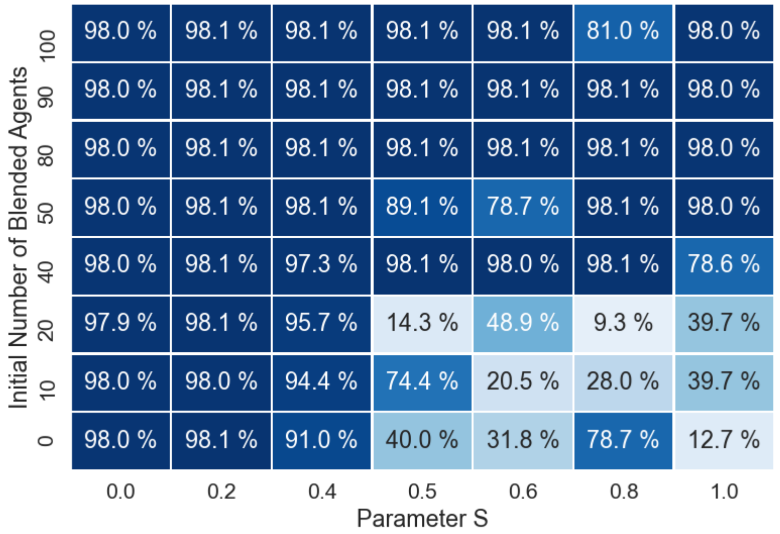


Figure 7.6: The average blending levels produced as a result of agents using their blending levels to determine their motivation and applying parameter S. Overall the average blending level recorded is high ($\geq 78\%$) as agents evolve their action rules and blending levels except for 10 and 20 when $S \geq 0.5$.

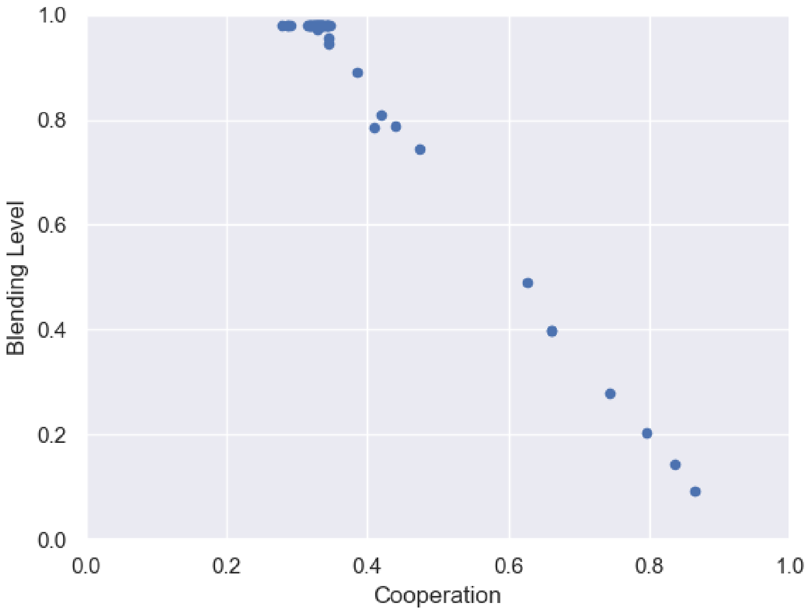
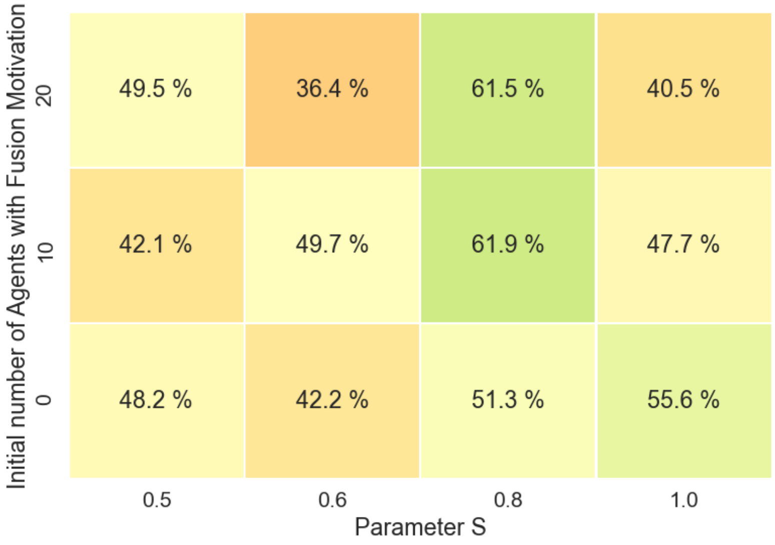


Figure 7.7: The correlation of blending levels and cooperation when agents use their blending levels to determine their motivation as described in Section 7.5. Each point represents a pair of (blending level, cooperation) from Figures 7.5 and 7.6.


Initial number of Agents with Fusion Motivation
Parameter S

	0.5	0.6	0.8	1.0
20	49.5 %	36.4 %	61.5 %	40.5 %
10	42.1 %	49.7 %	61.9 %	47.7 %
0	48.2 %	42.2 %	51.3 %	55.6 %

Figure 7.8: After a million generations, we note that average cooperation records a lower average than that of $100,000$ generations when the initial number of blended agents is low $(0, 10, 20)$ and $S \geq 0.5$. This is attributed to the higher blending average recorded as seen in Figure 7.9. The table shows the average cooperation recorded over generations when agents use their blending levels to determine their motivation. Parameter S imposes that when $S = 1$ blended agents cannot receive donations from independents agents, and when $S = 0$ blended agents can only receive donations from independent agents.

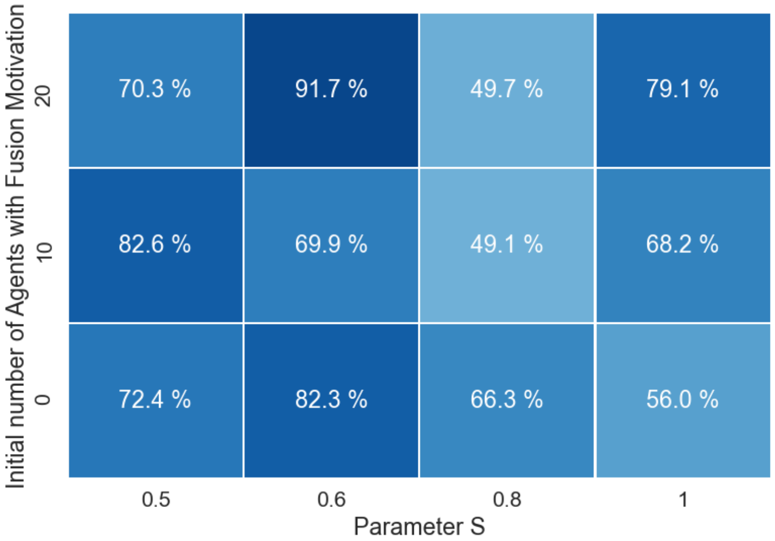


Figure 7.9: After a million generations, we note that blending levels record a higher average than that of $100,000$ generations when the initial number of blended agents is low $(0, 10, 20)$ and $s \geq 0.5$. This is attributed to mutation introducing higher blending levels to the population. The table shows the average blending levels when agents use their blending levels to determine their motivation. Parameter S imposes that when $S = 1$ blended agents cannot receive donations from independents agents, and when $S = 0$ blended agents can only receive donations from independent agents.

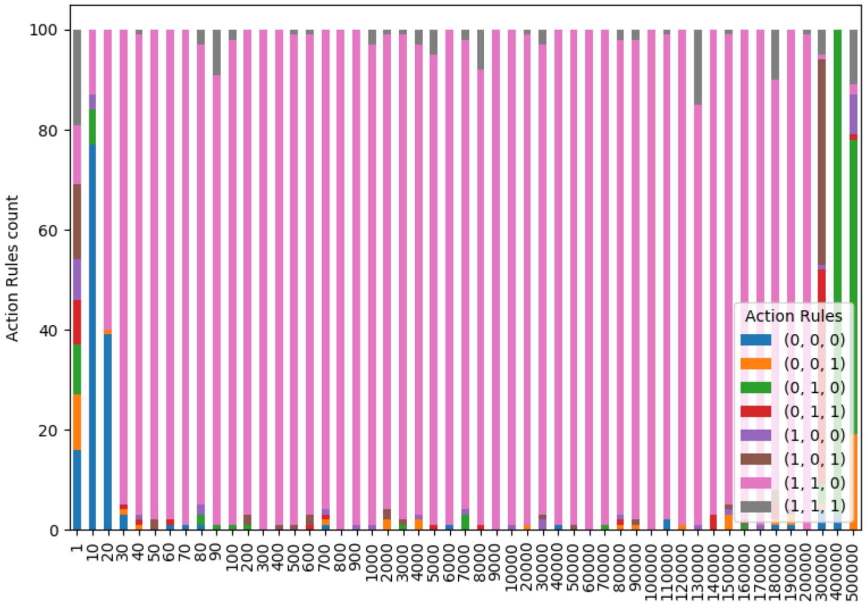


Figure 7.10: The distribution of action rules, in subsequent but not consecutive generations, produced as a result of agents using their blending levels to determine their motivation where all agents initially have a blending level of $w_i = 0$ (i.e. independent) and use economic motivation with $S = 0.6$.

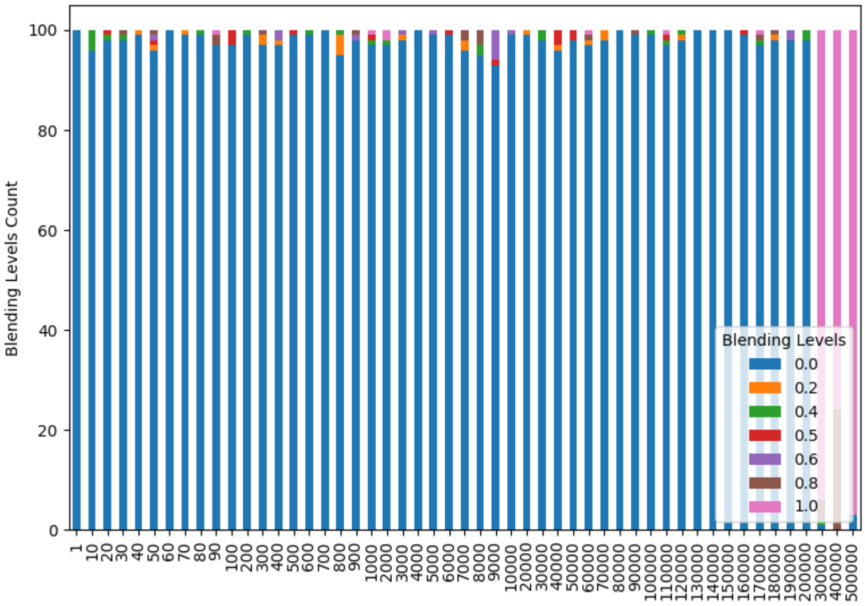


Figure 7.11: The distribution of blending levels, in subsequent but not consecutive generations, produced as a result of agents using their blending levels to determine their motivation where all agents initially have a blending level of $w_i = 0$ (i.e. independent) and use an economic motivation with $S = 0.6$.

7.6 Evolving Motivations in a Heterogeneous Population

In this section we investigate allowing agents to evolve their reproduction motivation along with action rules and blending levels. In this scenario, agents socially learn by adopting the action rules, blending levels, and the parent agent's motivation. Motivations are based on a probabilistic scale from the set $MV = \{0, 0.2, 0.4, 0.5, 0.6, 0.8, 0.9, 1\}$ the distribution of the scale was chosen to ensure that a number of probabilities are selected. The scenario allows us to determine further the effects of reproduction based on economic and fusion motivation without it being dictated to the agent through their blending level, such as in Section 7.5 or through the experiment setting, such as in Sections 6.5 and 7.4. In this scenario, we continue to apply mutation on action rules and on blending levels with a probability of 1% as discussed in Section 6.4. In addition, we apply the same rate of mutation on motivation, allowing for a change in motivation after assignment. We continue to use Parameter S in the same fashion as previous experiments, in which it dictates the probability of the chosen donor to be a blended agent. Algorithm 7 illustrates the reproduction phase that corresponds to this experiment.

The experiment in this section makes several assumptions in addition to those identified in Section 7.3.1. At the first generation, all independent agents, where $w_i = 0$, are assumed to have an economic motivation while dependent agents, where $w_i > 0$, are assigned a motivation probability from the set $\{0.2, 0.4, 0.5, 0.6, 0.8, 1\}$; this corresponds to Figure 7.12. The remaining components of the experiment follow the pattern outlined in Subsection 7.3.1 and Algorithm 3. Once a generation of interactions is complete, at the reproduction step, agents socially learn from others based on their relative success as measured by their motivation, either being economic or fusion (lines 24-28 in Algorithm 7). If an agent has an economic motivation, they will base their reproduction on the payoff of others. On the other hand, if an agent has fusion motivation, they will base reproduction on agents with high blending levels, as outlined in Section 7.3. Regardless of the motivation used, agents inherit their parent's action rules, blending level, and motivation to be used at the end of the next generation. Note that the motivation probability used is different from the blending level assigned to agents as they are separate parameters.

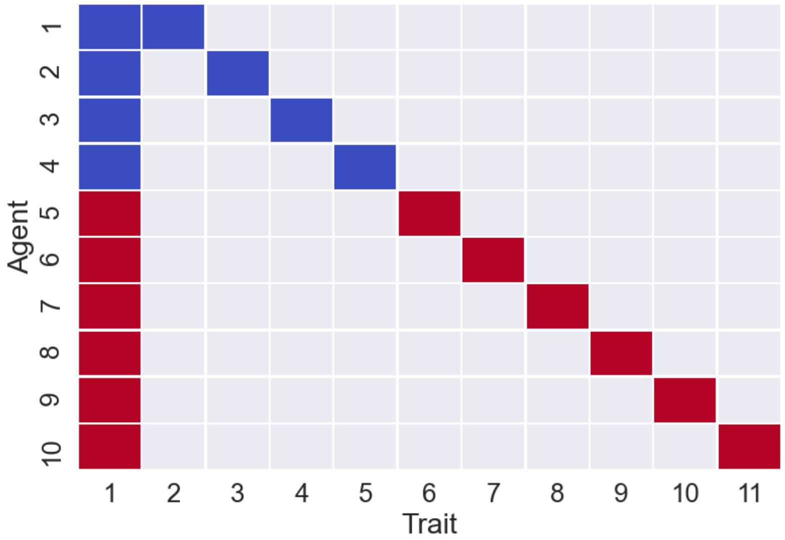


Figure 7.12: A simple visualisation of 10 agents, where agents may share a trait while maintaining different motivations, where blue represents agents with blending levels $w_i > 0$ and fusion motivation. Red represents independent agents $w_i = 0$, with economic motivation.

After updating agents' action rules, blending levels, and the motivation probability, mutation takes place. Mutation changes these parameters with a probability of $\frac{1}{100}$, the rate being sufficient to alter the population's structure as demonstrated in [198] and [199] and used in previous experiments in this thesis. Specifically, after an agent i has been assigned a new set of action rules, a blending level and a motivation probability mutation occurs. Blending level mutation randomly assigns agents one of seven blending levels identified in Subsection 6.3.2, including $w_i = 0$. Action rule mutation randomly assigns one of the eight action rules specified in Subsection 3.4.3. Motivation mutation randomly assigns the agent with a motivation probability from the set MV, with 0 being economic motivation and 1 being fusion motivation. Note that mutation is applied to both independent and blended agents.

The results, as shown in Figures 7.13 and 7.14, indicate that when agents evolve their motivation, cooperation records low levels and blending records high levels. The cooperation achieved is lower than any previous scenarios examined.

Figure 7.14 shows that all agents evolve towards blending levels higher than 90% regardless of their starting configuration. Agents that adopt fusion motivation quickly become highly blended as instructed by their motivation. In contrast, agents that have a lower proportion of fusion motivation will copy others based on their payoff. Within economic motivation scenarios, highly blended defector agents tend to be the agents with the highest payoff. As such, other agents with economic motivation will inherit high blending levels and defector strategies. This trend takes place in a few generations as shown in Figure 7.16.

Figure 7.13 shows that when agents evolve their motivation, cooperation records low levels averaging below 7%. This is attributed to agents adopting different probabilities of motivation as can be seen in Figure 7.15. This leads to the defector strategy being dominant as agents with a percentage economic motivation will primarily inherit from highly blended defectors, as shown in Figure 7.17. At the same time, the presence of fusion motivation guarantees that highly blended agents are maintained in the population, which impedes the population from going back to low blending levels, as shown in Figure 7.16, thus allowing shirkers to take advantage of the high blending levels in the population. As fusion motivation does not consider any action rule favourable to agents, agents probabilistically copy the agents' action rules with the highest blending level, which allows for the defective strategies to spread widely.

Algorithm 7 Algorithm for Indirect Reciprocity Based on the Reputation of Traits and Evolving Motivations

```
Require: Number of iterations $m$; number of generations $M$; set of agents $A$; set of traits $T$; set of traits $T$; set of binary action rules $AR = (s_i, u_i, d_i)$; set of blending levels $W$; cost $c$; benefit $b$; in-group interaction probability $S$; mutation rate of action rules $\mu_A$; mutation rate of blending levels $\mu_B$; the set of agents that share at least one trait with agent $j$ $N_j$;
for $M$ generations do                                      ▷ Evolutionary simulation
    Set $r_t = 0\ \forall t \in T$
    for $m$ iterations do
        $j \leftarrow U(A)$                                 ▷ Select recipient (see Section 3.5.1)
        $p \leftarrow U(0,1)$
        if $p < S$ and $|N_j| > 0$ then
            $i \leftarrow U(N_j)$                           ▷ Select random in-group donor
        else if $p \geq S$ and $|\bar{N_j}| > 0$ then
            $i \leftarrow U(\bar{N_j})$                     ▷ Select random out-group donor
        else
            $i \leftarrow U(A - \{j\})$                     ▷ Select random donor
        end if
        ▷ Apply action Rules (see Sections 3.4.3 and 3.5.2)
        $r^i = (1 - w_i)r(t_i) + w_i r(g)$                  ▷ $i$'s reputation is derived from $w_i$
        if ($r^i$=$r^j$ and $s_i = 1$)                       ▷ Compare equal
                or ($r^i < r^j$ and $u_i = 1$)               ▷ Compare upwards
                or ($r^i > r^j$ and $d_i = 1$) then          ▷ Compare downwards
            $r_t \leftarrow min(5, r_t + 1)$                ▷ $i$ donates, increase reputation
        else                                                ▷ $i$ defects
            if $r^j \geq r^i$ then                          ▷ Detect unjustified defection.
                $r_t \leftarrow max(-5, r_t - 1)$           ▷ Decrease reputation
            end if
        end if
    end for                                 ▷ Reproduction stage (see Sections 3.5.3 and 7.3)
    for $i \in A$ do
        if $U([0,1]) < mv_i$ then
            $j \leftarrow R(A, w)$                          ▷ Roulette wheel based on blending
        else
            $j \leftarrow R(A, f)$                          ▷ Roulette wheel based on fitness
        end if
        $(s'_i, u'_i, d'_i) \leftarrow (s_j, u_j, d_j)$     ▷ $i$ copies $j$'s action rules
        $w'_i \leftarrow w_j$                               ▷ $i$ copies $j$'s blending level
        $mv'_i \leftarrow mv_j$                             ▷ $i$ copies $j$'s motivation
        if $U([0,1]) < \mu_A$ then
            $(s'_i, u'_i, d'_i) \leftarrow U(AR)$           ▷ Mutate action rules
        end if
        if $U([0,1]) < \mu_B$ then
            $w'_i \leftarrow U(W)$                          ▷ Mutate blending level
        end if
        if $U([0,1]) < \mu_M$ then
            $mv'_i \leftarrow U(MV)$                                    ▷ Mutate motivation
        end if
    end for
    for $i \in A$ do
        $(s_i, u_i, d_i) \leftarrow (s'_i, u'_i, d'_i)$              ▷ Update action rules for all agents
        $w_i \leftarrow w'_i$                                ▷ Update blending level for all agents
        $mv_i \leftarrow mv'_i$                                  ▷ Update motivation for all agents
    end for
end for
```

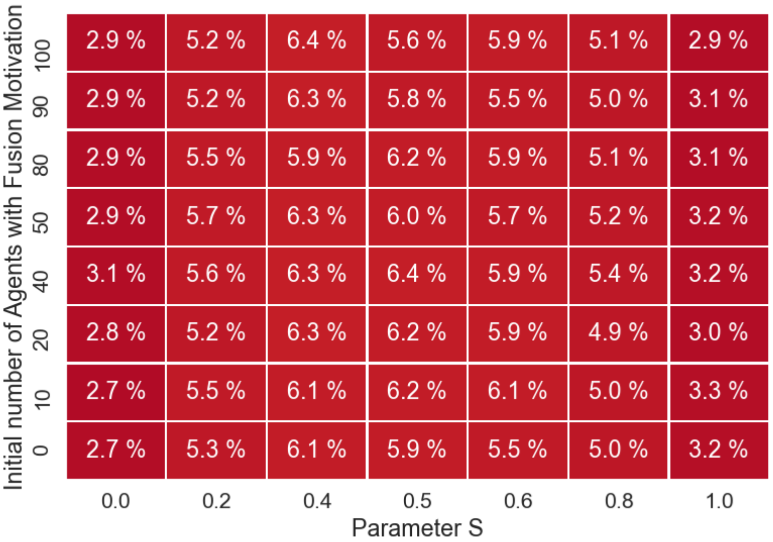
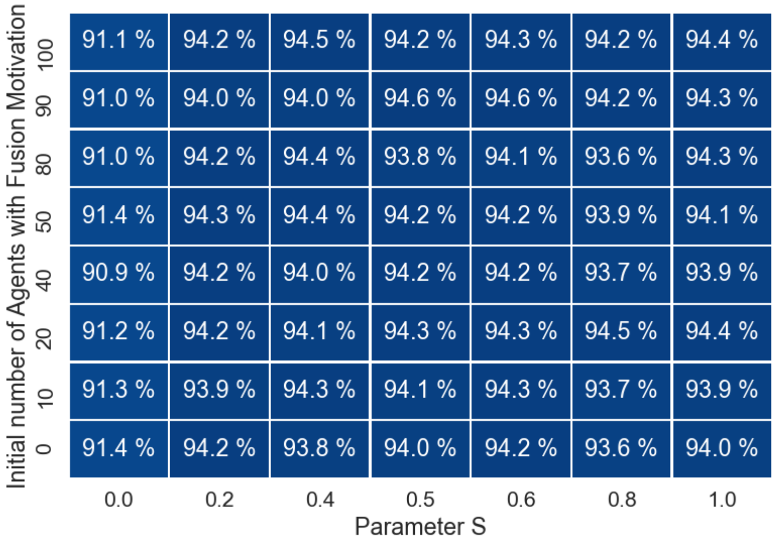


Figure 7.13: Low cooperation recorded as a result of agents evolving their motivation along with their blending level and action rules while enforcing parameter S. Cooperation is impacted heavily as agents evolve their motivations. A mutation of 1% is applied on motivation, blending levels and action rules.



Initial number of Agents with Fusion Motivation	0.0	0.2	0.4	0.5	0.6	0.8	1.0
100	91.1 %	94.2 %	94.5 %	94.2 %	94.3 %	94.2 %	94.4 %
90	91.0 %	94.0 %	94.0 %	94.6 %	94.6 %	94.2 %	94.3 %
80	91.0 %	94.2 %	94.4 %	93.8 %	94.1 %	93.6 %	94.3 %
50	91.4 %	94.3 %	94.4 %	94.2 %	94.2 %	93.9 %	94.1 %
40	90.9 %	94.2 %	94.0 %	94.2 %	94.2 %	93.7 %	93.9 %
20	91.2 %	94.2 %	94.1 %	94.3 %	94.3 %	94.5 %	94.4 %
10	91.3 %	93.9 %	94.3 %	94.1 %	94.3 %	93.7 %	93.9 %
0	91.4 %	94.2 %	93.8 %	94.0 %	94.2 %	93.6 %	94.0 %

Figure 7.14: High blending levels are recorded as a result of agents evolving their motivation along with their blending level and action rules while enforcing parameter S. Cooperation is impacted heavily as agents evolve their motivations. A mutation of 1% is applied on motivation, blending levels and action rules.

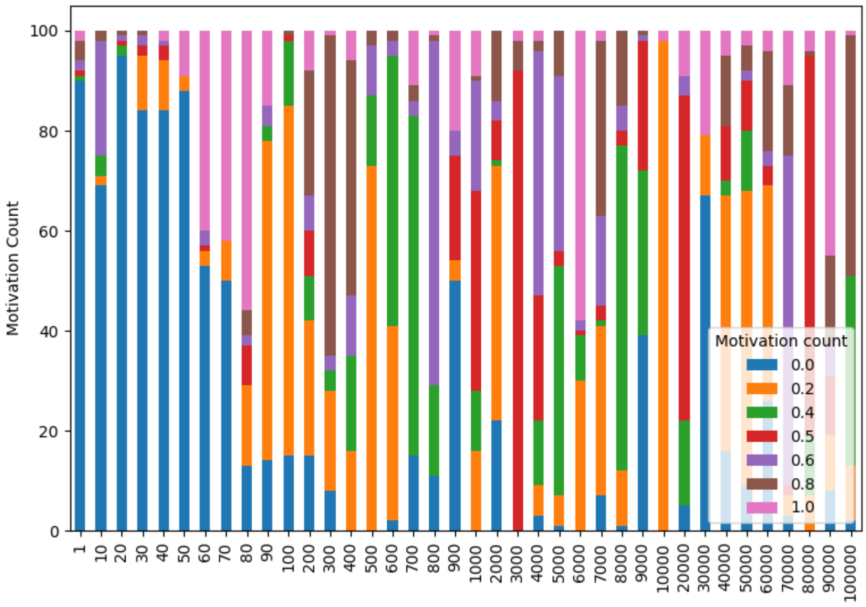


Figure 7.15: The distribution of the motivations produced, over subsequent but not consecutive generations, as a result of agents evolving their motivations where we initially assigned 90 agents economic motivation shows that no dominant motivation probability can dominate the population.

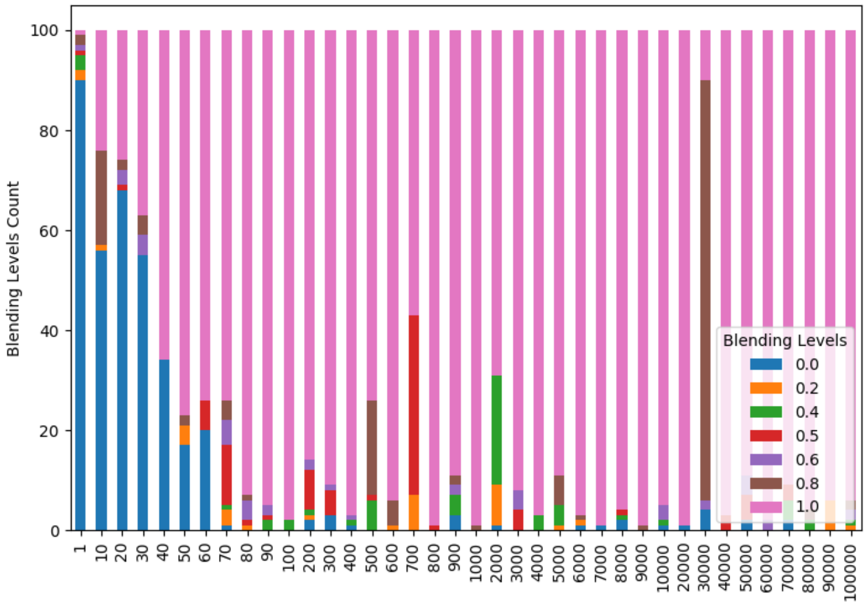


Figure 7.16: The distribution of the blending levels produced, over subsequent but not consecutive generations, where we initially assigned ninety agents as independent ($w_i = 0$) while allowing agents to evolve their motivations. Agents quickly evolve towards higher blending levels within a few generations. The figure shows the blending levels of agents of the same run and scenario as Figure 7.15.

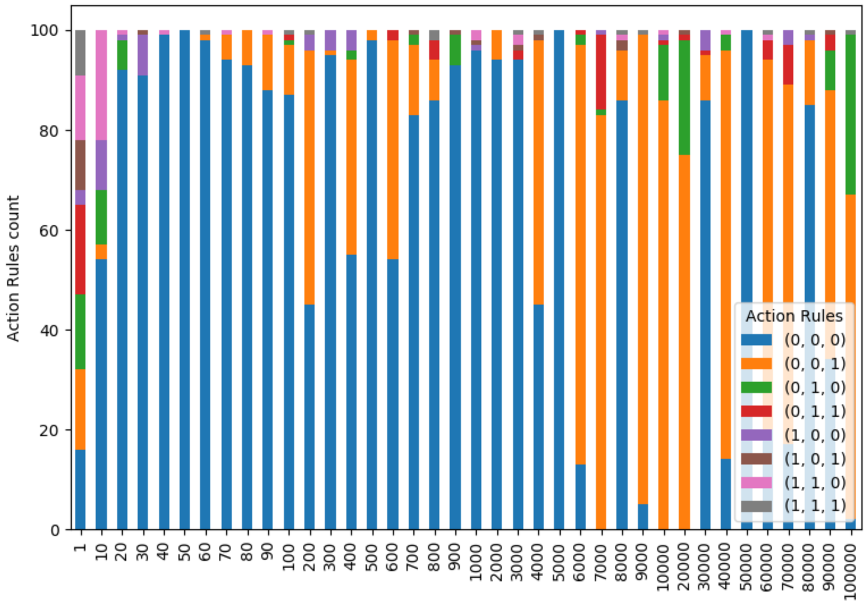


Figure 7.17: The distribution of action rules over subsequent but not consecutive generations as agents evolve their motivations shows that the defector action rule ($s_i, u_i, d_i = (0, 0, 0)$) dominates the population at the early generations. This trend of dominance is interrupted through mutation in later generations by introducing other action rules. The figure shows the action rules of agents of the same run and scenario as Figure 7.15.

7.7 Discussion

The chapter's focus on fusion motivation is a shift from previous chapters where individuals used explicit (i.e., economic) motivation based on payoff. In this chapter, individuals additionally use fusion motivation and can base success on perceptions of identity, showing preferences to copy those that are strongly aligned with a group's identity.

In Chapter 6, a new generation of agents probabilistically copied the action rules, traits and blending levels of other agents based on payoff. In this chapter, agents probabilistically copy others based on identity (blending levels). The motivation for this has originated in psychology, where individuals' behaviours in groups have been examined [174] and observed to be heavily motivated by identity, and the extent of fusion with a group's identity.

Our results have shown that allowing agents to reproduce based on fusion motivation drives cooperation to suffer significantly due to the randomness of the action rules inherited by agents. Our results in Section 7.4 show that adequate, but not high, levels of cooperation ($48 - 51\%$) are recorded when all agents reproduce with fusion motivation. The outcome, as mentioned earlier, is attributed to the random allocation of action rules at the first generation and the subsequent randomness in inheriting the action rules by agents in the following generations.

To further compare economic and fusion motivations, we developed two scenarios where agents evolve their motivations. Section 7.5 allowed agents to base their motivation on the blending level they have. In other words, as agents evolved their blending levels, their probability of using fusion motivation evolved as well. Our results in Section 7.5 showed that as agents adopt higher blending levels (higher than 90%), fusion motivation dominates the population, causing a drop in cooperation. Initially, when all cases of this scenario were examined, a small number of cases presented an alternative view. This outcome is attributed to the initial high number of independent agents with $w_i = 0$ and the enforcement of $S \geq 0.5$. However, based on further analysis, by allowing longer generation runs, we have shown that fusion motivation and high blending levels continue to dominate the population, although at a slower pace. We note that once the population converges towards fusion motivation, it fails to restore economic motivation even with a mutation of 1% having been applied.

Section 7.6 showed how agents evolve motivation based on their parent's motivation with a small mutation applied for randomness. Our results in Section 7.6 showed that high blending levels, similar to Section 7.5, quickly dominate the population from early generations. The dominance results in low cooperation rates (below 7%). The low cooperation outcome is attributed to agents adopting defective action rules (strategies) in early generations. Once an action rule is dominant within the population, fusion motivation does not alter it, and the change in action rules only occurs due to mutation. This outcome was found for all cases examined regardless of the starting motivation.

7.8 Conclusions

This chapter has introduced a novel selection method based on an agent's identity, referred to as fusion motivation. Fusion motivation allows agents to evolve based on their identity rather than using the convention of economic payoff. This represents agents using an intrinsic motivation rather than an extrinsic motivation, which is the game-theoretic convention. This has been implemented at the reproduction step, and equates fitness to an agent's blending level, with different implementations possible. We implemented three different fusion motivation approaches to examine how it affects identity sharing and cooperation.

Fusion motivation shifts the selection process's association from payoff to identity, with value placed in copying others that have a high blending level towards the group. This chapter's results showed that fusion motivation heavily disrupts cooperation. However, it allows agents to value devotion to their group identity. The outcome is attributed to the motivation's dominance and its allowance for random action rules to exist in the population. Although we continue to use the standing assessment, defective strategies escape punishment as agents focus on adopting higher blending levels rather than maintaining higher payoff or reputation.

Our approach to fusion motivation was to allow agents to represent the human behaviour towards groups where motivation shifts between extrinsic (wealth) and intrinsic motivations (based on an internal value of the group to the individual). Furthermore, this approach is linked with the literature on identity fusion [175, 177], which studies how individuals value groups and fuse their identity with that of the group. Our model's agents have simple cognitive functions, and therefore the model they operate in cannot represent the complexity of human social life. However, interestingly, the results indicate that fusion motivation is a potentially powerful force in drawing agents away from economic incentives, with a challenge being presented in how agents may be incentivised to change strategy once they pursue intrinsic motivation. Valuing who an individual is, rather than what they achieve (in terms of economic payoff) allows de-prioritisation of rational economic rewards that can underpin survival. Without other mechanisms coming into play (e.g. different punishment mechanisms) [105, 139, 149], this can be detrimental to the long term prosperity of a population.

Overall, this chapter reaffirms that the consequences of intrinsic motivation on a population are potentially significant, and are likely to work in tandem, to some degree with economic motivation. This is worthy of future development in its own right, and identity is a valuable point on which to focus, given the rationale presented from psychology.

Chapter 8

Conclusions & Future Work

8.1 Introduction

Overall, the contribution of this thesis has concerned generalising a fundamental assumption that is widely used in cooperation systems: namely that each agent has a unique identity. We have developed and examined this by

- allowing each agent to compose an identity based on different "units" of identity, referred to as traits;
- introducing sharing of identity through traits being adopted by multiple agents;
- letting traits carry reputation, which is then inherited by those carrying the traits.

Our identity framework has been developed using indirect reciprocity, where agents are presented with the dilemma of donating to others without the guarantee of reciprocation. This is a fundamental and useful game to consider because it is based on reputation, which effectively represents a quality measure associated to an identity. This is also highly relevant to literature concerning human systems (psychology), theoretical biology and economics (game theory). In particular, the framework allows group identity to be considered, through groups having an identity (trait) which can be shared by group members, while also retaining elements of identity (i.e., other trait(s)) that are unique to themselves. This has allowed us to consider how concepts of identity in the human world (e.g., group identity, stereotyping) may be represented in our model and cause an impact.

As a result of this framework, we have been able to determine how changing the structure of identity alone, affects cooperation. This is useful because it quantifies the extent to which cooperation systems need to have functionality (e.g., second-order mechanisms) beyond the basic reputation system in order for cooperation to be sustained.

A number of aspects of our results are striking. Although we find that cooperation has the potential to be heavily disrupted by shared identity, such as in Section 4.4 when the number of agents sharing a single trait exceeded 15%, it is also apparent that considerable cooperation can emerge if sharing is structured in a particular way. For example, when sharing takes place in smaller groups as was the case in Section 4.4.1, or by allowing individuals to evolve their groups such as in Section 5.3. On the other hand, it has also been apparent that concepts that occur in the human world related to identity, such as stereotyping, Section 4.4, have an equally considerable impact in the simulated world of indirect reciprocity. These result in considerable reductions in the observed cooperation.

Another element of surprise from the results has been the discovery of the relationship to evolutionary set theory (See Section 5.3.1). To date, this has been considered as an alternative model to cooperation, as compared to a reputation system, with minimal components in the model. This model focuses on agent set memberships and their movement between sets, with moderate levels of cooperation observed (e.g., 50-70%, [122, 185]). However, in Section 5.3.1, we observe that as cooperation collapses due to sharing of traits, our model defaults to set based cooperation, as seen in [71, 186]. This is a significant observation that relates two alternative approaches to sustaining cooperation which have previously been considered unrelated, and indeed potentially orthogonal in the sense that these relate to different communities of interest (i.e., reputation systems and set theory). We believe this is the first time that a relationship between these different types of models has been established.

Further interesting observations have arisen from Chapters 6 and 7 where we have explored an alternative formulation for identity that allows agents to effectively have a trade-off in their identity between the extent they inherit the shared group trait, as opposed to their unique (non-shared) personal identity. This formulation is also interesting because it relates, at a simplified level, to the psychological concept of identity fusion [177], which is an important contemporary psychology explanation as to how group identity and individual identity may overlap (i.e., blend) and combine to motivate an individual's group-driven behaviour. Here we have explored a representation of this and found, for the first time, the relationship between cooperation and the extent of blending levels. In particular, in the absence of any additional social mechanisms, our findings (Chapter 7) re-affirms that balance between identities is needed for cooperation to emerge and be sustained.

Finally, insights due to alternative assumptions on motivating agent behaviours have considerable feedback. From the psychology literature (particularly social identity theory [182] and identity fusion theory [177]) it is clear that identity can be a motivating force in its own right for (human) agents, because of the value that group membership can bring to the individual. Therefore group identity was introduced as an attraction at the selection stage. Various formulations were considered, alongside the traditional economic motivation of payoff. This approach disrupts the economic assumption of perfect rationality that is commonly applied in game theory. The results showed that motivations related to identity were in fact a very powerful force, which is due to an inherent feedback loop that promotes identity across the population. Under these circumstances, cooperation appears to become randomised (i.e., 50%) while high blending levels take hold under a number of experimental conditions. These findings are consistent with general observations concerning identity fusion [177] in the sense that identity driven motivations can become deeply established and hard to reverse. These findings also reaffirm that there is considerable scope to further develop game theoretical modelling with a greater emphasis on behavioural motivations.

8.2 Limitations and Future Work

It is important to understand that our work represents a reductionist model of possible actions and responses, and as such, it only informs a limited aspect of possible activity. For example, a particular limitation is the cognition available to agents, such as the donor's ability to view their recipient's trait's reputation and compare it. This limitation of cognition means that the agents cannot represent human conditions, as humans are more complex and may consider different traits to be of higher importance than others and, as such, may view the reputations of others differently. However, on the other hand, the agent-based simulation approach allows for the control of assumptions so that factors can be considered in isolation. Furthermore, agent-based simulations are unable to cope with massive populations without additional resources. Nonetheless, agent-based models remain an important tool that enables us to explore cooperation problems and implement different scenarios.

In terms of further work, there are multiple important aspects. Firstly, as highlighted in Chapter 7, there is considerable scope to explore how bounded rationality and further human-level motivations can impact our understanding of the fundamentals of cooperation. This can further help to understand how social dynamics take hold in individual-level decision making and have collective effects. Additionally, blended identities, Chapter 6, could be further investigated by considering multiple groups to which an agent can subscribe and how this affects both their identity and, in return, cooperation. Finally, different sharing structures should be considered where agents may prioritise some traits when engaging with others.

It is also important to note that although this work is of a fundamental nature, as machines and devices become autonomous, they will need mechanisms through which they can understand and rationalise each other. Defining identity through traits, and making decisions based on these, could become increasingly important.

Bibliography

[1] Christoph Adami, Jory Schossau, and Arend Hintze. “Evolutionary game theory using agent-based methods”. In: *Physics of Life Reviews* 19 (2016), pp. 1–26. ISSN: 1571-0645. DOI: https://doi.org/10.1016/j.plrev.2016.08.015. URL: http://www.sciencedirect.com/science/article/pii/S1571064516300884.

[2] Icek Ajzen. “Attitudes, traits, and actions: Dispositional prediction of behavior in personality and social psychology”. In: *Advances in experimental social psychology*. Vol. 20. Elsevier, 1987, pp. 1–63.

[3] C Athena Aktipis. “Know when to walk away: contingent movement and the evolution of cooperation”. In: *Journal of theoretical biology* 231.2 (2004), pp. 249–260.

[4] Richard D Alexander. *The biology of moral systems*. Transaction Publishers, 1987.

[5] Tibor Antal, Hisashi Ohtsuki, John Wakeley, Peter D Taylor, and Martin A Nowak. “Evolution of cooperation by phenotypic similarity”. In: *Proceedings of the National Academy of Sciences* 106.21 (2009), pp. 8597–8600.

[6] Jacobien van Apeldoorn and Arthur Schram. “Indirect reciprocity; a field experiment”. In: *PloS one* 11.4 (2016).

[7] Dan Ariely, Anat Bracha, and Stephan Meier. “Doing good or doing well? Image motivation and monetary incentives in behaving prosocially”. In: *American Economic Review* 99.1 (2009), pp. 544–55.

[8] Scott Atran. “The devoted actor: unconditional commitment and intractable conflict across cultures”. In: *Current Anthropology* 57.S13 (2016), S192–s203.

[9] Robert Axelrod. “Chapter 33 Agent-based Modeling as a Bridge Between Disciplines”. In: ed. by L. Tesfatsion and K.L. Judd. Vol. 2. Handbook of Computational Economics. Elsevier, 2006, pp. 1565–1584. DOI: https://doi.org/10.1016/S1574-0021(05)02033-2. URL: http://www.sciencedirect.com/science/article/pii/S1574002105020332.

[10] Robert Axelrod. “Effective Choice in the Prisoner’s Dilemma”. In: *Journal of Conflict Resolution* 24.1 (1980), pp. 3–25. DOI: 10.1177/002200278002400101. eprint: https://doi.org/10.1177/002200278002400101. URL: https://doi.org/10.1177/002200278002400101.

[11] Robert Axelrod. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press, 1997. ISBN: 9780691015675. URL: http://www.jstor.org/stable/j.ctt7s951.

[12] Robert Axelrod. “The dissemination of culture: A model with local convergence and global polarization”. In: *Journal of conflict resolution* 41.2 (1997), pp. 203–226.

[13] Robert Axelrod and William D Hamilton. “The evolution of cooperation”. In: *science* 211.4489 (1981), pp. 1390–1396.

[14] Robert M Axelrod. *The evolution of cooperation*. eng. Rev. ed. New York: Basic Books, 2006. ISBN: 9780465005642.

[15] Albert Bandura and Richard H Walters. *Social learning theory*. Vol. 1. Prentice-hall Englewood Cliffs, NJ, 1977.

[16] B. Baranski, T. Bartz-Beielstein, R. Ehlers, T. Kajendran, B. Kosslers, J. Mehnen, T. Polaszek, R. Reimholz, J. Schmidt, K. Schmitt, D. Seis, R. Slodzinski, S. Steeg, N. Wiemann, and M. Zimmermann. “The Impact of Group Reputation in Multiagent Environments”. In: *2006 IEEE International Conference on Evolutionary Computation*. 2006, pp. 1224–1231. DOI: 10.1109/cec.2006.1688449.

[17] Ali Bazghandi. “Techniques, advantages and problems of agent based modeling for traffic simulation”. In: *International Journal of Computer Science Issues (IJCSI)* 9.1 (2012), p. 115.

[18] Adam Bear and David G. Rand. “Intuition, deliberation, and the evolution of cooperation”. In: *Proceedings of the National Academy of Sciences* 113.4 (2016), pp. 936–941. ISSN: 0027-8424.

[19] Roland Benabou and Jean Tirole. “Intrinsic and extrinsic motivation”. In: *The review of economic studies* 70.3 (2003), pp. 489–520.

[20] Cristina Bicchieri, Ryan Muldoon, and Alessandro Sontuoso. “Social Norms”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2018. Metaphysics Research Lab, Stanford University, 2018.

[21] Jason D. Boardman, Benjamin W. Domingue, and Jason M. Fletcher. “How social and genetic factors predict friendship networks”. In: *Proceedings of the National Academy of Sciences* 109.43 (2012), pp. 17377–17381. ISSN: 0027-8424. DOI: 10.1073/pnas.1208975109. eprint: https://www.pnas.org/content/109/43/17377.full.pdf. URL: https://www.pnas.org/content/109/43/17377.

[22] Galen V Bodenhausen and Robert S Wyer. “Effects of stereotypes in decision making and information-processing strategies.” In: *Journal of personality and social psychology* 48.2 (1985), p. 267.

[23] Eric Bonabeau. “Agent-based modeling: Methods and techniques for simulating human systems”. In: *Proceedings of the National Academy of Sciences* 99.suppl 3 (2002), pp. 7280–7287. ISSN: 0027-8424. DOI: 10.1073/pnas.082080899. eprint: https://www.pnas.org/content/99/suppl_3/7280.full.pdf. URL: https://www.pnas.org/content/99/suppl_3/7280.

[24] Dan Boneh and Xavier Boyen. “Efficient Selective-ID Secure Identity-Based Encryption Without Random Oracles”. In: *Advances in Cryptology - EUROCRYPT*

2004. Ed. by Christian Cachin and Jan L. Camenisch. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 223–238. ISBN: 978-3-540-24676-3.

[25] Robert Boyd, Herbert Gintis, and Samuel Bowles. "Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare". In: *Science* 328.5978 (2010), pp. 617–620. ISSN: 0036-8075. DOI: `10.1126/science.1183665`. eprint: `https://science.sciencemag.org/content/328/5978/617.full.pdf`. URL: `https://science.sciencemag.org/content/328/5978/617`.

[26] Robert Boyd and Peter J. Richerson. "Punishment allows the evolution of cooperation (or anything else) in sizable groups". In: *Ethology and Sociobiology* 13.3 (1992), pp. 171–195. ISSN: 0162-3095. DOI: `https://doi.org/10.1016/0162-3095(92)90032-Y`. URL: `http://www.sciencedirect.com/science/article/pii/016230959290032Y`.

[27] Robert Boyd and Peter J Richerson. "The evolution of indirect reciprocity". In: *Social networks* 11.3 (1989), pp. 213–236.

[28] Hannelore Brandt and Karl Sigmund. "The logic of reprobation: assessment and action rules for indirect reciprocation". In: *Journal of Theoretical Biology* 231.4 (2004), pp. 475–486. ISSN: 0022-5193. DOI: `https://doi.org/10.1016/j.jtbi.2004.06.032`. URL: `https://www.sciencedirect.com/science/article/pii/S0022519304002784`.

[29] Hannelore Brandt, Hisashi Ohtsuki, Yoh Iwasa, and Karl Sigmund. "A survey of indirect reciprocity". In: *Mathematics for Ecology and Environmental Sciences*. Springer, 2007, pp. 21–49.

[30] Joel G. Breman and Isao Arita. "The Confirmation and Maintenance of Smallpox Eradication". In: *New England Journal of Medicine* 303.22 (Nov. 1980), pp. 1263–1273. DOI: `10.1056/nejm198011273032204`. URL: `https://doi.org/10.1056/nejm198011273032204`.

[31] Marilynn B. Brewer. “The psychology of prejudice: Ingroup love or outgroup hate?” In: *Journal of Social Issues* 55.3 (1999), pp. 429–444. DOI: 10.1111/0022-4537.00126. URL: https://doi.org/10.1111/0022-4537.00126.

[32] Rogers Brubaker and Frederick Cooper. “Beyond ”Identity””. In: *Theory and Society* 29.1 (2000), pp. 1–47. ISSN: 03042421, 15737853. URL: http://www.jstor.org/stable/3108478.

[33] David Buckingham. *Introducing identity*. MacArthur Foundation Digital Media and Learning Initiative, 2008.

[34] Colin F Camerer. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press, 2011.

[35] Shelly Chaiken, Avika Liberman, Alice H Eagly, JS Uleman, and JA Bargh. “Unintended thought”. In: *Unintended thought* (1989), pp. 212–252.

[36] Mei huan Chen, Li Wang, Shi wen Sun, Juan Wang, and Cheng yi Xia. “Evolution of cooperation in the spatial public goods game with adaptive reputation assortment”. In: *Physics Letters A* 380.1 (2016), pp. 40–47. ISSN: 0375-9601. DOI: https://doi.org/10.1016/j.physleta.2015.09.047. URL: http://www.sciencedirect.com/science/article/pii/S0375960115008506.

[37] Y. Chen and K. J. R. Liu. “Indirect Reciprocity Game Modelling for Cooperation Stimulation in Cognitive Networks”. In: *IEEE Transactions on Communications* 59.1 (2011), pp. 159–168.

[38] Andrew M. Colman. “Cooperation, psychological game theory, and limitations of rationality in social interaction”. In: *Behavioral and Brain Sciences* 26.2 (2003), 139–153. DOI: 10.1017/s0140525x03000050.

[39] Andrew M Colman, Lindsay Browning, and Briony D Pulford. “Spontaneous similarity discrimination in the evolution of cooperation”. In: *Journal of Theoretical Biology* 299 (2012), pp. 162–171.

[40] Karen Schweers Cook and Margaret Levi. *The limits of rationality*. University of Chicago Press, 2008.

[41] Rob Cover. *Digital identities: Creating and communicating the online self*. Academic Press, 2015.

[42] Darren Croft, Jens Krause, Safi Darden, Indar Ramnarine, Jolyon Faria, and Richard James. "Behavioural trait assortment in a social network: Patterns and implications". In: *Behavioral Ecology and Sociobiology* 63 (Aug. 2009), pp. 1495–1503. DOI: `10.1007/s00265-009-0802-x`.

[43] Andrew T. Crooks and Alison J. Heppenstall. "Introduction to Agent-Based Modelling". In: *Agent-Based Models of Geographical Systems*. Ed. by Alison J. Heppenstall, Andrew T. Crooks, Linda M. See, and Michael Batty. Dordrecht: Springer Netherlands, 2012, pp. 85–105. ISBN: 978-90-481-8927-4. DOI: `10.1007/978-90-481-8927-4_5`. URL: `https://doi.org/10.1007/978-90-481-8927-4_5`.

[44] Oliver Scott Curry, Matthew Jones Chesters, and Caspar J Van Lissa. "Mapping morality with a compass: Testing the theory of 'morality-as-cooperation'with a new questionnaire". In: *Journal of Research in Personality* 78 (2019), pp. 106–124.

[45] Oliver Scott Curry, Lee A Rowland, Caspar J Van Lissa, Sally Zlotowitz, John McAlaney, and Harvey Whitehouse. "Happy to help? A systematic review and meta-analysis of the effects of performing acts of kindness on the well-being of the actor". In: *Journal of Experimental Social Psychology* 76 (2018), pp. 320–329.

[46] Charles Darwin. *On the Origin of Species by Means of Natural Selection Or the Preservation of Favoured Races in the Struggle for Life.* H. Milford; Oxford University Press, 1859.

[47] Richard Dawkins. "The selfish gene Oxford university press". In: *New York, New York, USA* (1976).

[48] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. "The spreading of misinformation online". In: *Proceedings of the National Academy of Sciences* 113.3 (2016), pp. 554–559.

[49] Oxford English Dictionary. *"co-operation, n."*. URL: https://www.oed.com/view/Entry/41037?result=1{\&}amprskey=GApyTc{\&}amp.

[50] Oxford English Dictionary. *"identity, n."*. URL: https://www.oed.com/view/Entry/91004?redirectedFrom=identity.

[51] Michael Doebeli and Christoph Hauert. "Models of cooperation based on the Prisoner's Dilemma and the Snowdrift game". In: *Ecology Letters* 8.7 (2005), pp. 748–766. DOI: 10.1111/j.1461-0248.2005.00773.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1461-0248.2005.00773.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1461-0248.2005.00773.x.

[52] John F. Dovidio, Samuel L. Gaertner, and Ana Validzic. "Intergroup bias: Status, differentiation, and a common in-group identity." In: *Journal of Personality and Social Psychology* 75.1 (1998), pp. 109–120. DOI: 10.1037/0022-3514.75.1.109. URL: https://doi.org/10.1037/0022-3514.75.1.109.

[53] Lee Alan Dugatkin. *Cooperation among animals: an evolutionary perspective*. Oxford University Press on Demand, 1997.

[54] Julia Eberlen, Geeske Scholz, and Matteo Gagliolo. "Simulate this! An introduction to agent-based models and their power to improve your research practice". In: *International Review of Social Psychology* 30.1 (2017).

[55] Víctor M. Eguíluz, Martín G. Zimmermann, Camilo J. Cela-Conde, and Maxi San Miguel. "Cooperation and the Emergence of Role Differentiation in the Dynamics of Social Networks". In: *American Journal of Sociology* 110.4 (2005), pp. 977–

1008. DOI: 10.1086/428716. eprint: https://doi.org/10.1086/428716. URL: https://doi.org/10.1086/428716.

[56] Euel Elliott and L. Douglas Kiel. “Exploring cooperation and competition using agent-based modeling”. In: *Proceedings of the National Academy of Sciences* 99.suppl 3 (2002), pp. 7193–7194. ISSN: 0027-8424. DOI: 10.1073/pnas.102079099. eprint: https://www.pnas.org/content/99/suppl_3/7193.full.pdf. URL: https://www.pnas.org/content/99/suppl_3/7193.

[57] The Editors of Encyclopaedia Britannica. *Natural selection*. Feb. 2020. URL: https://www.britannica.com/science/natural-selection.

[58] Alison Etheridge. *Some Mathematical Models from Population Genetics*. Springer Berlin Heidelberg, 2011.

[59] Ernst Fehr. “Human behaviour: don’t lose your reputation”. In: *Nature* 432.7016 (2004), pp. 449–450.

[60] Ernst Fehr and Urs Fischbacher. “The nature of human altruism”. In: *Nature* 425.6960 (2003), pp. 785–791.

[61] Ernst Fehr and Simon Gächter. “Altruistic punishment in humans”. In: *Nature* 415.6868 (2002), pp. 137–140.

[62] M. Feldman, C. Papadimitriou, J. Chuang, and I. Stoica. “Free-riding and whitewashing in peer-to-peer systems”. In: *IEEE Journal on Selected Areas in Communications* 24.5 (2006), pp. 1010–1019.

[63] Michal Feldman and John Chuang. “Overcoming free-riding behavior in peer-to-peer systems”. In: *ACM sigecom exchanges* 5.4 (2005), pp. 41–50.

[64] Michal Feldman and John Chuang. “The evolution of cooperation under cheap pseudonyms”. In: *Seventh IEEE International Conference on E-Commerce Technology (CEC’05)*. Ieee. 2005, pp. 284–291.

[65] Leon Festinger. “A theory of social comparison processes”. In: *Human relations* 7.2 (1954), pp. 117–140.

[66] Ronald Aylmer Fisher. *The genetical theory of natural selection,* Oxford: The Clarendon Press, 1930.

[67] Michael A. Fishman. “Indirect reciprocity among imperfect individuals”. In: *Journal of Theoretical Biology* 225.3 (2003), pp. 285–292. ISSN: 0022-5193. DOI: `https://doi.org/10.1016/S0022-5193(03)00246-7`. URL: `http://www.sciencedirect.com/science/article/pii/S0022519303002467`.

[68] Kevin R Foster, Tom Wenseleers, and Francis LW Ratnieks. “Kin selection is the key to altruism”. In: *Trends in ecology & evolution* 21.2 (2006), pp. 57–60.

[69] James H Fowler. “Altruistic punishment and the origin of cooperation”. In: *Proceedings of the National Academy of Sciences* 102.19 (2005), pp. 7047–7049.

[70] James H. Fowler, Jaime E. Settle, and Nicholas A. Christakis. “Correlated genotypes in friendship networks”. In: *Proceedings of the National Academy of Sciences* 108.5 (2011), pp. 1993–1997. ISSN: 0027-8424. DOI: `10.1073/pnas.1011687108`. eprint: `https://www.pnas.org/content/108/5/1993.full.pdf`. URL: `https://www.pnas.org/content/108/5/1993`.

[71] Feng Fu, Corina E Tarnita, Nicholas A Christakis, Long Wang, David G Rand, and Martin A Nowak. “Evolution of in-group favoritism”. In: *Scientific reports* 2 (2012), p. 460.

[72] Feng Fu, Martin A Nowak, Nicholas A Christakis, and James H Fowler. “The evolution of homophily”. In: *Scientific reports* 2 (2012), p. 845.

[73] Drew Fudenberg and Eric Maskin. “The folk theorem in repeated games with discounting or with incomplete information”. In: *A Long-Run Collaboration On Long-Run Games*. World Scientific, 2009, pp. 209–230.

[74] Adam D Galinsky and Gordon B Moskowitz. “Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism.” In: *Journal of personality and social psychology* 78.4 (2000), p. 708.

[75] Shiping Gao, Te Wu, and Long Wang. “Evolution of global cooperation and ethnocentrism in group-structured populations”. In: *Physics Letters A* 382.31 (2018), pp. 2027–2043.

[76] Julián García and Jeroen C.J.M. [van den Bergh]. “Evolution of parochial altruism by multilevel selection”. In: *Evolution and Human Behavior* 32.4 (2011), pp. 277–287. ISSN: 1090-5138. DOI: https://doi.org/10.1016/j.evolhumbehav.2010.07.007. URL: http://www.sciencedirect.com/science/article/pii/S1090513810000784.

[77] Gerd Gigerenzer and Henry Brighton. “Homo Heuristicus: Why Biased Minds Make Better Inferences”. In: *Topics in Cognitive Science* 1.1 (2009), pp. 107–143. DOI: 10.1111/j.1756-8765.2008.01006.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1756-8765.2008.01006.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-8765.2008.01006.x.

[78] Gerd Gigerenzer and Wolfgang Gaissmaier. “Heuristic Decision Making”. In: *Annual Review of Psychology* 62.1 (2011). Pmid: 21126183, pp. 451–482. DOI: 10.1146/annurev-psych-120709-145346. eprint: https://doi.org/10.1146/annurev-psych-120709-145346. URL: https://doi.org/10.1146/annurev-psych-120709-145346.

[79] Nigel Gilbert. *Agent-based models*. Vol. 153. Sage Publications, Incorporated, 2019.

[80] J. Glover. *I: The Philosophy and Psychology of Personal Identity*. Penguin Books, 1988.

[81] Kurt Gray, David G. Rand, Eyal Ert, Kevin Lewis, Steve Hershman, and Michael I. Norton. “The Emergence of “Us and Them” in 80 Lines of Code: Modeling Group Genesis in Homogeneous Populations”. In: *Psychological Science* 25.4 (2014). Pmid: 24590382, pp. 982–990. DOI: 10.1177/0956797614521816.

eprint: https://doi.org/10.1177/0956797614521816. URL: https://doi.org/10.1177/0956797614521816.

[82] Stuart Hall. “Who needs identity”. In: *Questions of cultural identity* 16.2 (1996), pp. 1–17.

[83] W D Hamilton, R Axelrod, and R Tanese. “Sexual reproduction as an adaptation to resist parasites (a review)”. In: *Proceedings of the National Academy of Sciences* 87.9 (1990), pp. 3566–3573. ISSN: 0027-8424. eprint: https://www.pnas.org/content/87/9/3566.full.pdf. URL: https://www.pnas.org/content/87/9/3566.

[84] William D Hamilton. “The evolution of altruistic behavior”. In: *The American Naturalist* 97.896 (1963), pp. 354–356.

[85] William D Hamilton. “The genetical evolution of social behaviour. II”. In: *Journal of theoretical biology* 7.1 (1964), pp. 17–52.

[86] Ross A. Hammond and Robert Axelrod. “The Evolution of Ethnocentrism”. In: *The Journal of Conflict Resolution* 50.6 (2006). Full publication date: Dec., 2006, pp. 926–936. URL: http://www.jstor.org/stable/27638531.

[87] Joacim Hansson. *Libraries and identity: the role of institutional self-image and identity in the emergence of new types of libraries*. Elsevier, 2010.

[88] Amy K. Heger and Lowell Gaertner. “Testing the identity synergy principle: Identity fusion promotes self and group sacrifice”. In: *Self and Identity* 17.5 (Sept. 2018), pp. 487–499. ISSN: 15298876. DOI: 10.1080/15298868.2017.1422538. URL: https://www.tandfonline.com/action/journalInformation?journalCode=psai20.

[89] Jospeh Henrich and Robert Boyd. “Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas”. In: *Journal of Theoretical Biology* 208.1 (2001), pp. 79–89. ISSN: 0022-5193. DOI: https://doi.org/10.1006/jtbi.2000.2202.

URL: http://www.sciencedirect.com/science/article/pii/S0022519300922021.

[90] Kevin Hoffman, David Zage, and Cristina Nita-Rotaru. "A survey of attack and defense techniques for reputation systems". In: *ACM Computing Surveys (CSUR)* 42.1 (2009), p. 1.

[91] Michael A. Hogg. "Social Identity and Social Comparison". In: *Handbook of Social Comparison: Theory and Research*. Ed. by Jerry Suls and Ladd Wheeler. Boston, MA: Springer US, 2000, pp. 401–421. ISBN: 978-1-4615-4237-7. DOI: 10.1007/978-1-4615-4237-7_19. URL: https://doi.org/10.1007/978-1-4615-4237-7_19.

[92] Michael A Hogg. "Social identity theory". In: *Understanding peace and conflict through social identity theory*. Springer, 2016, pp. 3–17.

[93] Michael A. Hogg. *The social psychology of group cohesiveness : from attraction to social identity*. English. Harvester Wheatsheaf New York ; Sydney, 1992, xi, 185 p. : ISBN: 0814734995 0745010636 0745010628.

[94] Matthew J. Hornsey. "Social Identity Theory and Self-categorization Theory: A Historical Review". In: *Social and Personality Psychology Compass* 2.1 (2008), pp. 204–222. DOI: 10.1111/j.1751-9004.2007.00066.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-9004.2007.00066.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-9004.2007.00066.x.

[95] Nicolas Houy. "Evolution of cooperation with similarity to an archetype". In: *Journal of theoretical biology* 332 (2013), pp. 78–88.

[96] E. Lance Howe, James J. Murphy, Drew Gerkey, and Colin Thor West. "Indirect Reciprocity, Resource Sharing, and Environmental Risk: Evidence from Field Experiments in Siberia". In: *Plos One* 11.7 (July 2016), pp. 1–17. DOI: 10.1371/journal.pone.0158940. URL: https://doi.org/10.1371/journal.pone.0158940.

[97] Lise Jans, Tom Postmes, and Karen I Van der Zee. "The induction of shared identity: The positive role of individual distinctiveness for groups". In: *Personality and Social Psychology Bulletin* 37.8 (2011), pp. 1130–1141.

[98] Vincent AA Jansen and Minus Van Baalen. "Altruism through beard chromodynamics". In: *Nature* 440.7084 (2006), pp. 663–666.

[99] Marco A. Janssen. "Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents". In: *Journal of Economic Behavior & Organization* 65.3 (2008), pp. 458–471. ISSN: 0167-2681. DOI: `https://doi.org/10.1016/j.jebo.2006.02.004`. URL: `http://www.sciencedirect.com/science/article/pii/S0167268106001934`.

[100] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[101] Michihiro Kandori. "Social Norms and Community Enforcement". In: *The Review of Economic Studies* 59.1 (Jan. 1992), pp. 63–80. ISSN: 0034-6527. DOI: `10.2307/2297925`. eprint: `https://academic.oup.com/restud/article-pdf/59/1/63/4356950/59-1-63.pdf`. URL: `https://doi.org/10.2307/2297925`.

[102] Peter M. Kappeler and Carel P. Van Schaik. *Cooperation in primates and humans: Mechanisms and evolution*. Springer, 2006, pp. 1–349. ISBN: 9783540282778. DOI: `10.1007/3-540-28277-7`.

[103] Kerry Kawakami, David M Amodio, and Kurt Hugenberg. "Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization". In: *Advances in experimental social psychology*. Vol. 55. Elsevier, 2017, pp. 1–80.

[104] Herbert C Kelman. "Compliance, identification, and internalization three processes of attitude change". In: *Journal of conflict resolution* 2.1 (1958), pp. 51–60.

[105] Max M Krasnow, Andrew W Delton, Leda Cosmides, and John Tooby. “Group cooperation without group selection: Modest punishment can recruit much cooperation”. In: *PloS one* 10.4 (2015).

[106] Kevin Lai, Michal Feldman, Ion Stoica, and John Chuang. “Incentives for cooperation in peer-to-peer networks”. In: *Workshop on economics of peer-to-peer systems*. 2003, pp. 1243–1248.

[107] Nicolas Lanchier. “The Axelrod model for the dissemination of culture revisited”. In: *The Annals of Applied Probability* 22.2 (Apr. 2012). ISSN: 1050-5164. DOI: 10.1214/11-aap790. URL: http://dx.doi.org/10.1214/11-AAP790.

[108] Olof Leimar and Peter Hammerstein. “Evolution of cooperation through indirect reciprocity”. In: *Proc. of the Royal Society B: Biological Sciences* 268.1468 (2001), pp. 745–753.

[109] Kun Li, Rui Cong, and Long Wang. “Stochastic evolutionary dynamics in minimum-effort coordination games”. In: *Physics Letters A* 380.34 (2016), pp. 2595–2602.

[110] Erez Lieberman, Christoph Hauert, and Martin A Nowak. “Evolutionary dynamics on graphs”. In: *Nature* 433.7023 (2005), p. 312.

[111] Gladys Loranger-Merciris, Laure Barthes, Alexandra Gastine, and Paul Leadley. “Rapid effects of plant species diversity and identity on soil microbial communities in experimental grassland ecosystems”. In: *Soil Biology and Biochemistry* 38.8 (2006), pp. 2336–2343. ISSN: 0038-0717. DOI: https://doi.org/10.1016/j.soilbio.2006.02.009. URL: http://www.sciencedirect.com/science/article/pii/S0038071706001374.

[112] Arthur Lupia, Mathew D. McCubbins, and Samuel L. Popkin, eds. *Elements of Reason: Cognition, Choice, and the Bounds of Rationality*. Cambridge Studies in Public Opinion and Political Psychology. Cambridge: Cambridge University Press, 2000. ISBN: 9780521653299. DOI: 10.1017/CBO9780511805813.

URL: https://www.cambridge.org/core/books/elements-of-reason/70ABD7CA28761475322742AA7E258B81.

[113] Gianluca Manzo and Toby Matthews. “Potentialities and limitations of agent-based simulations”. In: *Revue française de sociologie* 55.4 (2014), pp. 653–688.

[114] Hazel Markus and Elissa Wurf. “The dynamic self-concept: A social psychological perspective”. In: *Annual review of psychology* 38.1 (1987), pp. 299–337.

[115] James A.R. Marshall. “The donation game with roles played between relatives”. In: *Journal of Theoretical Biology* 260.3 (2009), pp. 386–391. ISSN: 0022-5193. DOI: https://doi.org/10.1016/j.jtbi.2009.07.008. URL: http://www.sciencedirect.com/science/article/pii/S0022519309003178.

[116] David Masad and Jacqueline Kazil. “MESA: an agent-based modeling framework”. In: *14th PYTHON in Science Conference*. 2015, pp. 53–60.

[117] Jorg J.M. Massen and Sonja E. Koski. “Chimps of a feather sit together: chimpanzee friendships are based on homophily in personality”. In: *Evolution and Human Behavior* 35.1 (2014), pp. 1–8. ISSN: 1090-5138. DOI: https://doi.org/10.1016/j.evolhumbehav.2013.08.008. URL: http://www.sciencedirect.com/science/article/pii/S1090513813000925.

[118] Naoki Masuda. “Ingroup favoritism and intergroup cooperation under indirect reciprocity based on group reputation”. In: *Journal of Theoretical Biology* 311 (2012), pp. 8–18. ISSN: 0022-5193. DOI: https://doi.org/10.1016/j.jtbi.2012.07.002. URL: https://www.sciencedirect.com/science/article/pii/S0022519312003360.

[119] M. McLure Wasko and S. Faraj. ““It is what one does”: why people participate and help others in electronic communities of practice”. In: *The Journal of Strategic Information Systems* 9.2 (2000), pp. 155–173. ISSN: 0963-8687. DOI: https://doi.org/10.1016/S0963-8687(00)00045-7. URL:

https://www.sciencedirect.com/science/article/pii/S0963868700000457.

[120] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a Feather: Homophily in Social Networks". In: *Annual Review of Sociology* 27.1 (2001), pp. 415–444. DOI: 10.1146/annurev.soc.27.1.415. eprint: https://doi.org/10.1146/annurev.soc.27.1.415. URL: https://doi.org/10.1146/annurev.soc.27.1.415.

[121] Manfred Milinski, Dirk Semmann, and Hans-Jürgen Krambeck. "Reputation helps solve the 'tragedy of the commons'". In: *Nature* 415.6870 (2002), pp. 424–426.

[122] Charles G Nathanson, Corina E Tarnita, and Martin A Nowak. "Calculating evolutionary dynamics in structured populations". In: *PLoS computational biology* 5.12 (2009), e1000615.

[123] Martin Nowak and Karl Sigmund. "A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game". In: *Nature* 364.6432 (1993), p. 56.

[124] Martin A Nowak. "Five rules for the evolution of cooperation". In: *science* 314.5805 (2006), pp. 1560–1563.

[125] Martin A Nowak and Robert M May. "Evolutionary games and spatial chaos". In: *Nature* 359.6398 (1992), p. 826.

[126] Martin A Nowak and Sébastien Roch. "Upstream reciprocity and the evolution of gratitude". In: *Proceedings of the Royal Society B: Biological Sciences* 274.1610 (2007), pp. 605–610. DOI: 10.1098/rspb.2006.0125. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.2006.0125. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2006.0125.

[127] Martin A Nowak and Karl Sigmund. "Evolution of indirect reciprocity". In: *Nature* 437.7063 (2005), pp. 1291–1298.

[128] Martin A Nowak and Karl Sigmund. “Evolution of indirect reciprocity by image scoring”. In: *Nature* 393.6685 (1998), pp. 573–577.

[129] Martin A Nowak, Corina E Tarnita, and Edward O Wilson. “The evolution of eusociality”. In: *Nature* 466.7310 (2010), p. 1057.

[130] P. J. Oakes and J. C. Turner. “Social categorization and intergroup behaviour: Does minimal intergroup discrimination make social identity more positive?” In: *European Journal of Social Psychology* 10.3 (1980), pp. 295–301. DOI: `10.1002/ejsp.2420100307`. URL: `https://doi.org/10.1002/ejsp.2420100307`.

[131] Hisashi Ohtsuki and Yoh Iwasa. “How should we define goodness?—reputation dynamics in indirect reciprocity”. In: *Journal of theoretical biology* 231.1 (2004), pp. 107–120.

[132] Hisashi Ohtsuki and Yoh Iwasa. “The leading eight: social norms that can maintain cooperation by indirect reciprocity”. In: *Journal of Theoretical Biology* 239.4 (2006), pp. 435–444.

[133] Hisashi Ohtsuki, Yoh Iwasa, and Martin A Nowak. “Indirect reciprocity provides only a narrow margin of efficiency for costly punishment”. In: *Nature* 457.7225 (2009), pp. 79–82.

[134] Hisashi Ohtsuki, Christoph Hauert, Erez Lieberman, and Martin A Nowak. “A simple rule for the evolution of cooperation on graphs and social networks”. In: *Nature* 441.7092 (2006), p. 502.

[135] Elinor Ostrom, James Walker, and Roy Gardner. “Covenants with and without a Sword: Self-Governance Is Possible”. In: *American Political Science Review* 86.2 (1992), 404–417. DOI: `10.2307/1964229`.

[136] Oxford University Press (OUP). *Reciprocity — Definition of Reciprocity by Lexico*. 2019. URL: `https://www.lexico.com/en/definition/reciprocity` (visited on 05/28/2020).

[137] Oxford University Press (OUP). *Reputation — Definition of Reputation by Lexico*. URL: https://www.lexico.com/en/definition/reputation (visited on 05/29/2020).

[138] Oxford University Press (OUP). *Trait — Definition of Trait by Lexico*. URL: https://www.lexico.com/en/definition/trait (visited on 05/31/2020).

[139] Hiroki Ozono, Yoshio Kamijo, and Kazumi Shimizu. "Punishing second-order free riders before first-order free riders: The effect of pool punishment priority on cooperation". In: *Scientific reports* 7.1 (2017), pp. 1–9.

[140] Jorge M Pacheco, Francisco C Santos, and Fabio AC C Chalub. "Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity". In: *PLoS computational biology* 2.12 (2006).

[141] Jorge M. Pacheco, Arne Traulsen, Hisashi Ohtsuki, and Martin A. Nowak. "Repeated games and direct reciprocity under active linking". In: *Journal of Theoretical Biology* 250.4 (2008), pp. 723–731. ISSN: 0022-5193. DOI: https://doi.org/10.1016/j.jtbi.2007.10.040. URL: http://www.sciencedirect.com/science/article/pii/S0022519307005450.

[142] Karthik Panchanathan and Robert Boyd. "A tale of two defectors: the importance of standing for evolution of indirect reciprocity". In: *Journal of Theoretical Biology* 224.1 (2003), pp. 115–126.

[143] Karthik Panchanathan and Robert Boyd. "Indirect reciprocity can stabilize cooperation without the second-order free rider problem". In: *Nature* 432.7016 (2004), pp. 499–502.

[144] Derek Parfit. "Personal Identity". In: *The Philosophical Review* 80.1 (1971), pp. 3–27. ISSN: 00318108, 15581470. URL: http://www.jstor.org/stable/2184309.

[145] Derek Parfit. "Personal Identity and Rationality". In: *Synthese* 53.2 (1982), pp. 227–241. ISSN: 00397857, 15730964. URL: http://www.jstor.org/stable/20115799.

[146] Elizabeth Pennisi. “On the Origin of Cooperation”. In: *Science* 325.5945 (2009), pp. 1196–1199. ISSN: 0036-8075. DOI: 10.1126/science.325_1196. eprint: https://science.sciencemag.org/content/325/5945/1196.full.pdf. URL: https://science.sciencemag.org/content/325/5945/1196.

[147] Steven Pinker. “The False Allure of Group Selection”. In: *The Handbook of Evolutionary Psychology*. American Cancer Society, 2015. Chap. 36, pp. 1–14. ISBN: 9781119125563. DOI: https://doi.org/10.1002/9781119125563.evpsych236. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119125563.evpsych236. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119125563.evpsych236.

[148] F. Pérez, B. E. Granger, and J. D. Hunter. “Python: An Ecosystem for Scientific Computing”. In: *Computing in Science Engineering* 13.2 (2011), pp. 13–21.

[149] Nichola J. Raihani, Alex Thornton, and Redouan Bshary. “Punishment and cooperation in nature”. In: *Trends in Ecology & Evolution* 27.5 (2012), pp. 288–295. ISSN: 0169-5347. DOI: https://doi.org/10.1016/j.tree.2011.12.004. URL: https://www.sciencedirect.com/science/article/pii/S016953471200002X.

[150] David G Rand and Martin A Nowak. “Human cooperation”. In: *Trends in Cognitive Sciences* 17.8 (2013), pp. 413–425.

[151] David G Rand, Hisashi Ohtsuki, and Martin A Nowak. “Direct reciprocity with costly punishment: Generous tit-for-tat prevails”. In: *Journal of theoretical biology* 256.1 (2009), pp. 45–57.

[152] Rick L Riolo, Michael D Cohen, and Robert Axelrod. “Evolution of cooperation without reciprocity”. In: *Nature* 414.6862 (2001), pp. 441–443.

[153] Gilbert Roberts. “Evolution of direct and indirect reciprocity”. In: *Proceedings of the Royal Society B: Biological Sciences* 275.1631 (2008), pp. 173–179.

[154] Angelo Romano, Daniel Balliet, and Junhui Wu. “Unbounded indirect reciprocity: Is reputation-based cooperation bounded by group membership?” In: *Journal of Experimental Social Psychology* 71 (2017), pp. 59–67. ISSN: 0022-1031. DOI: https://doi.org/10.1016/j.jesp.2017.02.008. URL: http://www.sciencedirect.com/science/article/pii/S0022103116307041.

[155] Fernando P Santos, Jorge M Pacheco, and Francisco C Santos. “Evolution of cooperation under indirect reciprocity and arbitrary exploration rates”. In: *Scientific reports* 6 (2016), p. 37517.

[156] Fernando P Santos, Francisco C Santos, and Jorge M Pacheco. “Social norm complexity and past reputations in the evolution of cooperation”. In: *Nature* 555.7695 (2018), p. 242.

[157] A Santos-Longhurst. “Intrinsic Motivation: How to Pick Up Healthy Motivation Techniques”. In: *URL: https://www.healthline.com/health/intrinsicmotivation [11.02. 2019]* (2019).

[158] Ingrid Seinen and Arthur Schram. “Social status and group norms: Indirect reciprocity in a repeated helping experiment”. In: *European Economic Review* 50.3 (2006), pp. 581–602. ISSN: 0014-2921. DOI: https://doi.org/10.1016/j.euroecorev.2004.10.005. URL: http://www.sciencedirect.com/science/article/pii/S0014292104001138.

[159] Karl Sigmund. “Complex adaptive systems and the evolution of reciprocation”. In: *AIP Conference Proceedings* 574.1 (2001), pp. 29–37. DOI: 10.1063/1.1386816. eprint: https://aip.scitation.org/doi/pdf/10.1063/1.1386816. URL: https://aip.scitation.org/doi/abs/10.1063/1.1386816.

[160] Karl Sigmund. *The calculus of selfishness*. Vol. 6. Princeton University Press, 2010.

[161] J Maynard Smith. “Group selection and kin selection”. In: *Nature* 201.4924 (1964), p. 1145.

[162] J Maynard Smith and George R Price. "The logic of animal conflict". In: *Nature* 246.5427 (1973), pp. 15–18.

[163] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982. DOI: 10.1017/cbo9780511806292.

[164] Anand Sriraman, Jonathan Bragg, and Anand Kulkarni. "Worker-Owned Cooperative Models for Training Artificial Intelligence". In: *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17 Companion. Portland, Oregon, USA: Association for Computing Machinery, 2017, 311–314. ISBN: 9781450346887. DOI: 10.1145/3022198.3026356. URL: https://doi.org/10.1145/3022198.3026356.

[165] Charles Stangor, Rajiv Jhangiani, and Hammond Tarry. *Principles of social psychology*. BCcampus, 2014.

[166] Claude M. Steele, Steven J. Spencer, and Joshua Aronson. "Contending with group image: The psychology of stereotype and social identity threat". In: vol. 34. Advances in Experimental Social Psychology. Academic Press, 2002, pp. 379–440. DOI: https://doi.org/10.1016/S0065-2601(02)80009-0. URL: http://www.sciencedirect.com/science/article/pii/S0065260102800090.

[167] Jan E. Stets and Chris F. Biga. "Bringing Identity Theory into Environmental Sociology". In: *Sociological Theory* 21.4 (2003), pp. 398–423. DOI: 10.1046/j.1467-9558.2003.00196.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1467-9558.2003.00196.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1467-9558.2003.00196.x.

[168] Jeffrey R. Stevens, Fiery A. Cushman, and Marc D. Hauser. "Evolving the Psychological Mechanisms for Cooperation". In: *Annual Review of Ecology, Evolution, and Systematics* 36.1 (2005), pp. 499–518. DOI: 10.1146/annurev.ecolsys.36.113004.083814. eprint: https://doi.org/10.1146/

annurev.ecolsys.36.113004.083814. URL: https://doi.org/10.1146/annurev.ecolsys.36.113004.083814.

[169] A.J. Stewart. *New take on game theory offers clues on why we cooperate*. 2015. URL: https://theconversation.com/new-take-on-game-theory-offers-clues-on-why-we-cooperate-38130http://theconversation.com/new-take-on-game-theory-offers-clues-on-why-we-cooperate-38130 (visited on 05/15/2020).

[170] Alexander J. Stewart and Joshua B. Plotkin. "Collapse of cooperation in evolving games". In: *Proceedings of the National Academy of Sciences* 111.49 (2014), pp. 17558–17563. ISSN: 0027-8424. DOI: 10.1073/pnas.1408618111. eprint: https://www.pnas.org/content/111/49/17558.full.pdf. URL: https://www.pnas.org/content/111/49/17558.

[171] Robert Sugden. *The economics of rights, co-operation and welfare*. Blackwell Oxford, 1986.

[172] Shinsuke Suzuki and Eizo Akiyama. "Reputation and the evolution of cooperation in sizable groups". In: *Proc. of the Royal Society B: Biological Sciences* 272.1570 (2005), pp. 1373–1377.

[173] Shinsuke Suzuki and Hiromichi Kimura. "Indirect reciprocity is sensitive to costs of information transfer". In: *Scientific reports* 3 (2013), p. 1435.

[174] William B. Swann, Michael D. Buhrmester, Angel Gómez, Jolanda Jetten, Brock Bastian, Alexandra Vázquez, Amarina Ariyanto, Tomasz Besta, Oliver Christ, Lijuan Cui, Gillian Finchilescu, Roberto González, Nobuhiko Goto, Matthew Hornsey, Sushama Sharma, Harry Susianto, and Airong Zhang. "What makes a group worth dying for? Identity fusion fosters perception of familial ties, promoting self-sacrifice." In: *Journal of Personality and Social Psychology* 106.6 (2014), pp. 912–926. DOI: 10.1037/a0036089. URL: https://doi.org/10.1037/a0036089.

[175] William B. Swann, Jolanda Jetten, Ãngel Gómez, Harvey Whitehouse, and Brock Bastian. “When group membership gets personal: A theory of identity fusion”. In: *Psychological Review* 119.3 (2012), pp. 441–456. ISSN: 0033295x. DOI: 10.1037/a0028589.

[176] William B Swann Jr and Michael D Buhrmester. “Identity fusion”. In: *Current Directions in Psychological Science* 24.1 (2015), pp. 52–57.

[177] William B. Swann Jr., Ángel Gómez, D. Conor Seyle, J. Francisco Morales, and Carmen Huici. “Identity fusion: The interplay of personal and social identities in extreme group behavior.” In: *Journal of Personality and Social Psychology* 96.5 (2009), pp. 995–1011. DOI: 10.1037/a0013668. URL: https://doi.org/10.1037/a0013668.

[178] Henri Tajfel. “Cognitive aspects of prejudice”. In: *Journal of biosocial science* 1.S1 (1969), pp. 173–191.

[179] Henri Tajfel. “Experiments in intergroup discrimination”. In: *Scientific American* 223.5 (1970), pp. 96–103.

[180] Henri Tajfel. *Human groups and social categories: Studies in social psychology*. Cup Archive, 1981.

[181] Henri Tajfel. “Social stereotypes and social groups.” In: Key readings in social psychology. New York, NY, US: Psychology Press, 2001, pp. 132–145. ISBN: 0-86377-678-7 (Hardcover); 0-86377-679-5 (Paperback).

[182] Henri Tajfel and John C Turner. “An integrative theory of intergroup conflict”. In: *The social psychology of intergroup relations* 33.47 (1979), p. 74.

[183] Henri Tajfel, M. G. Billig, R. P. Bundy, and Claude Flament. “Social categorization and intergroup behaviour”. In: *European Journal of Social Psychology* 1.2 (1971), pp. 149–178. DOI: https://doi.org/10.1002/ejsp.2420010202. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2420010202. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2420010202.

[184] Nobuyuki Takahashi and Rie Mashima. “The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity”. In: *Journal of Theoretical Biology* 243.3 (2006), pp. 418–436.

[185] Corina E Tarnita, Nicholas Wage, and Martin A Nowak. “Multiple strategies in structured populations”. In: *Proceedings of the National Academy of Sciences* 108.6 (2011), pp. 2334–2337.

[186] Corina E Tarnita, Tibor Antal, Hisashi Ohtsuki, and Martin A Nowak. “Evolutionary dynamics in set structured populations”. In: *Proceedings of the National Academy of Sciences* 106.21 (2009), pp. 8601–8604.

[187] Shelley E Taylor. “A categorization approach to stereotyping”. In: *Cognitive processes in stereotyping and intergroup behavior* 832114 (1981).

[188] Arne Traulsen and Martin A Nowak. “Chromodynamics of cooperation in finite populations”. In: *PLoS One* 2.3 (2007), e270.

[189] Arne Traulsen and Martin A Nowak. “Evolution of cooperation by multilevel selection”. In: *Proceedings of the National Academy of Sciences* 103.29 (2006), pp. 10952–10955.

[190] Arne Traulsen, Torsten Röhl, and Manfred Milinski. “An economic experiment reveals that humans prefer pool punishment to maintain the commons”. In: *Proceedings of the Royal Society B: Biological Sciences* 279.1743 (2012), pp. 3716–3721. DOI: 10.1098/rspb.2012.0937. eprint: https://royalsocietypublishin org/doi/pdf/10.1098/rspb.2012.0937. URL: https://royalsocietypubli org/doi/abs/10.1098/rspb.2012.0937.

[191] Sabine Trepte and Laura S Loy. “Social identity theory and self-categorization theory”. In: *The international encyclopedia of media effects* (2017), pp. 1–13.

[192] Robert L Trivers. “The evolution of reciprocal altruism”. In: *The Quarterly review of biology* 46.1 (1971), pp. 35–57.

[193] John C Turner, Michael A Hogg, Penelope J Oakes, Stephen D Reicher, and Margaret S Wetherell. *Rediscovering the social group: A self-categorization theory.* Basil Blackwell, 1987.

[194] Susan Tyler. "Social Categorization & Stereotyping". In: *Human Behavior and the Social Environment I* (2020).

[195] Tom R Tyler and Steven L Blader. "Identity and cooperative behavior in groups". In: *Group processes & intergroup relations* 4.3 (2001), pp. 207–226.

[196] Claus Wedekind and Manfred Milinski. "Cooperation through image scoring in humans". In: *Science* 288.5467 (2000), pp. 850–852.

[197] Roger M Whitaker, Gualtiero B Colombo, and Yarrow Dunham. "The evolution of strongly-held group identities through agent-based cooperation". In: *Scientific Reports* 11.1 (2021), pp. 1–16.

[198] Roger M Whitaker, Gualtiero B Colombo, and David G Rand. "Indirect reciprocity and the evolution of prejudicial groups". In: *Scientific reports* 8.1 (2018), p. 13247.

[199] Roger M Whitaker, Gualtiero B Colombo, Stuart M Allen, and Robin IM Dunbar. "A dominant social comparison heuristic unites alternative mechanisms for the evolution of indirect reciprocity". In: *Scientific reports* 6 (2016), p. 31459.

[200] Harvey Whitehouse, William Swann, Gordon Ingram, Karolina Prochownik, Johnathan Lanman, Timothy M Waring, Karl Frost, Douglas Jones, Zoey Reeve, and Dominic Johnson. "Three wishes for the world". In: *Cliodynamics: The Journal of Theoretical and Mathematical History* 4.2 (2013).

[201] George Christopher Williams. *Adaptation and natural selection: A critique of some current evolutionary thought*. Princeton university press, 2008.

[202] David Sloan Wilson. "A theory of group selection". In: *Proc. of the national academy of sciences* 72.1 (1975), pp. 143–146.

[203] David Sloan Wilson and Elliott Sober. “Reintroducing group selection to the human behavioral sciences”. In: *Behavioral and brain sciences* 17.4 (1994), pp. 585–608.

[204] Sewall Wright. “Evolution In Mendelian Populations”. In: *Genetics* 16.2 (1931), pp. 97–159. ISSN: 0016-6731. eprint: `https://www.genetics.org/content/16/2/97.full.pdf`. URL: `https://www.genetics.org/content/16/2/97`.

[205] Sewall Wright. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*. Vol. 1. 8. na, 1932.

[206] Jane Wu, Erin Paeng, Kari Linder, Piercarlo Valdesolo, and James C. Boerkoel. “Trust and Cooperation in Human-Robot Decision Making”. In: *AAAI Fall Symposia*. 2016.

[207] Junhui Wu, Daniel Balliet, and Paul AM Van Lange. “Reputation, gossip, and human cooperation”. In: *Social and Personality Psychology Compass* 10.6 (2016), pp. 350–364.

[208] Vero Copner Wynne-Edwards. *Animal dispersion: in relation to social behaviour.* Tech. rep. 1962.